

INSTITUTO CARLOS CHAGAS
Pós-Graduação em Biociências e Biotecnologia

DIOGO BORGES LIMA

Algoritmo para identificação de peptídeos covalentemente ligados e analisados por espectrometria de massas

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Biociências e Biotecnologia, do Instituto Carlos Chagas, como parte dos requisitos necessários à obtenção do título de Doutor em Ciências em Biociências e Biotecnologia.

Orientadores: Dr. Paulo Costa Carvalho
Dr. Fabio Cesar Gozzo

CURITIBA/PR
Janeiro de 2016

ALGORITMO PARA IDENTIFICAÇÃO DE PEPTÍDEOS COVALENTEMENTE
LIGADOS E ANALISADOS POR ESPECTROMETRIA DE MASSAS

Diogo Borges Lima

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO CARLOS CHAGAS
DA FUNDAÇÃO OSWALDO CRUZ (FIOCRUZ) COMO PARTE DOS
REQUISITOS NECESSÁRIOS PARA À OBTENÇÃO DO GRAU DE DOUTOR
EM CIÊNCIAS EM BIOCÊNCIAS E BIOTECNOLOGIA.

Examinada por:

Dr. Beatriz Gomes Guimarães

Dr. Marco Aurélio Krieger

Dr. Adriana Franco Paes Leme

CURITIBA, PR - BRASIL

JANEIRO DE 2016.

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

L732 Lima, Diogo Borges

Algoritmo para identificação de peptídeos covalentemente ligados e analisados por espectrometria de massas / Diogo Borges Lima. – Curitiba, 2015.
xix, 215 f. : il. ; 30 cm.

Tese (Doutorado) – Instituto Carlos Chagas, Pós-Graduação em Biociências e Biotecnologia, 2015.

Bibliografia: f. 60-68

1. *Crosslinking*. 2. Proteômica estrutural. 3. Bioinformática. I. Título.

CDD 572.65

*“C'est le temps que tu as perdu pour
ta rose qui fait ta rose si importante.”*

Antoine de Saint-Exupér

Agradecimentos

Agradeço, em primeiro lugar, a Deus pelas oportunidades que me foram dadas e às pessoas que conheci, as quais me proporcionaram a evolução do aprendizado.

Meus agradecimentos também à minha mãe, Mônica Maria Borges Lima, e aos meus avós, Rita do Amaral Borges e Benedito de Aragão Borges por sempre terem me apoiado ao longo de minha vida, e por terem me educado, para que hoje eu pudesse ser a pessoa que sou. Não posso deixar de agradecer à minha irmã, Aline Thaís Borges Lima, por ter me dado forças nessa jornada do doutorado, e aos meus sobrinhos Jonatas Borges Lima Stallone e Davi Borges Lima Stallone pelas alegrias e diversões que tivemos juntos.

Ao Dr. Paulo Costa Carvalho, pois além de ser um excepcional orientador, é um grande amigo que embarcou junto comigo nesse desafio multidisciplinar, compartilhando sua experiência e depositando confiança e muita dedicação; ao Dr. Felipe da Veiga Leprevost por fazer parte do grupo de proteômica computacional o qual faço parte, e me ajudar em vários momentos no desenvolvimento desta tese; ao Dr. Valmir Carneiro Barbosa, por me auxiliar no desenvolvimento deste trabalho científico; ao orientador Dr. Fabio Cesar Gozzo pelo grande apoio, atenção, amizade e dedicação na realização do projeto de doutorado.

Aos antigos e novos membros dos laboratórios que realizo cooperação, os quais são: Laboratório de Espectrometria de Massas Dalton/IQ – Unicamp, em especial: Tatiani de Lima e Mariana Fioramonte; Laboratório de Proteômica e Engenharia de Proteínas – ICC/Fiocruz, em especial: Dra. Tatiana Brasil; Laboratório de Toxinologia – IOC/Fiocruz, em especial: Dr. Jonas Perales, Dr. Richard Valente, Dra. Ana Gisele Ferreira, Dra. Giselle Brunoro, Dra. Aline Garcia, Dra. Karina Rebello, Dra. Monique Trugilho, Dr. Francisco Neto, Dr. Donat Chapeaurouge, Dr. André Ferreira, Dra. Surza Rocha, Carolina Nicolau, Joelma Saldanha, Luciana Girão, Marcelle Caminha, Monique Costa, Viviane Tostes e Viviane Bastos; e Unidade Proteômica/IQ – UFRJ, em especial: Dr. Gilberto Domont, Dr. Fábio Nogueira, Dr. Magno Junqueira, Dra. Livia Goto, Erika Velasquez, Gabriel Duarte, Nina Daddario, Jimmy Esneider, Livia Zamagna e Rafael Melani; pela ajuda e apoio em cada momento de meu doutorado.

À Dra. Juliana de Saldanha da Gama Fischer pela amizade e a troca de conhecimentos científicos em proteômica.

À Dra. Priscila Ferreira de Aquino pela amizade e troca de conhecimentos científicos em proteômica.

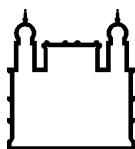
Aos amigos que me ajudaram e me deram forças para conseguir concluir mais esta etapa acadêmica. Agradeço em especial: Vitor Campos, Luana Machado, Mariana Areas, Elton Junior, Débora Andrade, Bruno Ferreira, Marcos Serpa, Angélica Corte, Ewellyn Barbosa, Vitor Rodrigues, Diego Cardoso, Jaqueline Velasco, Flávia Oliveira, Isabel Vilela, Danielle de Sá, Paulo de Souza, Denise Lima, Rodrigo Queixada, Aline Netto, Bruno Luiz, Mariana Apoteker, Diego Imenes, Rodrigo Campos, Pamela Figueiredo, Tiago Costa, Franciana Rosa, Nathalia Telles, Jaqueline Burlandi, Felipe Ribeiro, Priscila Hiraiwa e Ize Bittencourt.

À equipe EMMECAM *Challenging Team* por sempre confiar em mim e me dar forças para alcançar meus objetivos, em especial, ao professor Mário Mendonça.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela ajuda financeira, através da bolsa de estudos; ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), ao Programa de Apoio à Pesquisa Estratégica em Saúde (Papes) da Fiocruz, à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), e à *Microsoft Research*, pelo apoio financeiro para que esta tese pudesse ser realizada.

Aos membros presentes na banca desta tese.

E por fim, a todos que de alguma maneira contribuiu para que esta tese de doutorado pudesse ser concluída.



Ministério da Saúde

FIOCRUZ

Fundação Oswaldo Cruz

INSTITUTO CARLOS CHAGAS

ALGORITMO PARA IDENTIFICAÇÃO DE PEPTÍDEOS COVALENTEMENTE LIGADOS E ANALISADOS POR ESPECTROMETRIA DE MASSAS

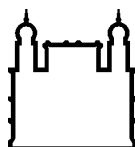
RESUMO

TESE DE DOUTORADO

Diogo Borges Lima

O estudo de estruturas e interações proteicas é uma importante área de pesquisa para se entender as funções das proteínas. No entanto, essa é também uma das áreas de grandes desafios experimentais, devido à inerente complexidade atômica de proteínas e peptídeos. Os métodos de elucidação estrutural de alta resolução (*e.g.* difração de raios-X e RMN) são hoje os considerados “padrões-ouro” para esses tipos de estudos. No entanto, uma grande parte das proteínas e seus respectivos complexos não são passíveis de serem resolvidos por esses métodos, motivando o desenvolvimento de novas técnicas para a caracterização estrutural de proteínas e seus complexos. Neste sentido, a espectrometria de massas acoplada à técnica de *cross-linking* (XL-MS) é uma grande promessa, devido às suas características intrínsecas, tais como alta sensibilidade e ampla aplicabilidade. Neste trabalho, desenvolveu-se um *software* com aplicações pioneiras, denominado SIM-XL, capaz de identificar peptídeos covalentemente ligados e analisados por espectrometria de massas, a fim de caracterizar estruturas de proteínas, bem como de complexos proteínas-proteínas e proteína-peptídeo. Esse *software* faz uso de técnicas de reconhecimento de padrões para resolver um gargalo na modelagem proteica e interação proteína-proteína. Portanto, o algoritmo aqui apresentado, traz benefícios imediatos nas áreas de biologia e biotecnologia e indiretamente, em diversas outras áreas, como por exemplo, no desenvolvimento de novos fármacos.

Palavras-chave: *crosslinking*, proteômica estrutural, bioinformática



Ministério da Saúde

FIOCRUZ

Fundação Oswaldo Cruz

INSTITUTO CARLOS CHAGAS

ALGORITHM FOR IDENTIFYING CROSS-LINKED PEPTIDES AND ANALYZED BY MASS SPECTROMETRY

ABSTRACT

THESIS OF DOCTORATE

Diogo Borges Lima

The study of protein structures and interactions is an important area of development for understanding the function of proteins. However, this is also an area of great experimental challenge, due to the inherent atomic complexity of proteins and peptides. The methods of structural elucidation of high-resolution (*e.g.* X-ray diffraction and NMR) are currently considered the “gold standard” for these types of studies. However, many proteins are not amenable to being solved by these methods; thus motivating the development of new techniques for structural characterization of proteins and their complexes. In this regard, mass spectrometry coupled by cross-linking technique (XL-MS) poses as a promise to overcome these limitations as it provides a high sensitivity and wide applicability. Here we present SIM-XL, a software pioneer in many ways, capable of identifying cross-linked peptides analyzed by mass spectrometry and thus ultimately aiding in structural characterization and in determining protein-protein interactions. Our software uses pattern recognition strategies to address a bottleneck in protein modeling and protein-protein interaction. As such, various fields related to biology and biotechnology suffer an immediate benefit from this work, and other areas, say, the development of new drugs, are indirectly benefited as well.

Key-words: cross-link, proteomics, bioinformatics

Sumário

LISTA DE FIGURAS	XII
LISTA DE TABELAS	XVI
LISTA DE ABREVIATURAS, SÍMBOLOS E UNIDADES	XVII

1	INTRODUÇÃO	1
1.1	COMPLEXOS PROTEICOS	2
1.2	PROTEÔMICA	4
1.2.1	PROTEÔMICA ESTRUTURAL E INTERAÇÃO PROTEICA	5
1.2.2	ESPECTROMETRIA DE MASSAS (MS) APLICADA À PROTEÔMICA	6
1.3	LIGAÇÃO COVALENTE DE PEPTÍDEOS PARA AUXILIAR NA DETERMINAÇÃO DE INTERAÇÃO PROTEÍNA-PROTEÍNA E A ESTRUTURA PROTEICA POR ESPECTROMETRIA DE MASSAS (<i>CROSS-LINKING</i>)	10
1.4	ALGORITMOS PARA OBTENÇÃO DA SEQUÊNCIA DE PEPTÍDEOS ANALISADOS POR ESPECTROMETRIA DE MASSAS	15
1.5	O FUNCIONAMENTO DE UMA FERRAMENTA DE BUSCA COM ABORDAGEM PSM	18
1.6	COMPLEXIDADE DO ESPAÇO DE BUSCA	20
2	JUSTIFICATIVA	21
3	OBJETIVOS	22
3.1	OBJETIVO GERAL	22
3.2	OBJETIVOS ESPECÍFICOS	22
4	METODOLOGIA	23
4.1	FLUXO DE TRABALHO DO ALGORITMO DE IDENTIFICAÇÃO	23
4.1.1	IDENTIFICAÇÃO PRÉVIA DE PEPTÍDEOS LINEARES COM E SEM <i>MONO-LINKS</i>	24
4.1.2	REDUÇÃO DINÂMICA DO BANCO DE DADOS	24
4.1.3	FILTRAGEM DE ESPECTROS CARACTERÍSTICOS DE <i>CROSS-LINKING</i>	26
4.1.4	IDENTIFICAÇÃO DOS PEPTÍDEOS COVALENTEMENTE LIGADOS	27
4.1.4.1	Indexação do banco de dados de peptídeos	27
4.1.4.2	O algoritmo minhoca	28
4.1.4.3	Preditor de espectro teórico para peptídeos com <i>interlink</i>	28
4.1.4.4	Preditor de espectro teórico para peptídeos com <i>intralink</i>	30
4.1.4.5	Otimizações no espectro teórico	31
4.1.4.6	Métrica de comparação entre espectros teóricos e experimentais.	32
4.1.5	AVALIAÇÃO DOS RESULTADOS DA BUSCA	33
4.1.5.1	RANSAC	33
4.1.6	AUXÍLIO NA MODELAGEM ESTRUTURAL DE PROTEÍNAS E INTERAÇÕES PROTEICAS	35
4.2	<i>SPECTRUM IDENTIFICATION MACHINE FOR CROSS-LINKED PEPTIDES</i>	35
4.2.1	PARÂMETROS	36
4.2.1.1	Modos de operação	36
4.2.2	LINGUAGEM DE PROGRAMAÇÃO	39
4.2.2.1	<i>Model-View-Controller</i> – MVC	39
4.2.3	INTERFACE GRÁFICA	40
4.2.4	LEITURA DOS ESPECTROS EXPERIMENTAIS	41
4.2.5	GERAÇÃO DO ARQUIVO DE RESULTADOS	41
4.2.5.1	Mapa bidimensional de interação proteína-proteína	42
4.2.5.1.1	Visualizador em barras	42
4.2.5.1.2	Visualizador circular	42

4.2.5.1	Relatório dinâmico de resultados	44
4.2.5.2	Mapa de calor das interações proteicas	44
4.2.5.3	Visualização dos <i>links</i> em estruturas terciárias ou quaternárias	46
4.2.5.4	Visualizador dinâmico de espectros	46
4.3	REPOSITÓRIO DE ARMAZENAMENTO DE DADOS – <i>PRIDE</i>	47
4.4	O <i>SIM-XL</i> E O <i>PATTERNLAB FOR PROTEOMICS</i>	48
5	<u>RESULTADOS</u>	50
6	<u>DISCUSSÃO E CONCLUSÕES</u>	53
7	<u>PRODUÇÃO CIENTÍFICA E/OU COLABORAÇÕES</u>	55
7.1	COLABORAÇÕES EM ANDAMENTO	55
7.1.1	<i>PATTERNLAB FOR PROTEOMICS</i> v.4.0	56
7.2	ARTIGOS PUBLICADOS EM COAUTORIA	57
8	<u>PERSPECTIVAS</u>	59
9	<u>REFERÊNCIAS BIBLIOGRÁFICAS</u>	60
10	<u>ANEXOS I – ARTIGOS PUBLICADOS</u>	69

Lista de Figuras

FIGURA 1.1 – PROTEÍNAS PERTENCENTES AO COMPLEXO NADH UBIQUINONA REDUTASE AO LONGO DA EVOLUÇÃO DOS ORGANISMOS. (GABALDÓN ET AL., 2005).....	3
FIGURA 1.2 – REPRESENTAÇÃO ESQUEMÁTICA DE UMA REDE DE INTERAÇÃO DE <i>H. SAPIENS</i> ENVOLVENDO 401 PROTEÍNAS LIGADAS ATRAVÉS DE 911 INTERAÇÕES. (STELZL ET AL., 2005).....	4
FIGURA 1.3 – FLUXOGRAMA DAS ETAPAS DE UM ESPECTRÔMETRO DE MASSAS [FIGURA MODIFICADA A PARTIR DE (CARVALHO; BARBOSA, 2010)].....	6
FIGURA 1.4 – REPRESENTAÇÃO DE UMA IONIZAÇÃO POR <i>ELECTROSPRAY</i> DE PARTÍCULAS EM SOLUÇÃO. [FIGURA ADAPTADA DE (KEBARLE; VERKERK, 2009)]	7
FIGURA 1.5 – REPRESENTAÇÃO ESQUEMÁTICA DOS TIPOS DE FRAGMENTAÇÃO ENTRE OS AMINOÁCIDOS QUE PODEM OCORRER ATRAVÉS DO PROCESSO DE COLISÃO CELULAR CRIADA POR <i>ROEPSTORFF-FOHLMANN-BIEMANN</i> . AS LINHAS EM VERMELHO INDICAM POSSÍVEIS REGIÕES DE FRAGMENTAÇÃO ENTRE UM DETERMINADO AMINOÁCIDO DE UM PEPTÍDEO DE TAMANHO <i>N</i> . NOS ÍONS <i>A</i> , <i>B</i> E <i>C</i> , A CARGA FICA RETIDA NO N-TERMINAL, ENQUANTO QUE NOS ÍONS <i>X</i> , <i>Y</i> E <i>Z</i> , A CARGA FICA NO C-TERMINAL.....	8
FIGURA 1.6 – REPRESENTAÇÃO DOS ÍONS FRAGMENTOS DA SÉRIE <i>B</i> EM UM ESPECTRO, REPRESENTANDO A SEQUÊNCIA PEPTÍDICA AEP TIR.	9
FIGURA 1.7 – REPRESENTAÇÃO DOS ÍONS FRAGMENTOS DA SÉRIE <i>Y</i> EM UM ESPECTRO, FORMANDO A SEQUÊNCIA AEP TIR. A INTERPRETAÇÃO DESTA SÉRIE É FEITA NO SENTIDO C-TERMINAL PARA O N-TERMINAL, OU SEJA, LENDO O ESPECTRO DA DIREITA PARA A ESQUERDA.....	9
FIGURA 1.8 – ESPECTRO DE MASSAS CONTENDO AS SÉRIES <i>B</i> E <i>Y</i> , ASSIM COMO OUTROS PICOS REPRESENTADO OS RUÍDOS.....	9
FIGURA 1.9 – AGENTE DE LIGAÇÃO CRUZADA DO TIPO HOMOBIFUNCIONAL (DSS), O QUAL POSSUI GRUPOS AMINO REATIVOS NHS-ÉSTER IDÊNTICOS EM AMBOS OS LADOS DA CADEIA ESPAÇADORA. ESSA, POR SUA VEZ, TEM TAMANHO DE 11.4 Å, QUE É A DISTÂNCIA MÁXIMA ENTRE DUAS MOLÉCULAS CONJUGADAS.[FIGURA ADAPTADA DE (“DSS (DISUCCINIMIDYL SUBERATE)”)].....	12
FIGURA 1.10 – AGENTE DE LIGAÇÃO CRUZADA DO TIPO HETEROBIFUNCIONAL (SULFO-SMCC), O QUAL POSSUI NO LADO ESQUERDO DA CADEIA ESPAÇADORA, UM GRUPO AMINO-REATIVO, SULFO-NHS-ÉSTER, E DO OUTRO LADO UM GRUPO REATIVO <i>SULFHIDRYL MALEIMIDE</i> . [FIGURA ADAPTADA DE (“SULFO-SMCC (SULFOSUCCINIMIDYL 4-(N-MALEIMIDOMETHYL)CYCLOHEXANE-1-CARBOXYLATE)”)].....	12
FIGURA 1.11 – AGENTE DE LIGAÇÃO CRUZADA CARACTERIZADA COMO <i>ZERO-LENGTH</i> , O QUAL FAZ A LIGAÇÃO ENTRE DUAS PROTEÍNAS SEM QUE HAJA UMA CADEIA	

ESPAÇADORA ENTRE ELAS. [FIGURA ADAPTADA DE (“EDC (1-ETHYL-3-(3-DIMETHYLAMINOPROPYL)CARBODIIMIDE HYDROCHLORIDE)”)].....	13
FIGURA 1.12 – ESQUEMA DE UM EXPERIMENTO TÍPICO DE LIGAÇÃO CRUZADA ACOPLADA À MS. UMA AMOSTRA PROTEICA É SUBMETIDA À AÇÃO DOS <i>CROSS-LINKERS</i> , ONDE APÓS A DIGESTÃO ENZIMÁTICA, TRÊS TIPOS DE PRODUTOS PODEM SER GERADOS: <i>CROSS-LINK</i> , <i>LOOP-LINK</i> OU <i>MONO-LINK</i> . NA ETAPA SEGUINTE, OCORRE UM ENRIQUECIMENTO DOS <i>CROSS-LINKS</i> , COMO POR EXEMPLO, UMA PURIFICAÇÃO; A AMOSTRA É INJETADA EM UM ESPECTRÔMETRO DE MASSAS DE ALTA RESOLUÇÃO, ONDE OS DADOS BRUTOS SERÃO GERADOS E ANALISADOS POSTERIORMENTE PELA FERRAMENTA COMPUTACIONAL SIM-XL, PODENDO ASSIM CARACTERIZAR O MODELO ESTRUTURAL DA PROTEÍNA. [FIGURA ADAPTADA DE (LEITNER ET AL., 2014) E HTTP://DALTONLAB.IQM.UNICAMP.BR/RESEARCH.HTML]	14
FIGURA 1.13 – METODOLOGIA DE SEQUENCIAMENTO DE ESPECTROS: <i>DE NOVO SEQUENCING</i>	17
FIGURA 1.15 – FLUXO DE TRABALHO DE UMA FERRAMENTA DE BUSCA QUE UTILIZA A ABORDAGEM <i>PEPTIDE SPECTRUM MATCHING</i> – PSM [FIGURA MODIFICADA A PARTIR DE (CARVALHO; BARBOSA, 2010)]......	19
FIGURA 4.1 – ESPECTRO DE FRAGMENTAÇÃO DERIVADO DE PEPTÍDEOS COVALENTEMENTE LIGADOS. EM DESTAQUE, OS ÍONS DIAGNÓSTICOS DE <i>M/Z</i> 222.149 E 305.222 QUE CARACTERIZAM A EXISTÊNCIA DA REAÇÃO COM UM AGENTE DE LIGAÇÃO CRUZADA.....	27
FIGURA 4.2 – ESTRUTURA MOLECULAR DOS ÍONS DIAGNÓSTICOS DE <i>M/Z</i> 222,1494; 239,1759 E 305,2229; OS QUAIS CARACTERIZAM ESPECTROS PROVENIENTES DE PEPTÍDEOS QUE REAGIRAM COVALENTEMENTE COM O DSS. [FIGURA RETIRADA DE (IGLESIAS ET AL., 2010)].....	27
FIGURA 4.3 – PREDITOR DO ESPECTRO TEÓRICO PARA PEPTÍDEOS COM <i>INTERLINK</i> . AS SÉRIES <i>B</i> E <i>Y</i> COMBINADAS COM AS CADEIAS α E β SÃO OBTIDAS DE MANEIRA A COMPREENDER A LIGAÇÃO COVALENTE ENTRE OS DOIS PEPTÍDEOS.	30
FIGURA 4.4 – PREDITOR DE ESPECTRO TEÓRICO PARA PEPTÍDEOS COM <i>INTRALINK</i> . AS SÉRIES QUE COMPÕEM ESTE ESPECTROS SÃO A <i>B</i> E <i>Y</i>	30
FIGURA 4.5 – INTERVALO DE <i>M/Z</i> A SER PROCURADO EM CADA TIPO DE CARGA DA SEQUÊNCIA PEPTÍDICA. COMO EXEMPLO, AO PROCURAR ÍONS DE CARGA 1 ⁺ , O INTERVALO SERÁ [1, 2.000], E DIFICILMENTE ÍONS QUE CONTENHAM <i>M/Z</i> ACIMA DE 2.000 SERÃO MONO CARREGADOS.....	31
FIGURA 4.6 – CÁLCULO DO <i>SCORE</i> : A PONTUAÇÃO OBTIDA NESSE ESPECTRO FOI DE 3.33 (CANTO SUPERIOR DIREITO), OBTIDO A PARTIR DO PRODUTO ESCALAR ENTRE O ESPECTRO TEÓRICO E O EXPERIMENTAL, IGUAL A 0.83, ACRESCIDO DOS 25 PICOS (15 DA SÉRIE <i>Y</i> , EM AZUL, E 10 DA SÉRIE <i>B</i> , EM VERMELHO) CONTIDOS NO ESPECTRO TEÓRICO E PRESENTES NO EXPERIMENTAL.	32

FIGURA 4.7 – ANOTAÇÃO PEPTÍDICA: A PARTIR DELA É POSSÍVEL OBSERVAR O NÚMERO DE RESÍDUOS EM CADA CADEIA PEPTÍDICA (α EM AZUL E β EM VERMELHO) QUE FORAM IDENTIFICADOS.....	33
FIGURA 4.8 – LINHAS RANSAC QUE É INTERPRETADA COMO O ERRO MÉDIO ENTRE OS ÍONS DO ESPECTRO EXPERIMENTAL E TEÓRICO. NESTE EXEMPLO, É POSSÍVEL OBSERVAR QUE O ERRO PPM ABSOLUTO É DE APROXIMADAMENTE 5 E QUE PRATICAMENTE TODOS OS ÍONS ESTÃO DENTRO DA REGIÃO DEMARCADA PELA LINHA CENTRAL RANSAC E PELAS LINHAS EXTREMAS QUE CORRESPONDEM TRÊS DESVIOS-PADRÃO PARA CIMA E PARA BAIXO.	34
FIGURA 4.9 – LISTA DE <i>CROSS-LINKINGS</i> GERADA PARA PROPOR FUTUROS MODELOS PROTEICOS E/OU INTERAÇÕES QUE OCORREM ENTRE AS PROTEÍNAS.	35
FIGURA 4.10 – VISUALIZAÇÃO PELO <i>PYMOL</i> DE UM MODELO DE UMA PROTEÍNA E OS <i>CROSS-LINKINGS</i> IDENTIFICADOS PELO ALGORITMO DE BUSCA.	35
FIGURA 4.11 – INTERAÇÃO DOS COMPONENTES DO MVC [<i>FIGURA RETIRADA DE (APPEL)</i>]	40
FIGURA 4.12 – INTERFACE PRINCIPAL DO SIM-XL, ONDE OS PRINCIPAIS PARÂMETROS DE BUSCA SÃO DEFINIDOS.....	41
FIGURA 4.13 – MAPA BIDIMENSIONAL EM FORMA DE BARRAS. ESSA FORMA DE VISUALIZAR OS RESULTADOS É INÉDITO EM FERRAMENTAS DE BUSCA DE <i>CROSS-LINKING</i>	42
FIGURA 4.14 – MAPA BIDIMENSIONAL EM FORMA CIRCULAR. ESSE TIPO DE VISUALIZAÇÃO É FACTÍVEL QUANDO MUITAS PROTEÍNAS SÃO IDENTIFICADAS, TORNANDO A INTERPRETAÇÃO MAIS CLARA E OBJETIVA.	43
FIGURA 4.15 – É POSSÍVEL DESTACAR DUAS PROTEÍNAS PARA QUE SE POSSA VISUALIZAR DE UMA MELHOR MANEIRA A INTERAÇÃO QUE AS DUAS TÊM ENTRE SI. E PODE-SE ALTERNAR PARA A VISUALIZAÇÃO EM BARRAS UMA PROTEÍNA ESPECÍFICA.	43
FIGURA 4.16 – RELATÓRIO DINÂMICO DE RESULTADOS. ATRAVÉS DELE É POSSÍVEL ANALISAR CADA IDENTIFICAÇÃO DETALHADAMENTE.....	44
FIGURA 4.17 – VISUALIZADOR DO MAPA DE CALOR DINÂMICO. COM ELE É POSSÍVEL OBSERVAR QUANTAS IDENTIFICAÇÕES CADA <i>LINK</i> OBTIVEU, REPRESENTADO POR CÉLULAS DIFERENTES. É POSSÍVEL SABER TAMBÉM, CADA IDENTIFICAÇÃO OBTIDA PARA O <i>LINK</i> APENAS CLICANDO SOBRE A CÉLULA DESEJADA.....	45
FIGURA 4.18 – MAPA DE CALOR OBTIDO ATRAVÉS DAS INTERAÇÕES ENTRE DUAS PROTEÍNAS. É POSSÍVEL VERIFICAR, ATRAVÉS DAS INTENSIDADES DE CORES EM CADA CÉLULA DO MAPA, EM QUAIS REGIÕES FORAM OBTIDAS UM MAIOR NÚMERO DE IDENTIFICAÇÕES PARA CADA <i>LINK</i> OBSERVADO.....	45
FIGURA 4.19 – GERAÇÃO DO <i>SCRIPT PYMOL</i> . COM ELE É POSSÍVEL CRIAR MODELOS 3D DE ESTRUTURAS TERCIÁRIAS DE PROTEÍNAS OU QUATERNÁRIAS, NO CASO DE COMPLEXOS PROTEICOS.	46

FIGURA 4.20 – VISUALIZADOR DE ESPECTRO DINÂMICO. É POSSÍVEL CUSTOMIZAR O ESPECTRO OBSERVADO A FIM DE OBTER UM MELHOR RESULTADO.	47
FIGURA 4.21 – <i>PATTERNLAB FOR PROTEOMICS</i> É UM AMBIENTE COMPUTACIONAL COMPOSTO POR VÁRIOS MÓDULOS PARA ANÁLISE DE PROTEÔMICA QUALITATIVA E QUANTITATIVA.	49
FIGURA 5.1 – IDENTIFICAÇÃO DE ESPECTROS DE ACORDO COM CADA FERRAMENTA ANALISADA. O CRUX FOI A FERRAMENTA QUE APRESENTOU A MENOR EFICIÊNCIA, IDENTIFICANDO CORRETAMENTE APENAS 50% DOS ESPECTROS APRESENTADOS; LOGO EM SEGUIDA, O PLINK NÃO APRESENTOU RESULTADOS ERRÔNEOS, ENTRETANTO, SÓ CONSEGUIU IDENTIFICAR 29 DOS 50 ESPECTROS APRESENTADOS. A FERRAMENTA DESENVOLVIDA NESSE TRABALHO, O SIM-XL, CONSEGUIU IDENTIFICAR 100% DOS ESPECTROS APRESENTADOS NO MODO MAIS OTIMIZADO.	51
FIGURA 5.2 – TEMPO DE PROCESSAMENTO, EM SEGUNDOS, DAS FERRAMENTAS DE BUSCA APÓS ANALISAR UM CONJUNTO DE DADOS RELACIONADOS À PROTEÍNA HSP90.	51
FIGURA 5.3 – EXEMPLOS DE PAÍSES QUE REALIZARAM <i>DOWNLOAD</i> DO SIM-XL, TOTALIZANDO MAIS DE 800 ATÉ O DIA 2 DE NOVEMBRO DE 2015.	52
FIGURA 5.4 – MAPA-MÚNDI REPRESENTANDO OS PAÍSES OS QUAIS FORAM REALIZADOS <i>DOWNLOAD</i> DO SIM-XL. QUANTO MAIS ESCURO O PAÍS É EXIBIDO, MAIOR É O NÚMERO DE <i>DOWNLOAD</i> FEITOS.	52

Lista de Tabelas

TABELA 1.1 – GRUPOS REATIVOS E OS RESPECTIVOS AGENTES DE LIGAÇÃO CRUZADA QUE REAGEM COM ELES.....	11
TABELA 1.2 – NÚMERO DE PEPTÍDEOS EM DIFERENTES ESPAÇOS DE BUSCA A PARTIR DA DIGESTÃO IN SILICO DE PROTEÍNAS DE <i>E. COLI</i> . O ESPAÇO DE BUSCA TOTALMENTE ESPECÍFICO É AQUELE COMPOSTO POR PEPTÍDEOS ONDE A CLIVAGEM ENZIMÁTICA OCORRE TANTO NO N-TERMINAL QUANTO NO C-TERMINAL DA CADEIA PEPTÍDICA. JÁ O SEMI-ESPECÍFICO É COMPOSTO POR PEPTÍDEOS TOTALMENTE ESPECÍFICOS ALÉM DE SER COMPOSTO POR AQUELES ONDE A CLIVAGEM ENZIMÁTICA OCORRE APENAS EM UMA EXTREMIDADE DA CADEIA, OU NO N-TERMINAL OU NO C-TERMINAL. E POR ÚLTIMO, O ESPAÇO NÃO-ESPECÍFICO É AQUELE COMPOSTO PELOS DOIS ESPAÇOS ANTERIORES MAIS OS PEPTÍDEOS CLIVADOS EM QUALQUER PARTE DE SUA CADEIA.....	20
TABELA 4.1 – ESPAÇO DE BUSCA TÍPICO DE UM EXPERIMENTO DE <i>CROSS-LINKING</i> ONDE TODOS OS PEPTÍDEOS COMBINAM ENTRE SI.....	25
TABELA 4.2 – ESPAÇO DE BUSCA APLICANDO A REDUÇÃO DINÂMICA, ONDE SERÁ COMPOSTO APENAS PELA COMBINAÇÃO DOS PEPTÍDEOS IDENTIFICADOS COM <i>DEAD-END</i> – K(DE), E OS PEPTÍDEOS QUE NÃO SOFRERAM NENHUMA MODIFICAÇÃO PÓS-TRADUCIONAL.....	26
TABELA 4.3 – INDEXAÇÃO DO BANCO DE SEQUÊNCIAS PROTEICAS A FIM DE AUMENTAR A VELOCIDADE DA BUSCA, NÃO DESPERDIÇANDO TEMPO EM GERAR ESPECTROS TEÓRICOS PARA PEPTÍDEOS CUJA MASSA ESTEJA FORA DO INTERVALO A SER BUSCADO.....	28
TABELA 4.4 – RESÍDUOS E MOLÉCULAS COM SUAS RESPECTIVAS MASSAS EM DA.....	38

Lista de abreviaturas, símbolos e unidades

ALC	Agentes de ligação cruzada
BS ³	Suberato de bis(sulfosuccinimidila), do inglês, <i>Bis(sulfosuccinimidyl)suberate</i>
CID	Dissociação por colisão induzida, do inglês, <i>Collision-Induced Dissociation</i>
d.d.p.	Diferença de potencial
Da	Dalton, unidade de massa atômica
DSS	Suberato de N,N-disuccinimidila, do inglês, <i>Disuccinimidyl Suberate</i>
DDA	Aquisição dependente dos dados, do inglês, <i>Data Dependent Acquisition</i>
DNA	Ácido desoxirribonucleico, do inglês, <i>deoxyribonucleic acid</i>
E	Aminoácido ácido glutâmico
ESI	Ionização por <i>electrospray</i> ; do inglês, <i>Electrospray Ionization</i>
ETD	Dissociação por transferência de elétrons, do inglês, <i>Electron Transfer Dissociation</i>
FT-ICR	Ressonância ciclôtrônica de íons por transformada de Fourier, do inglês, <i>Fourier transform ion cyclotron resonance</i>
G	Aminoácido glicina
GUI	Interface gráfica para o usuário, do inglês, <i>graphic user interface</i>
H	Aminoácido histidina
HCD	Dissociação por colisão de alta energia, do inglês, <i>Higher Energy Collision Dissociation</i>
HDL	Lipoproteína de alta densidade, do inglês, <i>High Density Lipoprotein</i>
HSP90	Proteína de choque térmico 90, do inglês, <i>Heat shock protein 90</i>
IA	Inteligência Artificial
iTRAQ	Marcadores isóbaros para a quantificação relativa e absoluta, do inglês, <i>Isobaric Tags for Relative and Absolute Quantification</i>
K	Aminoácido lisina

MALDI	Ionização por dessorção a <i>laser</i> assistida por matriz; do inglês, <i>Matrix-assisted laser desorption/ionization</i>
mg	Miligrama
MS	Espectrometria de Massas, do inglês, <i>Mass Spectrometry</i>
MS2 ou MS/MS	Varredura de íons produtos
<i>m/z</i>	Razão massa/carga
N	Aminoácido asparagina
NADH	Nicotinamida adenina dinucleótido hidreto
NHS	N-hidroxisuccinimida
O ₂	Molécula de oxigênio
P	Aminoácido prolina
pdf	Função de densidade de probabilidade, do inglês, <i>probability density function</i>
ppm	parte por milhão
PRIDE	Banco de dados de identificações proteômicas, do inglês, <i>Proteomics Identifications database</i>
PSM	<i>Peptide Spectrum Matching</i>
PTM	Modificação pós-traducional, do inglês, <i>post-translational modification</i>
Q	Aminoácido glutamina
Q-TOF	Quadrupolo-tempo de voo, do inglês, <i>Quadrupole-time-of-flight</i>
R	Aminoácido arginina
RMN	Ressonância Magnética Nuclear
RNA	Ácido ribonucleico, do inglês, <i>Ribonucleic acid</i>
S	Aminoácido serina
SIM-XL	<i>Spectrum Identification Machine for Cross-linked peptides</i>
Sulfo-SMCC	<i>Sulfosuccinimidyl 4-(M-maleimidomethyl)cyclohexane-1-carboxylate</i>
T	Aminoácido treonina
TOF-TOF	Tempo de voo-tempo de voo, do inglês, <i>Time-of-flight-Time-of-flight</i>
XL-MS	<i>Cross-linker</i> associado à espectrometria de massas

1 Introdução

A caracterização da estrutura de proteínas e o estudo de interações proteicas são fundamentais para o entendimento da função de proteínas. No entanto, esta área de pesquisa apresenta grandes desafios experimentais devido à complexidade molecular de proteínas e peptídeos. Atualmente, os métodos de alta resolução (*e.g.*, cristalografia de raios-X e ressonância magnética nuclear – RMN) são considerados os padrões-ouro nas análises estruturais. Apesar dessas técnicas serem bastante utilizadas por causa da precisão e do poder de resolução, elas possuem limitações. De uma maneira geral, uma parcela significativa de proteínas e seus respectivos complexos não são passíveis de análise por esses métodos. Alguns exemplos para ilustrar tal fato se deve que diversas proteínas não cristalizam ou geram cristais que difratam mal, o que limita o uso de técnicas cristalográficas. Outro desafio inclui o tamanho do complexo, uma vez que a técnica de espectroscopia de RMN normalmente é mais passível na caracterização de estruturas proteicas pequenas (até 50 kDa); vale ressaltar que a mesma também contribui com informações acerca da dinâmica de movimentos moleculares e das interfaces de interação. Além disso, ambas as técnicas requerem grande quantidade (na ordem de miligramas) de proteína com alto grau de pureza (ALBER et al., 2008; MERKLEY et al., 2013; STRONG et al., 2006).

Neste sentido, abordagens híbridas, as quais baseiam-se em métodos bioquímicos e biofísicos, vem sendo utilizadas cada vez mais na Biologia Estrutural, a fim de gerar dados complementares de baixa resolução. Do ponto de vista da modelagem molecular, técnicas como *small-angle X-ray scattering* – SAXS (LIPFERT; DONIACH, 2007; SVERGUN; KOCH, 2003) e crio-eletro microscopia – cryo-EM (CALLAWAY, 2015) podem gerar modelos bastante detalhistas – *near-atomic resolution*. (CHEN et al., 2015; SINZ et al., 2015).

A espectrometria de massas é considerada o padrão-ouro na caracterização, identificação e quantificação de peptídeos e proteínas. Ela também vem se consolidando como um importante método analítico para o mapeamento de interação proteína-proteína e a caracterização de conformações tridimensionais. Tal fato se deve porque a associação com a técnica de ligações cruzadas (XL-MS) pode gerar

dados estruturais complementares e com alta sensibilidade (na ordem de femtomols¹), sendo também uma técnica tolerante à heterogeneidade da amostra (BORCH et al., 2005). Nesta abordagem, proteínas e/ou seus complexos são estabilizados covalentemente através de reação com agentes de ligação cruzada (ALC ou *cross-linkers*), normalmente bifuncionais. Após digestão enzimática, os peptídeos com ALC são identificados por experimentos de MS/MS (varredura de íons produtos), gerando informações sobre a distância espacial entre eles. Tais restrições espaciais são subsequentemente usadas para auxiliar na elucidação, por exemplo, do enovelamento das proteínas, da topologia de complexos e até mesmo, no mapeamento da região de interação entre proteínas.

1.1 Complexos Proteicos

Proteínas são macromoléculas biológicas abundantes, e presentes em todas as partes de uma célula e possuem as mais diversas funções, tais como catálise (enzimas em geral), transporte (desde moléculas simples, como O₂, até outras proteínas), sinalização (receptores), defesa (anticorpos), entre outras. Essas funções, no entanto, nem sempre são desempenhadas por proteínas individualmente, mas sim por complexos proteicos e não-proteicos (*e.g.*, fosfatos, carboidratos, lipídeos etc.). Estas interações dinâmicas entre proteínas e seus ligantes são vitais para a manutenção dos processos biológicos, sendo descritas como o estudo da *sociologia molecular da célula* (BAI et al., 2008). Compreender os parâmetros moleculares que regem esses sistemas é de grande importância, a qual exige além da identificação dos parceiros de interação, também a análise da estequiometria dos complexos, a organização topológica e as regiões de interação e conformações. Pode-se dizer então, que existe na Biologia moderna um grande desafio intrínseco em correlacionar a função de proteínas, com suas respectivas estruturas terciárias e quaternárias (BENESCH et al., 2007). Na Figura 1.1 é mostrado um exemplo de evolução de complexos proteicos, onde há uma estrutura esquemática do complexo NADH ubiquinona redutase e de suas proteínas constituintes em alguns tipos de organismos ao longo do estágio evolucionário.

¹ femtomol: um bilionésimo de um milionésimo (10⁻¹⁵) de um mol

mais de cento e trinta mil interações binárias entre proteínas dentro de uma célula humana (DROIT et al., 2005). Portanto, nos últimos anos, tem se buscado mapear interações moleculares, ou o interatoma, nos mais variados organismos. Espécies como *Drosophila melanogaster* (BOEHR; WRIGHT, 2008), *Saccharomyces cerevisiae* (DYSON; WRIGHT, 2005), vírus da influenza A (ROBINSON et al., 2007) e *Homo sapiens* (KESKIN et al., 2008) tiveram suas redes de interação proteína-proteína publicadas. Há estudos que propõem que a complexidade de um organismo seja proporcional ao tamanho do seu interatoma, e não pelo número de proteínas contidas nele.

A importância funcional do interatoma é tão grande que estudos propõem que a complexidade de um organismo não é informada pelo seu número de proteínas, mas pelo tamanho de seu interatoma (CLARKE et al., 2011). A Figura 1.2 mostra o interatoma de *Homo sapiens*, contendo mais de 400 proteínas ligadas através de 911 interações.

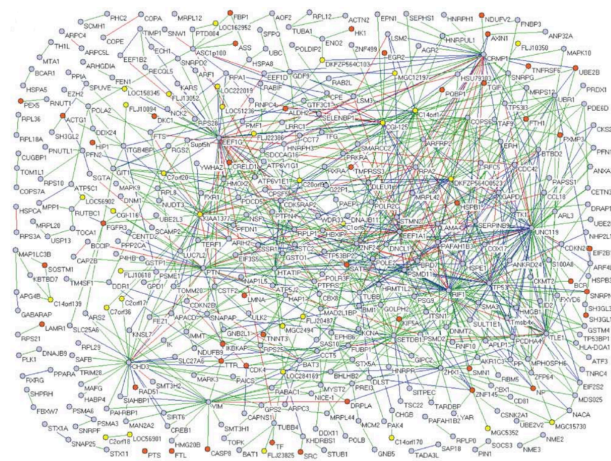


Figura 1.2 – Representação esquemática de uma rede de interação de *H. sapiens* envolvendo 401 proteínas ligadas através de 911 interações. (STELZL et al., 2005)

1.2 Proteômica

A proteômica é uma ciência multidisciplinar que une principalmente a biologia, química e a ciência da computação. Ela compreende o estudo em larga escala das proteínas presentes em um organismo, célula, tecido ou fluido biológico, entre outros, e como essas regem às mudanças temporais e a estímulos, permitindo assim, a identificação e quantificação de complexos proteicos através de instrumentos de alta resolução (EIDHAMMER et al., 2007). Dessa forma, estudos médicos, biomédicos e biotecnológicos vem sendo desenvolvidos a partir do uso dessa metodologia, a fim de estudar patologias, sistemas biológicos e, por conseguinte, o desenvolvimento de

novas tecnologias. Neste contexto, o termo “proteômica” foi primeiramente adotado na década de 90, fazendo uma analogia com o termo genômica, o qual refere-se ao estudo em larga escala dos genes (JAMES, 1997). À vista disso, em 1994, em sua tese de doutorado, Marc Wilkins cunhou o termo proteoma, fazendo a aglutinação da palavra proteína com genoma, a fim de referir-se ao completo conteúdo proteico expresso pelo genoma (WILKINS et al., 1996). Entretanto, mesmo após a tradução do RNA mensageiro pelo ribossomo, as proteínas podem sofrer modificações, alterando suas características estruturais e, conseqüentemente, a sua função. Essas modificações, por exemplo, que determinam a especificidade proteica, a localização e até que tipo de interação uma determinada proteína pode realizar com outra. Existem centenas de modificações pós-traducionais conhecidas e descritas em banco de dados; entre as mais estudadas pode-se exemplificar: a fosforilação, onde ocorre a adição de um grupo de fosfato (PO_4); a metilação, onde há a substituição de um átomo de hidrogênio por um grupo metil (CH_3); a sulfatação, onde ocorre a adição de uma ponte de sulfato; a formação de pontes dissulfeto, ocorrendo a ligação entre átomos de enxofre (S); a acetilação, onde ocorre a adição de um grupo acetila (CH_3CO); entre outras.

Para que se possa realizar um estudo do proteoma, são necessárias metodologias que possuam alta sensibilidade, reprodutibilidade, rapidez, facilidade e automação, fazendo com que esse tipo de estudo resulte em uma análise de custo relativamente alto. Entretanto, com o desenvolvimento de novas tecnologias para a separação de peptídeos e proteínas, a análise por espectrometria de massas, a quantificação por marcação com isóbaros e a análise dos dados pela bioinformática tem possibilitado o progresso nessa área (AEBERSOLD; MANN, 2003; YATES et al., 2009).

1.2.1 Proteômica estrutural e interação proteica

A proteômica estrutural inclui as análises das proteínas em larga escala. Ela é voltada na elucidação de estruturas proteicas e no auxílio da identificação de funções proteicas. A análise estrutural também auxilia no desenvolvimento de novos fármacos e em estudos de interação entre proteínas. A caracterização das interações ajuda a determinar as funções proteicas e também prover a topologia de complexos. Esse entendimento é alcançado utilizando diferentes tecnologias, tais como, as ditas padrões-ouro: cristalografia de raios-X e ressonância magnética nuclear – RMN; assim como outras tecnologias complementares, tais como SAXS, *cryon*-EM e, mais

recentemente, o uso de *cross-linking* em associação com a espectrometria de massas (RAO et al., 2014).

1.2.2 Espectrometria de Massas (MS) aplicada à Proteômica

A Espectrometria de Massas (MS) é uma das ferramentas analíticas que pode ser usada em estudos relacionados a diversas áreas, tais como biologia, medicina, biotecnologia, entre outras. Essa tecnologia é capaz de mensurar a razão massa/carga (m/z) de moléculas, sendo muito empregada em experimentos proteômicos, uma vez que é possível realizar diversos tipos de análises, como por exemplo, estudar modificações pós-tradicionais de proteínas, perfis de expressão proteicos, interações proteína-proteína, entre outros (FERREIRA et al., 2009).

O espectrômetro de massas é um equipamento que permite determinar massas atômicas com alta precisão a partir da ionização de moléculas para a forma gasosa, separando-as de acordo com a relação m/z . Existem diversos tipos de espectrômetros, mas todos são constituídos por fonte de ionização, um ou mais analisadores de massas, sistema de detecção e análise de dados (CAÑAS et al., 2006), como pode ser visualizado na Figura 1.3.

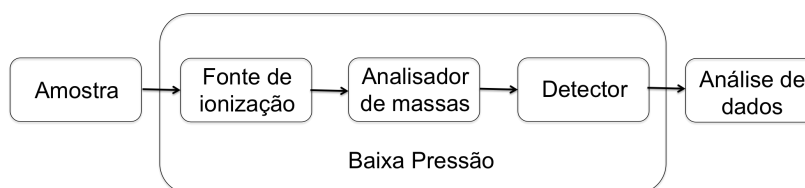


Figura 1.3 – Fluxograma das etapas de um espectrômetro de massas [Figura modificada a partir de (CARVALHO; BARBOSA, 2010)]

No fim da década de 80, com o desenvolvimento de técnicas de ionização suaves, como *Electrospray ionization* – ESI (MASAMICHI; B. FEN, 1983) e ionização por dessorção a *laser* assistida por matriz – MALDI (JUHASZ et al., 1993), a análise de peptídeos e proteínas por MS teve um grande avanço. Tais técnicas, permitem a ionização suave das macromoléculas com alta eficiência (MASAMICHI; B. FEN, 1983). A ionização por *electrospray* (ESI) é muito utilizada em vários instrumentos, produzindo íons a partir de uma solução. Por outro lado, o MALDI é utilizado principalmente com os analisadores de tempo de voo (TOF) e ioniza peptídeos que encontram-se co-cristalizados numa matriz composta de ácido de peso molecular conhecido (HILLENKAMP; KARAS, 1990; HILLENKAMP et al., 1991).

A técnica de ESI ioniza peptídeos que encontram-se em solução. Para isso, aplica-se uma d.d.p., ou seja, uma diferença de potencial, resultando em um campo

elétrico E entre a ponta do capilar e a entrada do espectrômetro de massas, fazendo com que um excesso da carga oposta (geralmente na forma de H^+) seja gerada. Esse excesso de carga na extremidade de um capilar leva a uma repulsão, gerando um fluxo da solução que se vaporiza em gotículas que contêm os peptídeos ionizados, produzindo um fino aerossol (FENN et al., 1989; KEBARLE; VERKERK, 2009). Assim, ao evaporar o solvente, o raio da gotícula diminui, fazendo com que a repulsão eletrostática entre os íons aumente. Quando essa repulsão supera a tensão superficial, ocorre a explosão de Coulomb, e as gotículas, então, se fragmentam em partículas ainda menores em última análise. (FENN et al., 1989), conforme pode ser observado na Figura 1.4. Os peptídeos ionizados, agora em fase gasosa, são direcionados até o analisador de massas, onde eles serão separados de acordo com a razão m/z .

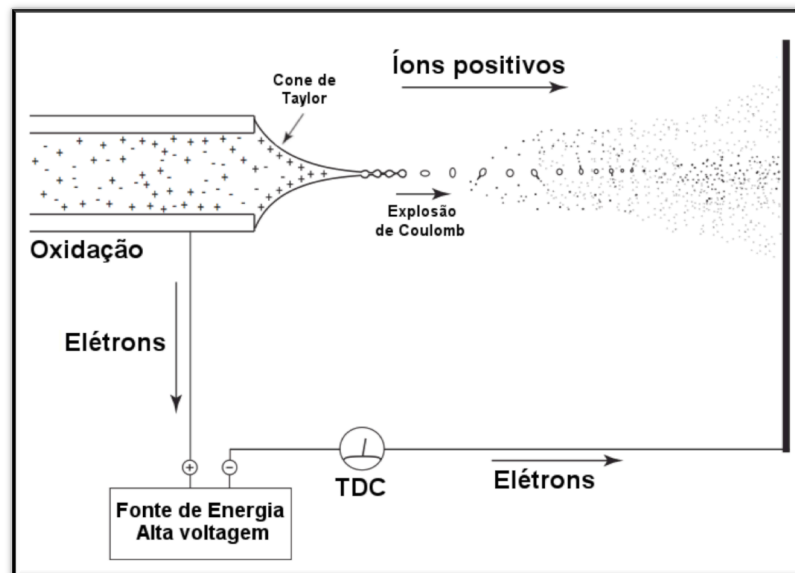


Figura 1.4 – Representação de uma ionização por *electrospray* de partículas em solução. [Figura adaptada de (KEBARLE; VERKERK, 2009)]

No analisador, os peptídeos são separados de acordo com razão m/z . Em seguida, a obtenção dos espectros de varredura de íons produtos, ou MS2, pode ser realizada. Geralmente, os dados podem ser adquiridos usando-se uma metodologia denominada de aquisição dependente de dados, ou *Data Dependent Acquisition* – DDA, no qual os íons precursores mais intensos são isolados e dissociados, geralmente com um gás inerte em uma câmara de colisão; os íons fragmentos resultantes dessas colisões são então analisados. Uma vez fragmentados, esses íons precursores são postos em uma lista de exclusão dinâmica (*dynamic exclusion list*)

por um intervalo de tempo, a fim de não serem analisados novamente, permitindo assim que íons precursores de baixa intensidade possam ser fragmentados.

A fim de facilitar a interpretação dos espectros de varredura de íons produtos de peptídeos, uma nomenclatura foi desenvolvida em 1984 dividindo os íons dos espectros em séries: *a*, *b* e *c* (para íons que pertencem ao lado N-terminal da molécula), e *x*, *y* e *z* (para aqueles que pertençam ao C-terminal) (ROEPSTORFF; FOHLMAN, 1984), que estão representados na Figura 1.5.

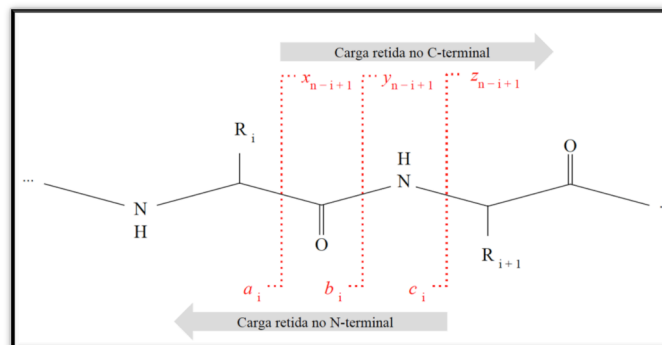


Figura 1.5 – Representação esquemática dos tipos de fragmentação entre os aminoácidos que podem ocorrer através do processo de colisão celular criada por Roepstorff-Fohlmann-Biemann. As linhas em vermelho indicam possíveis regiões de fragmentação entre um determinado aminoácido de um peptídeo de tamanho n . Nos íons a , b e c , a carga fica retida no N-terminal, enquanto que nos íons x , y e z , a carga fica no C-terminal.

Os íons mais frequentes nos espectros de peptídeos fragmentados por dissociação induzida por colisão (CID) são os do tipo b e y . Para interpretar os íons da série b é necessário ler o espectro da esquerda para a direita, ou seja, no sentido N-terminal para o C-terminal, onde a distância entre cada pico representa a massa molecular de um determinado resíduo de aminoácido, conforme demonstrado na Figura 1.6.

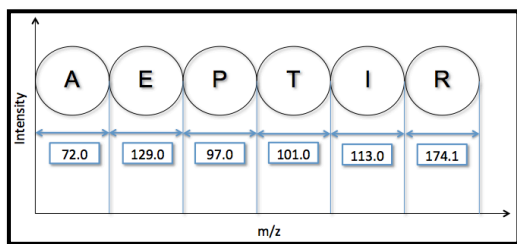


Figura 1.6 – Representação dos íons fragmentos da série *b* em um espectro, representando a sequência peptídica AEPTIR.

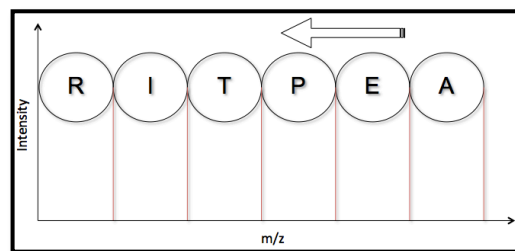


Figura 1.7 – Representação dos íons fragmentos da série *y* em um espectro, formando a sequência AEPTIR. A interpretação desta série é feita no sentido C-terminal para o N-terminal, ou seja, lendo o espectro da direita para a esquerda.

Por outro lado, os íons da série *y* são interpretados realizando a leitura do espectro na ordem inversa, no sentido C-terminal para o N-terminal (Figura 1.7). Entretanto, os íons fragmentos apresentam intensidades diferentes entre um pico e outro, e além disso, no espectro gerado também estão presentes picos oriundos de ruído, fazendo com que a interpretação do espectro seja uma tarefa árdua para o analista. A Figura 1.8 apresenta um espectro com os íons das séries *b* e *y* contendo também vários íons espúrios.

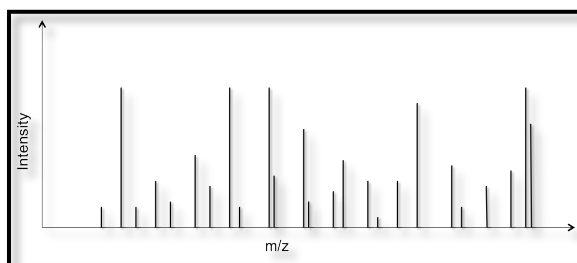


Figura 1.8 – Espectro de massas contendo as séries *b* e *y*, assim como outros picos representado os ruídos.

Finalmente, é no detector, que será analisada a corrente iônica originária da neutralização do íon do analito, através do sinal gerado no espectro de massas, concluindo assim a última etapa deste processo.

Paralelamente, houve uma grande evolução nos analisadores de massa, como aqueles dos tipos armadilhas de íons 3D ou linear, Q-TOF (quadrupolo-tempo de voo), TOF-TOF, FT-ICR (Ressonância Ciclotrônica de Íons por Transformada de Fourier) e *Orbitrap*. Nesse período, foi possível presenciar muitos estudos explorando os diferentes tipos de informações que podem ser obtidas pela análise de peptídeos e

proteínas por MS. O uso da MS para a análise estrutural de proteínas é especialmente atrativo, principalmente devido às suas vantagens intrínsecas, como alta sensibilidade, rapidez de análise e especificidade.

Os primeiros trabalhos mostravam que, em muitos casos, interações proteína-ligante não covalentes podiam ser mantidas em fase gasosa, desde que o processo de ionização fosse realizado em condições controladas (HECK; VAN DEN HEUVEL, 2004). Apesar dos primeiros resultados serem surpreendentes, esse método sofria grandes limitações instrumentais, uma vez que complexos maiores tendiam a se dissociar facilmente, seja por grandes diferenças de pressão às quais são submetidos os íons na análise por MS, seja pela natureza de suas forças de interação (e.g., interações hidrofóbicas são desprezíveis/inexistentes em fase gasosa). Além disso, detectores convencionais apresentavam limitações na faixa de maior m/z , dificultando a detecção dessas espécies. Atualmente a MS tem sido utilizada para a determinação da estequiometria de complexos macromoleculares (HECK; VAN DEN HEUVEL, 2004), a determinação da força de ligação das espécies constituintes de complexos (RUSS; LAMPEL, 2005), a análise de mudanças conformacionais devido à ação de ligantes (VAN DEN HEUVEL et al., 2006), a cinética de enovelamento e desenovelamento (LORENZEN et al., 2008) e a determinação da topologia molecular (ROSE et al., 2008). É importante mencionar que em MS também é observado uma abordagem denominada integrativa – a qual agrega várias técnicas diferentes de modo a compreender o modo de ação entre diversas proteínas que desempenham um determinado papel biológico, uma vez que as principais técnicas aplicadas à proteômica estrutural por si só não são capazes de fornecer todas as informações necessárias acerca de estruturas secundárias, terciárias e sítios de interação.

1.3 Ligação covalente de peptídeos para auxiliar na determinação de interação proteína-proteína e a estrutura proteica por espectrometria de massas (*cross-linking*)

O fenômeno de ligação cruzada (*cross-linking*) é caracterizado pela união de duas espécies a partir da formação de uma ligação covalente. Essas espécies podem representar diferentes classes químicas, como por exemplo, proteínas, ácidos nucleicos ou até mesmo partículas sólidas (WONG, 1993). Os agentes de ligação cruzada, ou ALC, são compostos orgânicos multifuncionais, contendo em geral dois ou três grupos reativos e unidos por uma cadeia espaçadora. Quando em contato com

proteínas, esses compostos são capazes de reagir com cadeias laterais dos aminoácidos, de acordo com suas especificidades (SINZ, 2006). Os ALC reagem com as cadeias laterais de dois resíduos que estejam espacialmente separados, no máximo, pela distância da cadeia espaçadora. Eles podem ser do tipo homobifuncional, onde as extremidades são compostas por grupos funcionais idênticos, como por exemplo, o DSS – *Disuccinimidyl Suberate*, conforme visto na Figura 1.9, ou heterobifuncional, onde as extremidades contém grupos funcionais diferentes, como pode ser visualizado na Figura 1.10.

Muitos reagentes de diferentes *cross-linkers* podem ser sintetizados pela incorporação de dois ou mais grupos reativos em uma determinada molécula. Quando combinados com tipos e/ou tamanhos diferentes, interpretados como espaçadores, uma vez que eles definem uma distância fixa entre os terminais reativos, o número de possíveis compostos de ALC é enorme. Na Tabela 1.1 são demonstradas os grupos reativos e seus respectivos agentes de ligação cruzada.

Tabela 1.1 – Grupos reativos e os respectivos agentes de ligação cruzada que reagem com eles

Grupo reativo	Reagentes de XL
Aminas	NHS ester, Imidoester, Éster Pentafluorofenila, Fosfina de Hidroximetila
Carboxilas com amina	Carbodiimida (e.g., EDC)
Sulfidrilas	Maleimida, Dissulfeto de piridina, Haloacetila (Bromo- or Iodo-)
Aldeídos	Alcoxiaminas, Hidrazidas
Fosfatos	Diazirine, Azidas aromáticas
Hidroxilas	Isocianato

Atualmente, os agentes de ligação cruzada mais utilizados são aqueles compostos por ésteres de NHS, por exemplo o DSS, e os derivados solúveis sulfo-NHS (e.g. BS³). Eles normalmente reagem com grupos aminas primárias, ou seja, com N-terminal e ϵ -amino de lisina, e com menos frequência, com cadeias laterais contendo hidroxila (principalmente serina), tendo uma certa facilidade de ser hidrolisado em meio aquoso. Para que a ligação covalente ocorra entre dois resíduos específicos, eles devem estar espacialmente próximos por um tempo suficientemente longo. Por estarem em grande maioria na amostra biológica e expostos na superfície da proteína, os resíduos de lisina constituem bons alvos nas reações de *cross-linking*

(LEITNER et al., 2010). A fim de mapear regiões de contato entre duas proteínas, agentes de *cross-linking* que catalisam a reação entre dois grupos espacialmente próximos ($\leq 3 \text{ \AA}$) podem ser utilizados sem que haja a presença de cadeias espaçadoras na estrutura. São os chamados *zero-length* (e.g. carbodiimidas) – Figura 1.11, os quais normalmente catalisam reações entre grupos amino e carboxila (MERKLEY et al., 2013).

Um grande progresso no desenvolvimento da técnica de ligação cruzada foi o acoplamento com a MS, pois essa permite ionizar e fragmentar espécies moleculares contendo a ligação cruzada e, conseqüentemente, obter a informação estrutural desejada. Isso é feito realizando a reação de ligação cruzada e em seguida, a digestão proteica do sistema-alvo, com uma enzima proteolítica. Os peptídeos resultantes podem ser então identificados por MS. Como o comprimento do *linker* ($0\text{-}20 \text{ \AA}^2$) é conhecido, assim como a especificidade da reação, então é possível determinar a partir dos experimentos, as distâncias máximas entre diferentes resíduos de aminoácidos e, conseqüentemente, quais deles estavam espacialmente próximos na estrutura nativa da proteína ou do complexo proteico. Os *cross-links* observados podem ser classificados em dois tipos diferentes: intramoleculares, os quais são aqueles sugestivos ao enovelamento da proteína, e intermoleculares, que indicam a presença de interação entre duas moléculas diferentes (SINZ, 2014).

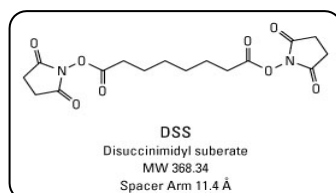


Figura 1.9 – Agente de ligação cruzada do tipo homobifuncional (DSS), o qual possui grupos amino reativos NHS-éster idênticos em ambos os lados da cadeia espaçadora. Essa, por sua vez, tem tamanho de 11.4 Å, que é a distância máxima entre duas moléculas conjugadas.[Figura adaptada de (“DSS (disuccinimidyl suberate)”)]

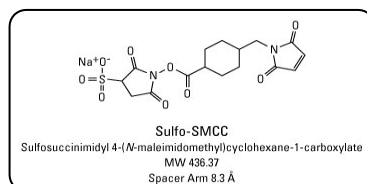


Figura 1.10 – Agente de ligação cruzada do tipo heterobifuncional (Sulfo-SMCC), o qual possui no lado esquerdo da cadeia espaçadora, um grupo amino-reativo, sulfo-NHS-

² Å: *ångström* é uma unidade de medida de comprimento que tem a seguinte relação com o metro: $1 \text{ \AA} = 10^{-10} \text{ m}$

éster, e do outro lado um grupo reativo *sulphydryl maleimide*. [Figura adaptada de (“Sulfo-SMCC (sulfosuccinimidyl 4-(N-maleimidomethyl)cyclohexane-1-carboxylate)”)]

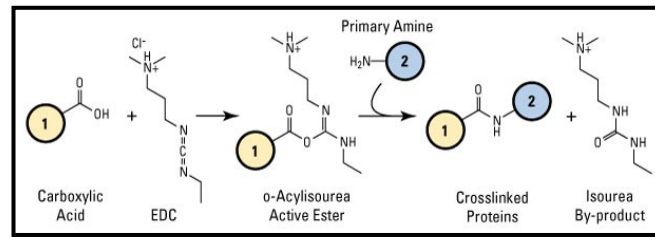


Figura 1.11 – Agente de ligação cruzada caracterizada como *zero-length*, o qual faz a ligação entre duas proteínas sem que haja uma cadeia espaçadora entre elas. [Figura adaptada de (“EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride)”)]

Após a digestão proteica, a amostra de experimentos de ligação covalente costuma ter uma grande complexidade, fazendo com que a interpretação dos dados seja de difícil interpretação manual. A maior parte dos peptídeos oriundos da digestão não sofre qualquer modificação química. Além disso, eles não fornecem quaisquer informações estruturais, aumentando também o intervalo dinâmico das amostras. A reação de *cross-linking* pode originar três tipos de produtos diferentes (Figura 1.12):

- ✓ *Dead-end, mono-link* ou tipo “0”: o peptídeo foi modificado com um *linker* hidrolisado em uma das extremidades. Esse tipo de reação é muito comum e, por mais que não imponham restrições espaciais, podem indicar o grau de acessibilidade dos grupos reativos
- ✓ *Intralink, loop-link* ou tipo “1”: o mesmo peptídeo reagiu com *linker* em dois resíduos diferentes.
- ✓ *Interlink, cross-link* ou tipo “2”: dois peptídeos foram interligados covalentemente com um *linker*. Neste caso específico, denomina-se a “cadeia α ”, aquela a qual contém o maior número de resíduos, enquanto aquela que contém o menor número, é denominada β .

Mais de uma modificação pode ocorrer ao mesmo tempo nos peptídeos, fazendo com que o grau de complexidade das análises aumente (SINZ, 2003).

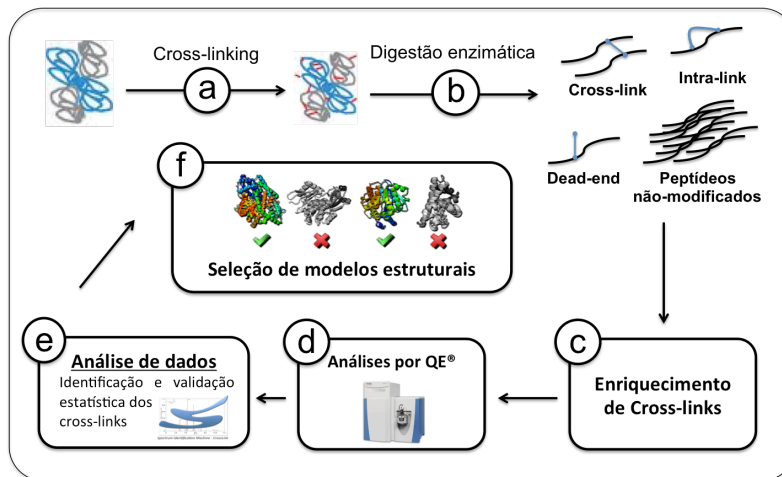


Figura 1.12 – Esquema de um experimento típico de ligação cruzada acoplada à MS. Uma amostra proteica é submetida à ação dos *cross-linkers*, onde após a digestão enzimática, três tipos de produtos podem ser gerados: *cross-link*, *loop-link* ou *mono-link*. Na etapa seguinte, ocorre um enriquecimento dos *cross-links*, como por exemplo, uma purificação; a amostra é injetada em um espectrômetro de massas de alta resolução, onde os dados brutos serão gerados e analisados posteriormente pela ferramenta computacional SIM-XL, podendo assim caracterizar o modelo estrutural da proteína. [Figura adaptada de (LEITNER et al., 2014) e <http://daltonlab.iqm.unicamp.br/research.html>]

Outra análise possível, envolve a identificação do entorno dos sítios específicos de interação, de forma a obter um nível superior de informação. A distância das cadeias espaçadoras define as restrições espaciais do conjunto de moléculas, o que permitiria obter, em princípio, estruturas mais confiáveis por meio de modelagem molecular. Dentre os desafios ainda a serem vencidos nesta técnica, temos a dificuldade de identificação e distinção dos peptídeos covalentemente ligados dentre os peptídeos lineares (que não sofreram ligação cruzada).

Uma metodologia bastante sensível utilizada para caracterizar estruturas proteicas é a análise das amostras de *cross-linking* por nano cromatografia líquida em coluna de fase reversa acoplada a um espectrômetro de massas com fonte nano *electrospray* (nESI). Os instrumentos do tipo *Orbitrap* (ZUBAREV; MAKAROV, 2013) têm sido cada vez mais utilizados nestes tipos de análises por permitirem a aquisição do perfil de espectro de massas e o espectro de massas em tandem em alta resolução, aumentando a confiabilidade dos resultados. Utilizando essa metodologia, várias estruturas de baixa resolução foram modeladas, com vários graus de complexidade. Foram caracterizadas estruturas de proteínas até então não resolvidas por métodos cristalográficos, como é o caso da apolipoproteína A-I humana (SILVA et al., 2005), além de estruturas de homodímeros (protease humana C1r) (LACROIX

et al., 1997), heterodímeros (gp43/gp45 da holoenzima DNA-polimerase do bacteriófago T4, Ffh/Fts Y de *Escherichia coli*) (CHU et al., 2004) e de grandes complexos proteicos, como o composto pela RNA polimerase II (Pol II) e o fator de iniciação TFIIIF, totalizando quinze subunidades e aproximadamente 670 kDa em massa (CHEN et al., 2010).

Desta forma, a ligação cruzada acoplada à MS é uma técnica de grande potencial para o estudo de estruturas superiores de proteínas, o que é particularmente interessante em função do grande número de proteínas que não são passíveis de análise pelas técnicas tradicionais de cristalografia de raios X e RMN. Mesmo que os dados estruturais gerados por MS não atijam a mesma precisão (*i.e.*, $< 3 \text{ \AA}$), as vantagens intrínsecas do método, como sensibilidade, rapidez e aplicação a qualquer proteína, o tornam muito atraente para se estudar sistemas proteicos de interesse.

1.4 Algoritmos para obtenção da sequência de peptídeos analisados por espectrometria de massas

Na proteômica, existem algumas metodologias canônicas para identificação de espectros oriundos de peptídeos, como o *De novo sequencing* e o *Peptide Spectrum Matching* – PSM.

O *De novo sequencing*, mostrado na Figura 1.13, é a metodologia utilizada quando não se possui o sequenciamento do organismo em questão, logo não há um banco de dados de sequências proteicas. Esta técnica é a mais propícia a erros por gerar múltiplas interpretações de um mesmo espectro. Resumidamente, a técnica produz um grafo, cujos nós correspondem a m/z 's de picos espectrais. Uma vez gerado o grafo, um caminho ótimo entre o primeiro e o último nó é traçado a fim de obter a sequência peptídica mais apropriada. Para realizar esse caminho, algumas regras, as quais têm como objetivo pontuar o caminho que possivelmente será a solução, são levadas em consideração a partir das características dos espectros analisados:

- 1) Perda neutra (amônia e/ou água)
 - a. Os íons fragmentos y e b que contém os resíduos aminoácidos R, K, Q e N podem perder uma amônia, representando a perda de 17Da.
 - b. Os íons fragmentos y e b que contém os resíduos aminoácidos S, T e E podem perder uma molécula de água, representando a perda de

18Da. No caso de ácido glutâmico, E, ele deverá estar no N-terminal do fragmento, para que possa ocorrer esta perda neutra.

2) Intensidade espectral

- a. A intensidade do íon b diminuirá quando os resíduos seguintes forem P, G ou H, K e R.
- b. A clivagem interna pode ocorrer nos resíduos P e H. Um fragmento interno é um fragmento que parece ser um peptídeo menor com P e/ou H em seu terminal amino.
- c. É comum os íons $b - y$ e $y - b$ trocarem de intensidade quando um P é encontrada na sequência. Isso também vale quando os resíduos básicos H, K ou R são encontrados na sequência.
- d. Quando uma clivagem ocorre antes ou depois de R, o pico referente a -17 (perda de amônia) pode ficar mais intenso que os correspondentes y e b .
- e. Quando um ácido aspártico aparece na sequência, os íons de uma série podem desaparecer.

3) Composição dos aminoácidos

- a. É possível observar íons imônios na parte baixa do espectro que pode dar uma dica que há um determinado aminoácido naquele peptídeo. Porém, se um íon imônio não é visto no espectro, não significa dizer que o aminoácido correspondente não estará também.

4) Isóbaros

- a. Leucina e Isoleucina são isóbaros e não podem ser diferenciados com uma energia de colisão baixa. Quando esta diferença de massa é observada no espectro, a marcação a ser feita é colocando um X, indicando que ali pode estar presente a Leucina ou a Isoleucina.
- b. Lisina e Glutamina têm massas muito próximas, 128.09496 e 128.05858, respectivamente. O $\Delta mass$ (diferença de massa) é 0.03638, que pode ser visualizado em um espectrômetro de alta resolução. Já nos de baixa resolução, uma acetilação pode ser feita para que a massa da lisina possa sofrer um deslocamento de 42u.

5) O sequenciamento poderá começar pelo último pico mais intenso;

6) Para saber se um peptídeo tríptico termina com Lisina ou Arginina, o primeiro pico da série y ($y1$) será observado, caso o valor for 147, refere-se à Lisina, caso for 175, Arginina;

7) Uma vez que se sabe a massa de um íon b ou y , o seu respectivo poderá ser calculado utilizando a fórmula:

a. $y = (M+H)^{1+} - b + 1$

b. $b = (M+H)^{1+} - y + 1$

Outras regras podem ser visualizadas no site da *IonSource* (“Rules of De Novo Sequencing,” 2015).

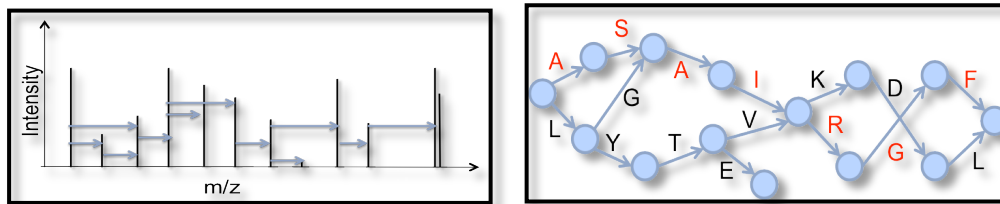


Figura 1.13 – Metodologia de sequenciamento de espectros: *De novo sequencing*.

Outra metodologia usada é o *Peptide Spectrum Matching*, ou PSM. Esta é tida como padrão-ouro da proteômica. Ela é a mais utilizada pelas ferramentas de busca, isso porque é a mais sensível nas identificações quando uma proteína está presente no banco de dados. Nessa metodologia, os espectros experimentais são confrontados contra os teóricos provenientes de um banco de dados. Todavia, é necessário especificar, a priori, as modificações pós-traducionais a serem consideradas.

Existem vários repositórios públicos onde é possível obter o banco de sequências, como por exemplo:

- ✓ *SWISS-PROT*: é um banco de dados de anotações de sequências proteicas. Contém informações adicionais sobre a função proteica, assim como conhecidas modificações pós-traducionais;
- ✓ *TrEMBL*: contém a maioria das traduções das entradas das sequências nucleotídicas que ainda não foram integradas ao *SWISS-PROT*;
- ✓ *PIR-International*: banco de dados de anotações de sequências proteicas;
- ✓ *NCBI nr*: contém sequências de DNA do *GenBank*, *SWISS-PROT* e do *PIR*;

- ✓ *UniProt*: essa é uma nova proposta de banco de dados. Ele junta os bancos SWISS-PROT, TrEMBL e PIR, armazenando sequências proteicas que foram verificadas manualmente.
- ✓ *NeXtProt*: contém informações sobre proteínas humanas, tais como funções, localização subcelular, expressão, interações, entre outras. Atualmente, a maior parte de suas informações são obtidas do *UniProt*.

A Figura 1.14 esquematiza o funcionamento da metodologia PSM.

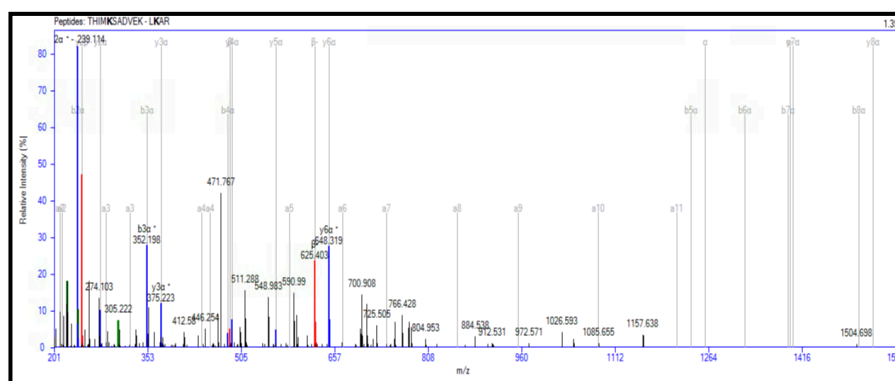


Figura 1.14 – Metodologia PSM (*Peptide Spectrum Matching*): em cinza, ao fundo, está o espectro teórico, originado de uma sequência peptídica contida no banco de dados, enquanto em preto e em colorido está o espectro experimental provido do espectrômetro de massas. Os picos contidos no espectro experimental que também pertencem ao espectro teórico ficam coloridos.

1.5 O funcionamento de uma ferramenta de busca com abordagem PSM

Após a aquisição de uma coleção de espectros de massas provenientes de um equipamento, faz-se necessário processá-la em uma ferramenta de busca proteômica, a fim de obter o sequenciamento dos espectros. Conforme explicado na seção 1.4, existem algumas formas canônicas para realizar a identificação do espectro, e a metodologia proposta nesta tese é a dita padrão-ouro, a *Peptide Spectrum Matching* ou PSM.

Uma ferramenta de busca que utiliza a abordagem PSM possui o seguinte fluxo de trabalho: primeiramente ela é alimentada por uma coleção de espectros provenientes de um instrumento analítico. A partir daí, uma comparação é realizada entre esses espectros, ditos experimentais, e os teóricos, ou seja, aqueles obtidos a partir da digestão *in silico*³ das sequências proteicas presentes no banco de dados,

³ *in silico*: corresponde à simulação computacional do feito ocorrido experimentalmente. Essa expressão originou-se a partir das expressões latinas *in vivo* e *in vitro*, geralmente utilizadas na Biologia (DANCHIN et al., 1991).

levando em consideração uma tolerância previamente especificada. Na ferramenta de busca será verificado qual o espectro experimental possui a maior semelhança com o teórico, de acordo com uma determinada métrica da ferramenta de busca. Aquele que apresentar a maior verossimilhança, será o PSM. A Figura 1.15 apresenta o diagrama de funcionamento de uma ferramenta de busca que utiliza a abordagem PSM.

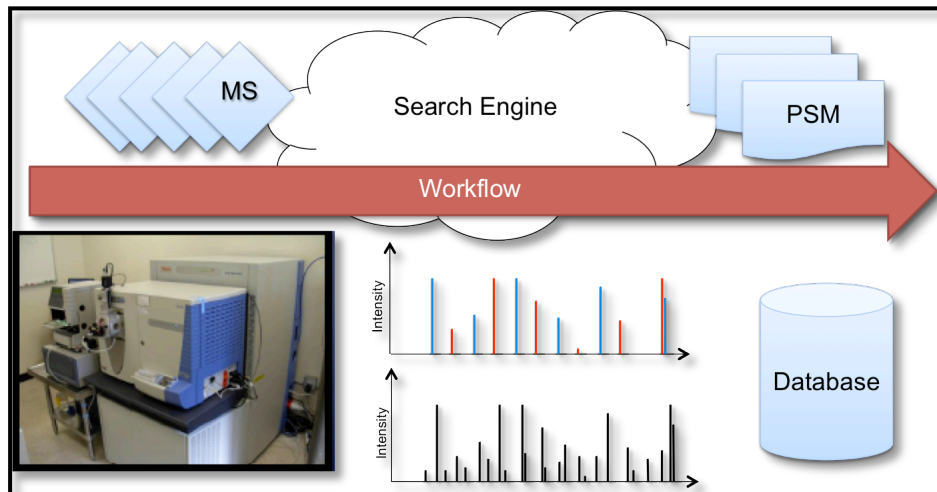


Figura 1.15 – Fluxo de trabalho de uma ferramenta de busca que utiliza a abordagem *Peptide Spectrum Matching* – PSM [Figura modificada a partir de (CARVALHO; BARBOSA, 2010)]

É na ferramenta de busca que está o preditor do espectro teórico, e é onde estão implementados métodos de computação a fim de tornar a geração do resultado mais rápido e mais preciso.

Contudo, os algoritmos existentes para analisar dados de peptídeos covalentemente ligados, como o pLink (YANG et al., 2012) e o Crux (MCILWAIN et al., 2014) foram desenvolvidos para *cross-linking* entre resíduos de lisinas. Apesar da lisina ser um resíduo razoavelmente abundante em proteínas, há a necessidade de se expandir a técnica de *cross-linking* para outros resíduos, de forma que se possa aumentar o número de *cross-links* em proteínas convencionais (resultando no aumento da qualidade dos dados estruturais), assim como permitir o estudo de proteínas pobres em resíduos de lisinas. Outra limitação das ferramentas atuais é a falta de uma interface gráfica para facilitar ao usuário na análise e validação dos dados, o que gera grandes dificuldades para a utilização dos mesmos. Logo, o desenvolvimento de uma ferramenta capaz de identificar peptídeos covalentemente ligados de forma ágil e tendo uma interface gráfica capaz de auxiliar na interpretação dos dados gerados, é fundamental para aumentar o poder preditivo no estudo de

interações proteína-proteína e na caracterização de estruturas proteicas por espectrometria de massas.

1.6 Complexidade do espaço de busca

Os avanços na espectrometria de massas resulta em equipamentos capazes de se aprofundar cada vez mais na proteômica de amostras complexas, devido a maior sensibilidade, resolução e maior velocidade na aquisição de espectros; por conseguinte, cada vez se faz necessário o desenvolvimento de algoritmos mais sofisticados e rápidos. Algoritmos esses que utilizam técnicas de reconhecimento de padrões probabilísticos e inteligência artificial para extrair dos resultados informações importantes para um determinado estudo. Um dos maiores desafios está na quantidade de dados a serem analisados, e também no espaço de busca a ser tratado, já que ele é inversamente proporcional à sensibilidade de identificação da ferramenta de busca, ou seja, quanto maior o espaço de busca, menor é o número de peptídeos e, por conseguinte, de proteínas identificadas (BORGES et al., 2013).

Quando se trata de dados de *cross-linking* a serem analisados, é imprescindível verificar o número de proteínas presentes no banco de sequências, uma vez que quando se procura por peptídeos do tipo “2”, ou os chamados interpeptídeos, o número de combinações aumenta de forma exponencial com o número de sequências no banco, e conseqüentemente, o espaço de busca crescerá proporcionalmente, como se pode verificar na Tabela 1.2.

Tabela 1.2 – Número de peptídeos em diferentes espaços de busca a partir da digestão in silico de proteínas de *E. coli*. O espaço de busca totalmente específico é aquele composto por peptídeos onde a clivagem enzimática ocorre tanto no n-terminal quanto no c-terminal da cadeia peptídica. Já o semi-específico é composto por peptídeos totalmente específicos além de ser composto por aqueles onde a clivagem enzimática ocorre apenas em uma extremidade da cadeia, ou no n-terminal ou no c-terminal. E por último, o espaço não-específico é aquele composto pelos dois espaços anteriores mais os peptídeos clivados em qualquer parte de sua cadeia.

Espaço de busca	Busca tradicional	Combinação 2 a 2
	#peptídeos	#peptídeos
Específico	566.070	160.217.339.415
Semi-específico	11.217.794	62.919.445.504.321
Não-específico	63.102.231	1.990.945.747.037.560

2 Justificativa

Os *softwares* voltados à identificação de peptídeos covalentemente ligados, tais como Crux (MCILWAIN et al., 2014), Stavrox (GÖTZE et al., 2012) e pLink (YANG et al., 2012), são computacionalmente custosos; e apresentam como resultado, arquivos textos que são de difícil análise e validação. Portanto, a maior parte da análise dos dados é voltado à validação manual dos resultados. Adicionalmente, tais ferramentas não apresentam interface gráfica interativa com o resultado; necessário, por exemplo, para a visualização das interações de um complexo proteico. Finalmente, nenhuma ferramenta, até então, era compatível com repositórios públicos de dados proteômicos, como por exemplo, o *PRIDE*.

Este trabalho descreve a primeira ferramenta de busca de peptídeos covalentemente ligados que apresenta como resultado, um mapa de interação bidimensional interativo, mostrando exatamente onde as proteínas identificadas fazem interação. Denominado *Spectrum Identification Machine for Cross-linked peptides* (SIM-XL), ele é capaz de realizar a busca em tempo essencialmente menor, comparado aos das ferramentas mais utilizadas, apresentando também uma maior sensibilidade. O SIM-XL também é a primeira ferramenta de busca compatível com o *PRIDE*, um dos repositórios públicos mais utilizados no mundo para armazenar dados proteômicos. A vantagem principal, é que o usuário pode importar ou exportar dados do repositório para o *software*, tornando-o mais atrativo. A ferramenta computacional também possui a facilidade de exportar todos os *cross-linkings* identificados em formato de tabelas, a fim de facilitar a interpretação dos dados obtidos. Ela também possibilita uma interação maior com o *PyMOL*, objetivando gerar um modelo proteico a partir das identificações realizadas. Finalmente, nossa ferramenta é a única capaz de analisar dados provenientes de homodímeros. Ela faz parte de um ambiente computacional integrado para proteômica denominado *PatternLab for Proteomics*, onde outros módulos, como por exemplo, a proteômica quantitativa, podem eventualmente serem utilizados para interpretações adicionais dos dados.

3 Objetivos

3.1 Objetivo Geral

Desenvolvimento de *software* capaz de identificar peptídeos covalentemente ligados e analisados por espectrometria de massas. O *software* deve ser compatível com *cross-linkers* comercialmente disponíveis e de novos ainda em desenvolvimento.

3.2 Objetivos Específicos

- Criação de um preditor de espectros teóricos para *cross-linkers* atuantes em aminoácidos básicos.
- Criação de um algoritmo para realizar a busca utilizando a abordagem *Peptide Spectrum Matching* (PSM).
- Desenvolvimento de uma interface gráfica fácil de ser utilizada.
- Desenvolvimento de um mapa bidimensional de interação proteica.
- Desenvolvimento de visualizador de espectro dinâmico.

4 Metodologia

Analisar dados provenientes de experimentos de *cross-linking* é uma tarefa computacionalmente difícil. Dentre os diversos motivos, destacamos que a possibilidade de combinações entre peptídeos cresce de forma quadrática de acordo com o número de proteínas presentes no banco de dados. Em contra partida, quanto maior o número de possibilidades a serem consideradas por um algoritmo, maior a probabilidade de erros, culminando em perda de sensibilidade (secção 1.6). Estudos de XL-MS não toleram falsos-positivos porque cada identificação contribui com uma informação estrutural de forma única (KAO et al., 2011); uma identificação errada pode resultar em um modelo falso ou apontar uma interação entre proteínas que não existe. Objetivando superar as limitações, hora apresentamos um algoritmo que demonstrou-se mais eficiente e sensível quando comparado aos existentes.

Tomamos como ponto de partida a ferramenta de busca desenvolvida durante meu mestrado, denominada *Spectrum Identification Machine* – SIM (BORGES et al., 2013), a fim de que ela pudesse processar espectros provenientes de experimentos de *cross-linking*. Para isso, uma nova interface gráfica foi desenvolvida a fim de contemplar parâmetros necessários para estudos voltados a XL-MS.

4.1 Fluxo de trabalho do algoritmo de identificação

Processar dados de *cross-linking* é extremamente custoso do ponto de vista computacional devido ao tamanho do espaço de busca e à quantidade de espectros gerados pelo instrumento analítico. Para que pudéssemos ter um resultado confiável e otimizar o processamento, um fluxo de trabalho foi desenvolvido e dividido nas seguintes etapas:

- Identificar peptídeos modificados pós-traducionalmente, contendo *mono-link*, ou *dead-end*, e também identificar peptídeos que não sofreram nenhuma modificação.
- A partir da identificação dos peptídeos no passo anterior, um novo banco de dados de sequências peptídicas é gerado, contendo apenas os peptídeos com a modificação pós-traducional e os peptídeos provenientes das proteínas identificadas.
- Filtrar os espectros experimentais, provenientes dos espectrômetros de massas, selecionando somente àqueles que são caracterizados pela fragmentação de peptídeos covalentemente ligados.

- Identificar os peptídeos covalentemente ligados através da abordagem *Peptide Spectrum Matching* – PSM.
- Avaliar a qualidade do resultado das identificações
- Auxiliar na modelagem estrutural de proteínas e/ou interações proteína-proteína que podem ser realizadas em um complexo.

4.1.1 Identificação prévia de peptídeos lineares com e sem *mono-links*

Ao estudar dados publicados, observamos que para toda identificação de peptídeos covalentemente ligados, era identificado também um dos peptídeos que foi modificado com um *linker* hidrolisado em uma das extremidades, chamado de *mono-link*, ou *dead-end*.

Considerando essas observações, introduzimos o conceito de busca em *tandem* para XL-MS, a qual, a dividimos em duas etapas: a primeira, objetiva a identificação de peptídeos que não reagiram com os agentes de ligação cruzada (ALC), que são os peptídeos lineares, porém sofreram a modificação *dead-end*. Essa busca tradicional em proteômica, exemplificada na Figura 1.15, é realizada utilizando o *Comet* (ENG et al., 2013), uma ferramenta computacional que utiliza também a abordagem PSM e possui sensibilidade e rapidez na identificação dos dados.

4.1.2 Redução dinâmica do banco de dados

A segunda etapa do conceito de busca em *tandem* consiste em gerar um espaço de busca a partir dos resultados do primeiro passo. Portanto, o algoritmo não mais considera a combinação de todas as possibilidades de peptídeos provenientes de um banco de sequências proteicas, mas agora apenas combinações que possuam ao menos um peptídeo com um *dead-end* identificado na primeira etapa. Essa redução implica em um aumento de sensibilidade e velocidade na busca por peptídeos covalentemente ligados.

Para exemplificar o funcionamento dessa segunda etapa, considera-se que o banco de dados contém apenas a sequência proteica: ABCDE**K**FGHIJKLMNOK**Q**Q**K**RSTKUVWXYZ, onde em negrito estão representados os possíveis sítios de reação do *cross-linker* homobifuncional DSS (Figura 1.9). Ao realizar a digestão *in silico* dessa cadeia polipeptídica com a enzima proteolítica tripsina, serão produzidos vários peptídeos, os quais, alguns exemplos são demonstrados a seguir:

- ✓ ABCDEKFGHIJK
- ✓ FGHIJKLMNOK
- ✓ LMNOKQQKR
- ✓ QQKRSTK
- ✓ STKUVWXYZ

Essa hidrólise *in silico* sob ação da tripsina acontece em resíduos de aminoácidos específicos, como a lisina (K) e a arginina (R), sempre no sentido c-terminal da sequência, o que significa que a proteína será clivada após esses aminoácidos. Entretanto, a ação da enzima nem sempre é eficiente, o que pode ocasionar em falhas de clivagem, ou *missed-cleavages*, e nesse exemplo, foi considerado que pudessem ocorrer até duas falhas como essas.

A Tabela 4.1 ilustra um espaço de busca representativo de um experimento de *cross-linking*, onde todos os peptídeos gerados a partir da digestão *in silico* da sequência proteica, combinam entre si, gerando um espaço que cresce de forma quadrática de acordo o número de peptídeos a serem considerados. Todavia, aplicando uma redução dinâmica, conforme explicado anteriormente, o espaço pode ser reduzido, contemplando apenas peptídeos com *dead-end* e sem modificação pós-traducional identificados na primeira etapa do conceito de busca em *tandem*. Logo, na Tabela 4.2, pode-se notar a redução do número de combinações e conseqüentemente a redução do espaço de busca comparado à Tabela 4.1.

Tabela 4.1 – Espaço de busca típico de um experimento de *cross-linking* onde todos os peptídeos combinam entre si.

	ABCDEKFGHIJK	FGHIJKLMNOK	LMNOKQQKR	QQKRSTK	STKUVWXYZ
ABCDEKFGHIJK	X				
FGHIJKLMNOK	X	X			
LMNOKQQKR	X	X	X		
QQKRSTK	X	X	X	X	
STKUVWXYZ	X	X	X	X	X

Tabela 4.2 – Espaço de busca aplicando a redução dinâmica, onde será composto apenas pela combinação dos peptídeos identificados com *dead-end* – K(DE), e os peptídeos que não sofreram nenhuma modificação pós-traducional.

	ABCDEK(DE)FGHIJK	FGHIJK(DE)LMNOK	LMNOKQQKR	QQKRSTK	STKUVWXYZ
ABCDEKFGHIJK	X				
FGHIJKLMNOK	X	X			
LMNOKQQKR	X	X			
QQKRSTK	X	X			
STKUVWXYZ	X	X			

Dessa forma, o que antes, em um banco de dados ilustrativo de um experimento de *cross-linking* tinha-se um espaço de busca composto por quinze peptídeos, após realizada a redução dinâmica, este número diminuiu para nove; em espaços de busca maiores, esta redução pode resultar em um aumento na sensibilidade das identificações, e em uma redução no tempo de processamento.

4.1.3 Filtragem de espectros característicos de *cross-linking*

A reação de agentes de ligação cruzada em peptídeos permanece durante toda separação cromatográfica, sendo dissociada apenas na etapa de fragmentação no espectrômetro de massas. Um dos efeitos desta dissociação é o aparecimento de íons diagnósticos, ou marcadores (i.e., *reporter ions*) (Figura 4.1), que caracterizam espectros apenas provenientes de peptídeos que sofreram a reação com o *cross-linker*. Exemplos de íons marcadores são: 222,149; 239,1759; 305,2229 para o DSS ou BS³ – Figura 4.2) (IGLESIAS et al., 2010).

O nosso algoritmo toma proveito desta informação, oferecendo como opção a consideração de analisar espectros experimentais que contenham ao menos um íon diagnóstico. Isso resulta em otimização do tempo de busca e redução na possibilidade de identificação de falsos-positivos, ou seja, espectros que não são provenientes de peptídeos covalentemente ligados.

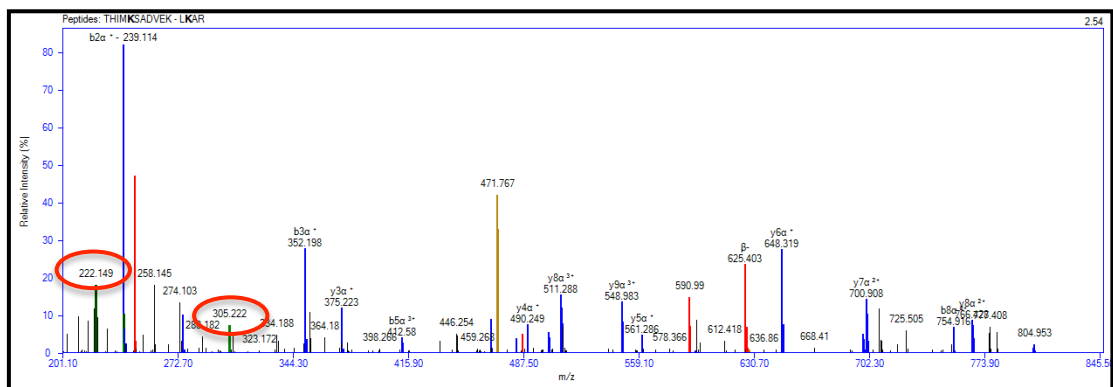


Figura 4.1 – Espectro de fragmentação derivado de peptídeos covalentemente ligados. Em destaque, os íons diagnósticos de m/z 222.149 e 305.222 que caracterizam a existência da reação com um agente de ligação cruzada.

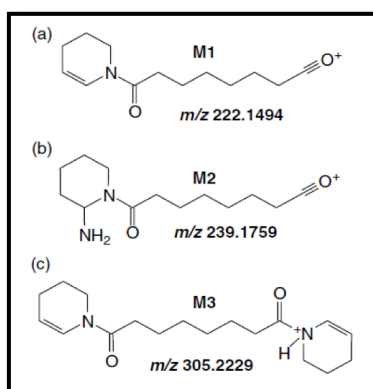


Figura 4.2 – Estrutura molecular dos íons diagnósticos de m/z 222,1494; 239,1759 e 305,2229; os quais caracterizam espectros provenientes de peptídeos que reagiram covalentemente com o DSS. [Figura retirada de (IGLESIAS et al., 2010)]

4.1.4 Identificação dos peptídeos covalentemente ligados

Após a redução dinâmica do espaço de busca e seleção de espectros contendo íons marcadores, o próximo passo consiste na identificação de peptídeos que sofreram a ação dos agentes de ligação cruzada conforme descrito a seguir.

4.1.4.1 Indexação do banco de dados de peptídeos

O algoritmo aqui apresentado fundamenta-se na comparação de espectros experimentais contra teóricos provenientes de um banco de sequências proteicas. A primeira etapa consiste-se em gerar um índice onde, para uma determinada massa, é provido quais sequências proteicas possuem peptídeos com massas próximas. Isso viabiliza o tempo computacional e o gerenciamento das combinações peptídicas na memória do computador. A Tabela 4.3 exemplifica uma indexação ilustrativa; isso permite, por exemplo, encontrar rapidamente quais sequências proteicas contêm peptídeo(s) cuja massa é aproximadamente 784,5298 Da. Em tempo, guardar todos

os peptídeos com suas respectivas massas na memória do computador é inviável, justificando esta forma proposta de indexação.

Tabela 4.3 – Indexação do banco de sequências proteicas a fim de aumentar a velocidade da busca, não desperdiçando tempo em gerar espectros teóricos para peptídeos cuja massa esteja fora do intervalo a ser buscado.

Intervalo de massas	Lista de sequências proteicas
784.4	Seq1, Seq2, Seq3
784.5	Seq4, Seq5
784.6	Seq6, Seq7, Seq8
...	...

4.1.4.2 O algoritmo minhoca

Após a indexação do banco, o próximo passo é identificar sequências peptídicas cujo seu espectro teórico seja semelhante ao espectro experimental em questão. Devido a explosão combinatória de possibilidades durante análises por XL-MS, foi necessário o desenvolvimento de um novo algoritmo denominado “algoritmo minhoca”. Esse é capaz de identificar sequências peptídicas cuja massa teórica esteja igual ao m/z do íon precursor de acordo com uma determinada tolerância. Diferentemente de algoritmos existentes que realizam a digestão *in silico* de sequências proteicas para identificar tais peptídeos, o algoritmo minhoca procura dentro da sequência proteica uma combinação de aminoácidos, representando o peptídeo, o qual sua massa teórica satisfaça a massa do íon precursor.

Uma vez encontrados as combinações de sequências peptídicas satisfazendo as restrições supracitadas, na etapa conseguinte, o algoritmo compara o espectro teórico contra o experimental utilizando uma métrica de distância a ser explicada na 32. Para isso, foi necessário desenvolver um preditor capaz de gerar um espectro teórico a partir de um ou mais peptídeos covalentemente ligados.

O preditor gera espectros teóricos contendo as séries b e y (secção 1.2.2), com as cadeias α e β (secção 1.3). Ele possui duas funções: a primeira gera espectros de peptídeos com *interlink*, e outro de peptídeos com *intralink*.

4.1.4.3 Preditor de espectro teórico para peptídeos com *interlink*

A Figura 4.3 mostra um espectro teórico da combinação peptídica DSSLPPHILEVIDKCGYKEPTPIQR – FGKPLGIR contendo a ligação cruzada no aminoácido lisina (K) na posição quatorze do peptídeo α , com o aminoácido lisina na

posição três do peptídeo β . A série $b\alpha$ que corresponde ao espectro teórico dessa combinação peptídica é composta pelos seguintes íons:

b1α : D = 116,0348	b5α : DSSLP = 500,2356
b2α : DS = 203,0668	...
b3α : DSS = 290,0988	b13α : DSSLPPHILEVID =
b4α : DSSL = 403,1829	1416,7374

Ao chegar na ligação covalente, o valor da relação m/z do íon é igual à massa do peptídeo α até o resíduo correspondente, acrescida da massa total do peptídeo β mais a massa do *cross-linker* (XL). Neste caso, o íon $b14\alpha$ ficaria assim:

$$\mathbf{b14\alpha: DSSLPPHILEVIDK + \text{massa total do peptídeo } \beta + \text{massa do XL} = 1544,8212 + 886,528 + 138,0681 = 2569,4173$$

$$\mathbf{b15\alpha: DSSLPPHILEVIDKC + \text{massa total do peptídeo } \beta + \text{massa do XL} = 2729,47$$

E assim por diante.

De forma análoga, a série β é composta pelos seguintes íons:

$$\mathbf{b1\beta: F = 148,0762}$$

$$\mathbf{b2\beta: FG = 205,0977}$$

$$\mathbf{b3\beta: FGK + \text{massa total do peptídeo } \alpha + \text{massa do XL} = 350,184 + 2874,471735 + 138,0681 = 3362,7458$$

$$\mathbf{b4\beta: FGKP + \text{massa total do peptídeo } \alpha + \text{massa do XL} = 3459,7986$$

$$\mathbf{b5\beta: FGKPL + \text{massa total do peptídeo } \alpha + \text{massa do XL} = 3572,8827$$

E assim por diante.

As séries $y\alpha$ e $y\beta$ são obtidas de forma análoga, porém no sentido C-terminal para o N-terminal (da direita para a esquerda).

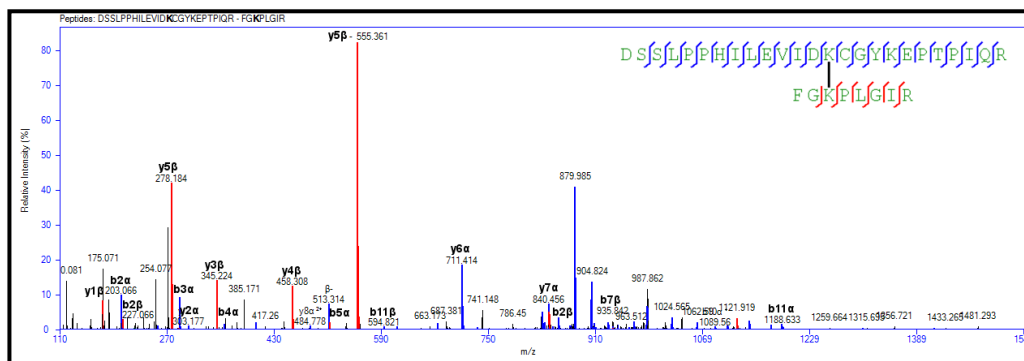


Figura 4.3 – Preditor do espectro teórico para peptídeos com *interlink*. As séries *b* e *y* combinadas com as cadeias α e β são obtidas de maneira a compreender a ligação covalente entre os dois peptídeos.

4.1.4.4 Preditor de espectro teórico para peptídeos com *intralink*

Quando o *linker* reage em um mesmo peptídeo, é caracterizado um *intralink* (secção 1.3). A Figura 4.4 mostra a sobreposição de um espectro teórico contra um experimental proveniente do peptídeo IKKLLKEDISQGVHISVYR, com ligação cruzada ocorrendo entre as lisinas (K) nas posições dois e cinco.

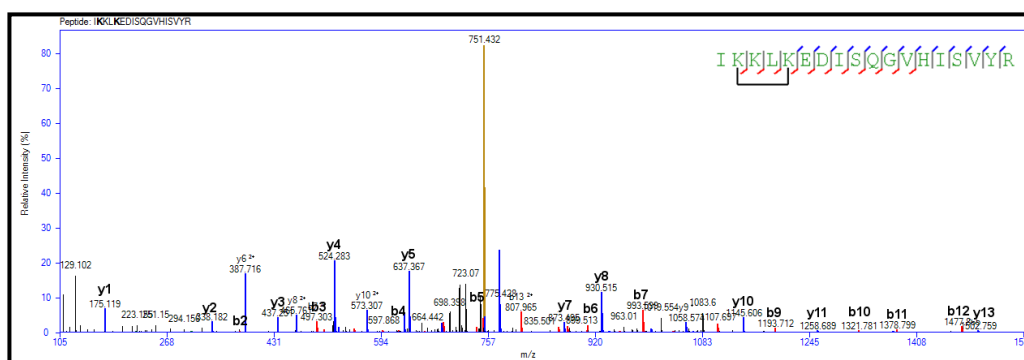


Figura 4.4 – Preditor de espectro teórico para peptídeos com *intralink*. As séries que compõem este espectros são a *b* e *y*.

Neste caso há apenas as séries *b* e *y*, não sendo criadas as combinações com as cadeias α e β . A série *b* é composta pelos seguintes íons:

$$b1: I = 114,0919$$

Ao chegar na reação do agente de ligação cruzada, formando a ligação covalente, a relação *m/z* referente a esse íon será igual à massa do peptídeo linear até o resíduo correspondente acrescida da massa do *cross-linker* (XL). Neste caso o íon *b2* será:

$$b2: IK = I + \text{massa do XL} = 242,1868 + 138,0681 = 380,2549$$

$$b3: IKK = IKK + \text{massa do XL} = 508,3499$$

b4: IKKL = IKKL + massa do XL = 621,434

E assim por diante.

A série y é obtida de forma análoga, porém no sentido C-terminal para o N-terminal. Neste caso ela seria composta pelos seguintes íons:

y1: R = 175,1195

y3: RYV = 437,2512

y2: RY = 338,1828

y4: RYVS = 524,2833

...

y14: RYVSIHVGQSIDEK + massa do XL = 1630,8853 + 138,0681 =
1768,9234

y15: RYVSIHVGQSIDEKL + massa do XL = 1882,0074

E assim por diante.

4.1.4.5 Otimizações no espectro teórico

Nosso grupo determinou regras, de forma probabilística, da existência de intervalos de relação m/z que são mais propícios a conterem íons com determinadas cargas. Por exemplo, peptídeos mono carregados, ou seja, com carga 1^+ , geralmente ocorrem quando a massa é inferior a 2.000 Da. A Figura 4.5 mostra o intervalo m/z para cada carga a ser procurada. Nota-se que, embora existam íons fora do intervalo a ser procurado, eles são de baixíssima quantidade.

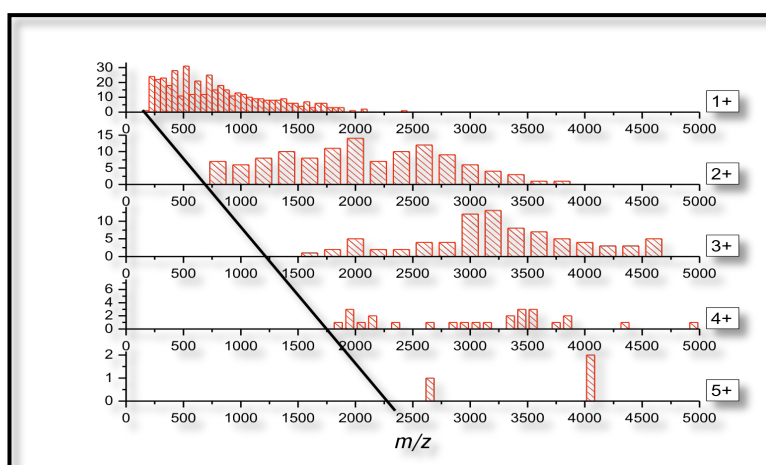


Figura 4.5 – Intervalo de m/z a ser procurado em cada tipo de carga da sequência peptídica. Como exemplo, ao procurar íons de carga 1^+ , o intervalo será [1, 2.000], e dificilmente íons que contêm m/z acima de 2.000 serão mono carregados.

Assim, ao refinar a predição do espectro teórico, torna-o mais limpo, evitando que íons espúrios sejam identificados erroneamente.

4.1.4.6 Métrica de comparação entre espectros teóricos e experimentais.

Uma métrica de verossimilhança é utilizado para avaliar a qualidade dos resultados. A primeira etapa consiste-se no cálculo do produto escalar entre o espectro teórico e o experimental, atribuindo assim um *primary score*, para cada espectro teórico, de acordo com a fórmula a seguir:

$$A \cdot B = \|\vec{A}\| \cdot \|\vec{B}\| \cdot \cos \phi$$

Onde \vec{A} é o vetor que representa os picos com as intensidades do espectro teórico, \vec{B} é o que representa os picos com as intensidades do espectro experimental, e $\cos \phi$ é o ângulo entre os dois vetores, o qual é 0° , dado pelo paralelismo deles. Os vetores são formados a partir da relação m/z dos espectros teórico e experimental. Como as intensidades são normalizadas, então essa primeira parcela do *score* resulta em um valor que pertence ao intervalo $[0, 1]$. Em seguida, a pontuação sofre uma bonificação de 0,1 para cada pico do espectro teórico que está contido no espectro experimental, conforme é mostrado na equação a seguir:

$$\text{score} = \text{produto escalar} + \text{bonificação por cada pico encontrado}$$

Como pode ser observado na Figura 4.6, a pontuação obtida por este espectro é 3,33 (canto superior direito da figura), que foi obtido através do produto escalar, igual a 0,83, acrescido de 2,5, o qual corresponde aos 25 picos encontrados no espectro teórico presentes no espectro experimental (15 picos da série y e 10 da série b). Como cada bonificação equivale a 0,1, então a pontuação final nesse caso é demonstrada a seguir:

$$\text{score} = 0,83 + (25 \text{ picos} \times 0,1) = 0,83 + 2,5 = 3,33$$

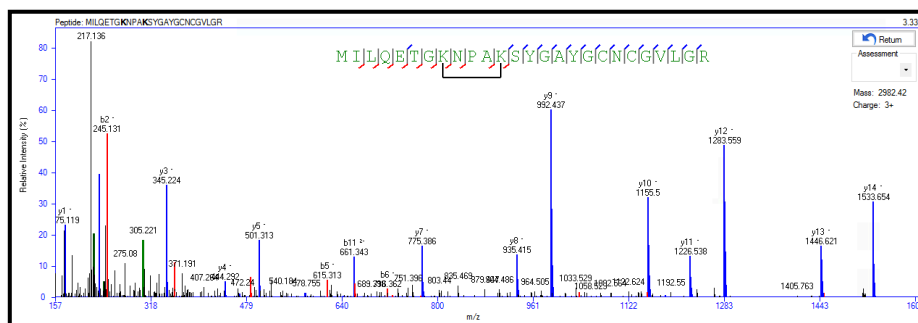


Figura 4.6 – Cálculo do *score*: a pontuação obtida nesse espectro foi de 3.33 (canto superior direito), obtido a partir do produto escalar entre o espectro teórico e o experimental, igual a 0.83, acrescido dos 25 picos (15 da série y , em azul, e 10 da série b , em vermelho) contidos no espectro teórico e presentes no experimental.

4.1.5 Avaliação dos resultados da busca

Após a busca ser realizada, a ferramenta oferece meios para avaliar os resultados. E para isso, a ferramenta disponibiliza a informação de vários fatores descritos a seguir. O primeiro, o *primary score*, reflete o quão semelhante a sequência peptídica (no caso do *intranlink*) ou a combinação peptídica (no caso do *interlink*) é com o espectro experimental. O segundo, reflete o número de vezes que o mesmo *link* aparece nas identificações, chamado de *spectral counting*. Ou seja, em quantos espectros diferentes houve a mesma identificação peptídica. Quanto maior o número observado com um *score* consistente, mais confiável é o *link*. Entretanto, isso depende também, da geração do espectros pelo espectrômetro de massas, isto é, quanto maior a resolução do instrumento analítico, maior será a precisão do espectro experimental gerado. Então, uma única identificação com um *score* relativamente alto, também poderá ser consistente.

Um terceiro fator considerado é o número de resíduos identificados em cada cadeia peptídica (cadeia α e β) – Figura 4.7, quanto maior, mais confiável fica a identificação. Entretanto, outros componentes podem vir a influenciar nesta métrica, como por exemplo, picos espectrais provenientes de ruído e a enzima utilizada. Em geral, a tripsina é a mais utilizada, gerando assim peptídeos de tamanho médio, o que é apropriado para identificação por espectrometria de massas em *tandem* (CANFIELD, 1963). Nestes casos, quando tem-se uma cadeia α muito grande, a tendência é que se tenha uma cadeia β pequena, fazendo com que o número de resíduos sequenciados na segunda cadeia seja baixo.

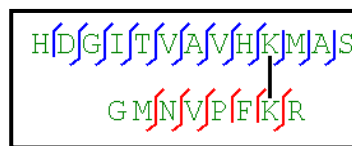


Figura 4.7 – Anotação peptídica: a partir dela é possível observar o número de resíduos em cada cadeia peptídica (α em azul e β em vermelho) que foram identificados.

4.1.5.1 RANSAC

Aqui, empregamos o regressor linear *Random Sample Consensus* – RANSAC (FISCHLER; BOLLES, 1981). Este método é capaz de gerar soluções cada vez mais aproximadas do esperado a medida que se permite um maior número de iterações, tendo como principal vantagem gerar um regressor que seja robusto a *outliers*. Ou seja, a partir de um número mínimo de pontos para instanciar os parâmetros livres,

ditos como p , e um conjunto de pontos totais P , tais que o número de pontos contido em P seja maior que p , seleciona-se arbitrariamente um subconjunto S a partir dos pontos contidos em P . Utiliza-se um modelo já pré-selecionado e instanciado, dito como N , para que se possa gerar SI , selecionando pontos em P que estejam dentro de um erro tolerável. É dito então que SI é um conjunto consenso de S . No caso do algoritmo de busca, o conjunto P é dado pelos íons do espectro experimental analisado, e p é compreendido pelo erro ppm (parte por milhão) de exatidão entre cada íon experimental e teórico de acordo com o produto escalar realizado. Ou seja, cada ponto visualizado na Figura 4.8 representa o erro ppm da relação m/z de cada íon do espectro experimental em relação ao espectro teórico.

O RANSAC é gerado a partir dos dez espectros melhores identificados de uma corrida, traçando um regressor linear, a fim de refletir uma linha de erro médio quadrado dos picos coincidentes entre os espectros teórico e o experimental. Na Figura 4.8 pode-se observar três linhas na cor azul delimitando uma região de erro médio quadrático; o RANSAC é representado pela linha central e as outras duas são obtidas a partir de três desvios-padrão para mais ou menos em relação à linha central. Espera-se que todas as identificações tenham linhas RANSAC bastante semelhantes, mostrando assim, o quão reprodutível e preciso o espectrômetro está.

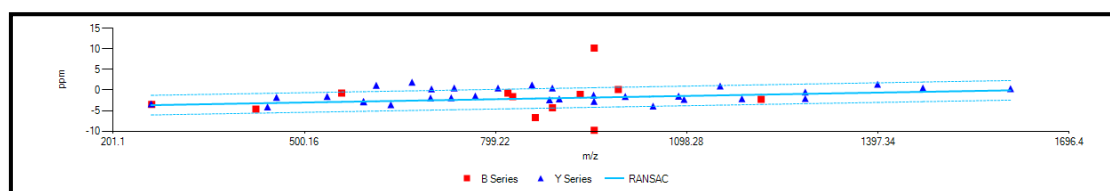


Figura 4.8 – Linhas RANSAC que é interpretada como o erro médio entre os íons do espectro experimental e teórico. Neste exemplo, é possível observar que o erro ppm absoluto é de aproximadamente 5 e que praticamente todos os íons estão dentro da região demarcada pela linha central RANSAC e pelas linhas extremas que correspondem três desvios-padrão para cima e para baixo.

4.1.6 Auxílio na modelagem estrutural de proteínas e interações proteicas

A seguir será descrito como nosso algoritmo auxilia na elaboração de modelo estrutural das proteínas e/ou as interações que elas realizam (Figura 1.12 f). Para isso, uma lista de peptídeos com *cross-linking* é produzido (Figura 4.9) para auxiliar na elaboração de modelos como exemplificado na Figura 4.10.

Scan Number	Measure M+H	Type Peptide Link	Primary Score	ppm	Peaks Matched a	Peaks Matched b	Peptide Sequence	Protein 1	Protein 2	Position of XL Residue 1	Position of XL Residue 2	Liked Spectra
14362	7332.542725	Inter	6.99035	9.099	45	23	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	56	31	
		Inter	6.893110	9.099	44	23	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	56	31	
		Inter	6.896640	9.099	44	23	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	60	31	
		Inter	6.892929	9.099	44	23	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	55	31	
		Inter	6.792005	9.099	44	22	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	56	31	
14907	5470.726290	Inter	6.673788	0.276	52	12	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	60	20	
		Inter	6.062760	0.276	46	12	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	56	20	
		Inter	5.960973	0.276	45	12	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	55	20	
		Inter	5.949355	0.276	45	12	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	63	20	
		Inter	5.829061	0.276	52	4	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	60	26	
14752	5123.497997	Inter	6.659510	4.224	61	3	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	60	64	
13696	5645.824110	Inter	6.110379	0.943	53	6	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	60	103	
		Inter	5.705055	0.943	53	2	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	60	96	
		Inter	5.394750	0.943	46	6	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	55	103	
		Inter	5.381976	0.943	46	6	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	63	103	
10989	4873.393517	Inter	5.986717	0.406	56	2	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	60	11	
		Inter	5.479477	0.406	51	2	TDEGALLSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN2	63	11	

Figura 4.9 – Lista de *cross-linkings* gerada para propor futuros modelos proteicos e/ou interações que ocorrem entre as proteínas.

Uma vez que as restrições de distâncias são verificadas, um *script* pode ser gerado para auxiliar a visualização dos *links* identificados em ferramentas como o *PyMOL* (“<https://www.pymol.org/>”), o qual tem como entrada modelos já pré estabelecidos do *Protein Data Bank – PDB* (BERMAN et al., 2000).

É possível também utilizar as restrições de distâncias para auxiliar na modelagem de estruturas proteicas através do *software RosettaCommons*® (“<https://www.rosettacommons.org/>”).



Figura 4.10 – Visualização pelo *PyMOL* de um modelo de uma proteína e os *cross-linkings* identificados pelo algoritmo de busca.

4.2 *Spectrum Identification Machine for Cross-linked peptides*

Disponibilizar uma ferramenta objetiva e simples para viabilizar identificações de peptídeos covalentemente ligados de uma forma rápida e confiável foi o desafio

proposto nesta tese; aqui materializado como um *software* denominado de *Spectrum Identification Machine for Cross-linked peptides* – SIM-XL (LIMA et al., 2015). O SIM-XL utiliza a abordagem PSM (secção 1.4) e é capaz de disponibilizar um resultado gráfico das interações proteínas-proteínas através de um mapa bidimensional.

Neste secção será detalhado o funcionamento do SIM-XL, mostrando as características e paradigmas utilizados.

4.2.1 Parâmetros

A ferramenta desenvolvida nesta tese permite realizar buscas por peptídeos covalentemente ligados de forma rápida e sensível. Entretanto, para que tal performance seja alcançada, a especificação de alguns parâmetros, intrínsecos do experimento em questão devem ser especificados. A seguir detalharemos os principais parâmetros; fazemos uma ressalva que como parte desta tese publicamos um protocolo (BORGES et al., 2015) a fim de detalhar todos os parâmetros presentes no SIM-XL.

4.2.1.1 Modos de operação

A SIM-XL proverá resultados mais rapidamente quando ativado o modo de redução dinâmica de banco de dados (secção 4.1.2), pois considerará apenas espectros contendo íons diagnósticos (secção 4.1.3) e portanto, processará apenas espectros derivados de peptídeos com ligação cruzada. A partir do momento que ambas as características sejam factíveis, é legítimo realizar a busca no modo *SIM-XL Dynamic DB Reduction with Reporter Ions*. Entretanto, caso os íons diagnósticos (*reporter ions*) não estejam presentes (devido, por exemplo, a utilização de um espectrômetro com analisador do tipo *ion trap*, o qual não gera íons de baixa relação massa/carga), é possível realizar a busca sem o *Dynamic DB Reduction mode*; todavia, o tempo de processamento será maior. E por fim, há o *SIM-XL Normal mode*, onde a ferramenta não realiza o espaço de busca e nem utiliza os íons diagnósticos para otimizar a busca. Na análise de dados, primeiramente, define-se um critério de tolerância de massas para íons precursores, o qual é medido em ppm (parte por milhão), a nível do MS1 (secção 1.2.2), o qual permite com que apenas espectros experimentais, cuja massa do íon precursor esteja dentro de uma tolerância previamente definido, sejam procurados. Adicionalmente, outra tolerância é estabelecida para o espectro de varredura dos íons produtos, ou MS2 (secção 1.2.2), com o objetivo de pontuar picos espectrais

coincidentes entre os espectros teóricos e o experimental (secção 32). Em seguida, o usuário define modificações pós-traducionais (PTMs) a serem consideradas na análise. No SIM-XL há algumas PTMs já pré-definidas, como por exemplo, a “carbamidometilação de cisteína” e a “oxidação da metionina” que implicam em um aumento de massa de 57,02146 Da na Cisteína e 15.9949 Da na Metionina, respectivamente. Essas modificações podem ser consideradas como fixa (i.e., em todos os aminoácidos estabelecidos pelo usuário), ou variáveis. Outro parâmetro pré-definido é a enzima proteolítica tripsina. Esse parâmetro define como a digestão *in silico* será realizada, ou seja, neste caso, “clivando sequências” após os aminoácidos K e R no sentido n-terminal para c-terminal da cadeia polipeptídica.

O banco de dados que contém as sequências proteicas deverá ser indicado na ferramenta de busca e também deverá ser selecionado qual agente de ligação cruzada foi utilizado. Outros parâmetros poderão ser customizados, conforme descritos no protocolo de utilização do SIM-XL (BORGES et al., 2015).

Cumprido a etapa de especificação dos parâmetro de busca, o SIM-XL fará uma serialização⁴ dos dados a fim de poder recuperar os parâmetros usados em uma próxima execução do programa. Em seguida, a ferramenta organizará um dicionário, isto é, uma estrutura de dados, na memória do computador, contendo como chaves os caracteres dos aminoácidos e como valores suas respectivas massas monoisotópicas⁵. As massas das modificações pós-traducionais que a proteína poderá sofrer também serão adicionadas ao dicionário. A criação desse dicionário facilitará no momento que a busca pela sequência peptídica é realizada, uma vez que o acesso à informação tem um tempo de complexidade $O(1)$ ⁶. A Tabela 4.4 mostra as massas em Da dos resíduos e das moléculas pré-configurados no *software*:

⁴ Serializar um objeto é colocar os valores nele contidos juntamente com suas propriedades de certa maneira que fiquem em série, daí o nome serial. Dessa forma, um objeto serializado terá os privilégios para que ele possa ser gravado em disco ou mesmo transmitido pela rede.

⁵ Massa monoisotópica corresponde à soma das massas dos átomos de uma molécula utilizando a massa do isótopo mais abundante. Para a grande maioria dos compostos orgânicos, a massa monoisotópica corresponde à massa do isótopo mais abundante.

⁶ Complexidade computacional é um ramo da teoria da computação que se concentra em classificar problemas de acordo com sua dificuldade. Quando o acesso à informação é de forma imediata, sem a necessidade de resolver cálculos aprofundados, dizemos que o tempo é $O(1)$.

Tabela 4.4 – Resíduos e moléculas com suas respectivas massas em Da.

Aminoácido / Molécula	Descrição	Massa
G	Glicina	57,0214637
A	Alanina	71,0371138
S	Serina	87,0320284
P	Prolina	97,0527638
V	Valina	99,0684139
T	Treonina	101,047678
C	Cisteína	103,009185
I	Isoleucina	113,084064
L	Leucina	113,084064
N	Asparagina	114,042927
D	Ácido Aspártico	115,026943
Q	Glutamina	128,058578
K	Lisina	128,094963
E	Ácido Glutâmico	129,042593
M	Metionina	131,040485
H	Histidina	137,058912
F	Fenilalanina	147,068414
U	Selenocisteína	150,95364
R	Arginina	156,101111
X ou J	Leucina ou Isoleucina	113,08406
Y	Tirosina	163,063329
W	Triptofan	186,079313
O	Pirrolisina - O 22º aminoácido	255,166692
H	Hidrogênio	1,007825032
O	Oxigênio	15,99491462
C	Carbono	12
N	Nitrogênio	14,00307401
NH ₃	Amina	17,0265491
CO	Monóxido de Carbono	27,99491462
H ₂ O	Água	18,01056469
B	Asparagina ou Ácido Aspártico	114,042927
Z	Ácido Glutâmico	128,058578

4.2.2 Linguagem de programação

A esquematização do *software* é parte fundamental para futuras manutenções e aprimoramentos na lógica da programação. Portanto, nós optamos por seguir o paradigma da orientação a objetos, que garante com que o código fonte seja robusto e organizado.

O SIM-XL segue o paradigma *Model-View-Controller* (MVC), que é um padrão bastante difundido na área de desenvolvimento de *softwares*. Tal metodologia permite que a ferramenta seja executada pela linha de comando e através de uma interface gráfica (GUI), facilitando o manuseio do *software* por usuários e também por *clusters* de processamento.

4.2.2.1 *Model-View-Controller* – MVC

Com o objetivo de separar a lógica de negócio da apresentação, e também da necessidade de organizar as linhas de código de sistemas grandes e complexos, foi criado o padrão *Model-View-Controller*, ou simplesmente MVC, permitindo a separação da lógica da aplicação de sua parte gráfica.

Dessa forma, o MVC é compreendido por três camadas (Figura 4.11):

1. **Model:** Responsável por reunir as informações que mostram o estado de um componente, além de informar para seus observadores sobre as mudanças ocorridas nos dados. É no *model* que se gerencia e definem-se as classes de domínio.
2. **View:** É a parte da aplicação que interage com o usuário. É nessa camada que há a ligação e interação com o modelo, especificando como os dados serão apresentados ao usuário.
3. **Controller:** É no *controller* que há o tratamento dos eventos do sistema, ou seja, é nele que as ações do usuário, realizados na camada *view*, serão capturados e processados para que o *model* seja modificado. É aqui que ocorre a validação e a filtragem dos dados realizado pelo usuário.

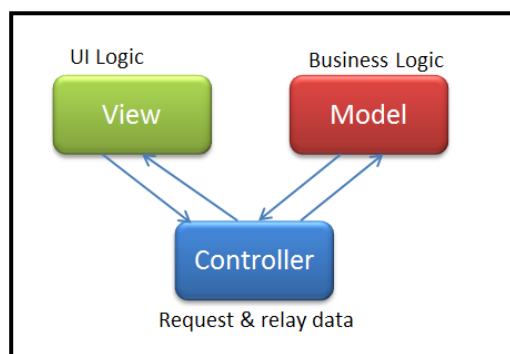


Figura 4.11 – Interação dos componentes do MVC [figura retirada de (APPEL)]

Após a escolha de adoção do MVC, optamos por usar a linguagem de programação C#, pertencente à plataforma .NET, que oferece diversas bibliotecas de auxílio no desenvolvimento do *software*. Essa linguagem é amplamente difundida, facilitando na integração de fóruns de dúvidas, fundamentais no ambiente de desenvolvimento e sendo a linguagem de escolha nos programas desenvolvidos pela Thermo®, uma das principais empresas que vende espectrômetros de massas. Finalmente, o C# dispõem de características únicas, como, por exemplo, o *Language Integrated Query (LINQ)*, que introduz uma série de instruções para manipular dados de forma rápida e simples.

A versão corrente do SIM-XL (v.1.1) possui mais de 900.000 linhas de código implementadas divididas em mais de 119 classes, obedecendo o paradigma MVC.

4.2.3 Interface gráfica

Com o objetivo de tornar a nossa ferramenta e seus resultados de fácil utilização e interpretação elaboramos uma interface gráfica – GUI (Figura 4.12 **Error! Reference source not found.**). A GUI permite a configuração dos parâmetros de busca de forma rápida e simples.

Ao executar o SIM-XL, uma tela é exibida ao usuário onde poderão ser configurados os parâmetros, tais como o arquivo do banco de dados de sequências proteicas no formato FASTA, o diretório onde estão presentes os arquivos oriundos do espectrômetro de massas e também o *cross-linker* utilizado no experimento, dentre outros explicados na seção 4.2.1.

Para cada operação realizada pela nossa ferramenta, são exibidas informações na aba *log*, permitindo o acompanhamento de todo o processo de busca. O *software* também dispõe de uma barra de progresso que permite uma estimativa de quanto tempo falta para que a operação em questão seja finalizada.

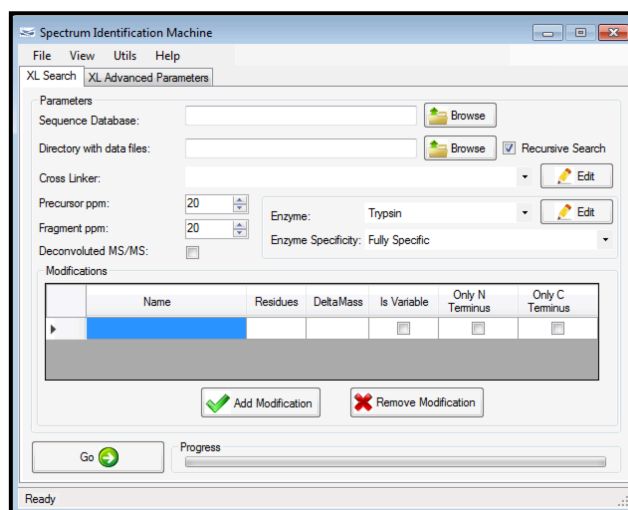


Figura 4.12 – Interface principal do SIM-XL, onde os principais parâmetros de busca são definidos.

4.2.4 Leitura dos espectros experimentais

A busca inicia-se pela leitura de um ou mais arquivos gerados pelo espectrômetro de massas, ou nos formatos *.ms2, *.mzML⁷ e *.mgf⁸. O SIM-XL possui um *parser*⁹ que interpreta esses tipos de arquivos, permitindo a comparação com os espectros teóricos. Para isso, informações como o número do espectro (*scan number*) o tipo de dissociação utilizado (por exemplo, CID, HCD, ETD etc.), assim como massa e estado de carga do íon precursor e de seus íons-fragmentos são interpretadas. Uma vez lidos, os dados experimentais são armazenados na memória para que sejam interpretados no momento da busca.

4.2.5 Geração do arquivo de resultados

Ao término do processamento da busca, os resultados são apresentados de forma gráfica e interativa, facilitando sua interpretação. A interface gráfica permite filtragem automática dos resultados de acordo com critérios de qualidade (secção 4.1.5) e também validação manual, através dos módulos que permitem sobrepor espectros teóricos contra experimentais. Finalmente, é apresentado de forma gráfica *links* que contribuem com informações relacionadas a estrutura da proteína e/ou interação proteína-proteína (secção 4.1.6). Além disso, a fim de evitar perdas da

⁷ O formato mzML é um padrão especificado pela *HUPO Proteomics Standards Initiative (MARTENS et al., 2011)*.

⁸ O formato MGF, ou *Mascot Generic Format*, é um padrão muito utilizado por diversas ferramentas de busca para gravar espectros MS2, e foi oriundo da ferramenta *Mascot* (KOENIG et al., 2008).

⁹ *Parser*, ou analisador sintático, é o responsável por analisar uma sequência de entrada para interpretar sua estrutura gramatical de acordo com um determinada gramática pré-estabelecida.

avaliação dos *links* identificados, os resultados podem ser gravados em arquivos com extensão *.simxlr para futuras avaliações. O arquivo permite recuperar as informações sobre os *links* identificados, o mapa bidimensional, o relatório dinâmico e mapa de calor das interações entre proteínas, descritos a seguir.

4.2.5.1 Mapa bidimensional de interação proteína-proteína

Um característica inédita proposta no SIM-XL está na exibição dos resultados dos *links* na forma de um mapa bidimensional, permitindo observar instantaneamente onde ocorrem as interações proteicas. Esta etapa é preliminar à criação de um modelo de estrutura terciária ou quaternária. Dessa forma, a interface gráfica permite avaliar manualmente cada *link* apresentado e em seguida, atribuir uma qualificação dentre cinco opções seguintes: Excelente, Bom, Médio, Ruim e Péssimo.

Os resultados de *intra-link* e *inter-link* são apresentados de forma separada, já que a pontuação em geral, para os *interlinks* é maior. Isso permite o estabelecimento de *scores* de corte diferenciados para essas classes de resultado.

4.2.5.1.1 Visualizador em barras

O visualizador em barras constitui uma das formas de apresentação dos resultados de forma gráfica. Nesse, proteínas são representadas como barras no mapa bidimensional e os *links* como linhas (Figura 4.13), e é possível reanjar as proteínas de certa a fim de obter uma melhor visualização.

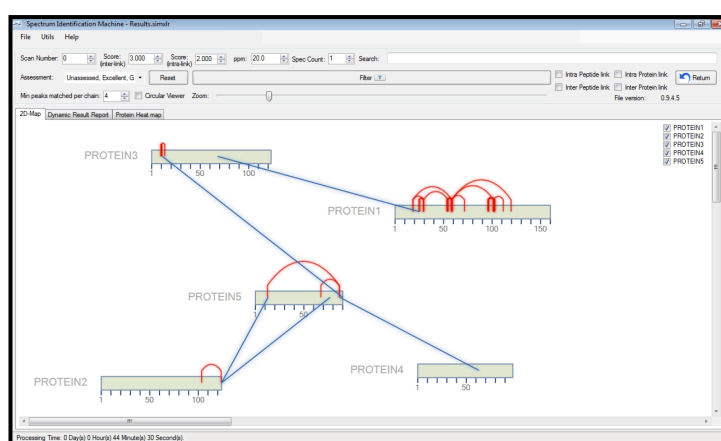


Figura 4.13 – Mapa bidimensional em forma de barras. Essa forma de visualizar os resultados é inédito em ferramentas de busca de *cross-linking*.

4.2.5.1.2 Visualizador circular

Para o caso onde três ou mais proteínas são identificadas, um novo visualizador foi desenvolvido, o circular. Nesses casos, geralmente ocorrem um

número maior de *links*, podendo tornar o resultado de difícil interpretação quando no modo em barras. Nesse visualizador, proteínas são dispostas na forma de arcos, inicialmente formando um círculo. O comprimento de cada arco é diretamente proporcional ao número de aminoácidos contidos na respectiva sequência (Figura 4.14). É possível destacar proteínas do círculo para facilitar a visualização de *intra e interlinks*. Para o caso das proteínas com sequências muito menores que as demais, o visualizador circular permite gerar uma imagem híbrida com o modo em barra para uma melhor interpretação (Figura 4.15).

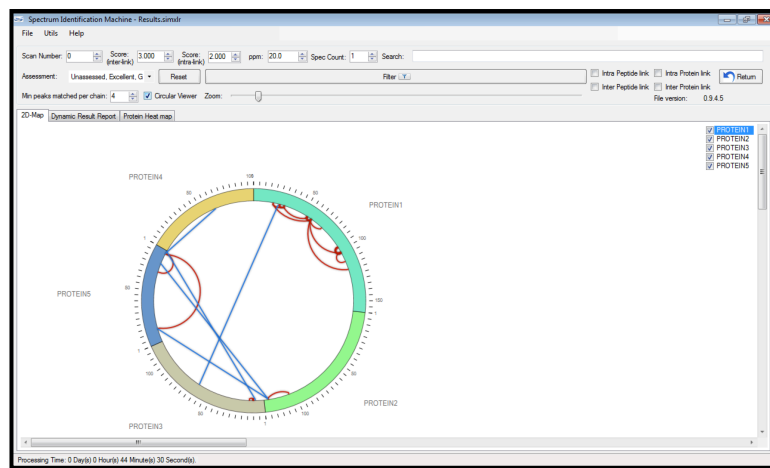


Figura 4.14 – Mapa bidimensional em forma circular. Esse tipo de visualização é factível quando muitas proteínas são identificadas, tornando a interpretação mais clara e objetiva.

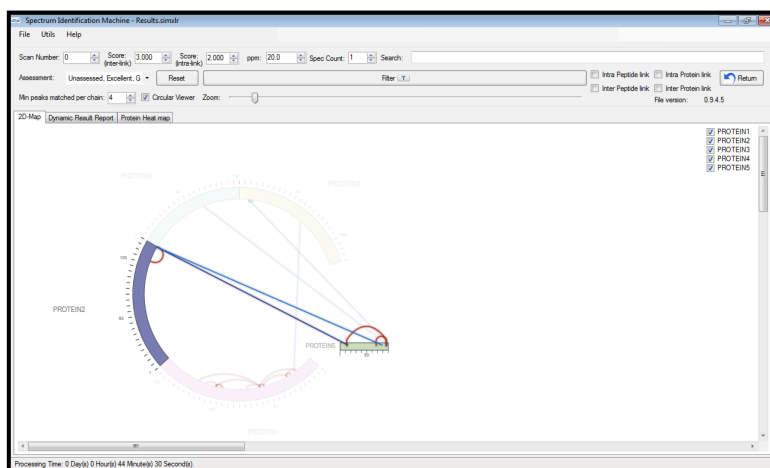
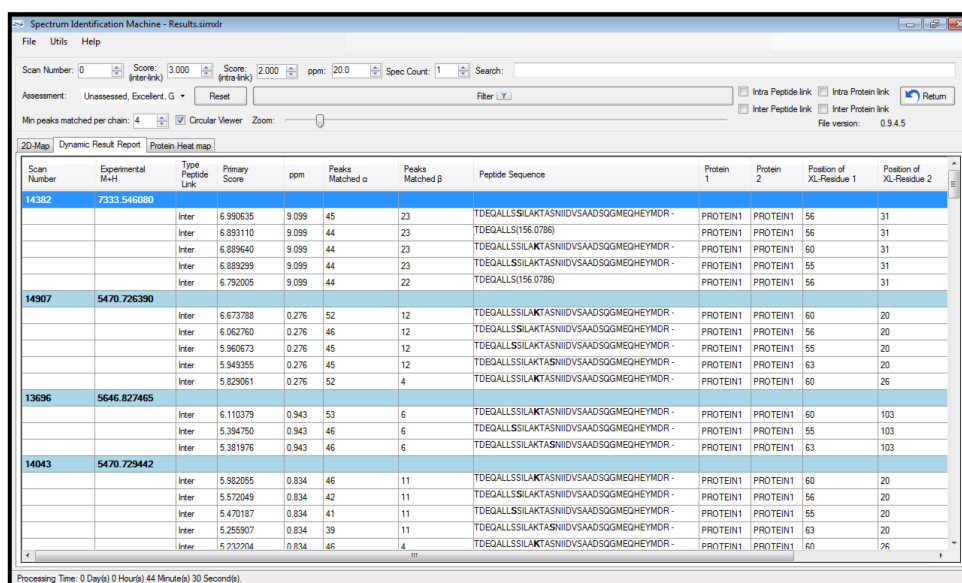


Figura 4.15 – É possível destacar duas proteínas para que se possa visualizar de uma melhor maneira a interação que as duas têm entre si. E pode-se alternar para a visualização em barras uma proteína específica.

4.2.5.1 Relatório dinâmico de resultados

As identificações também são exibidas em forma de relatório. No relatório dinâmico do SIM-XL, cada detalhe da identificação é apresentado, como por exemplo, o número do espectro caracterizando uma determinada ligação covalente, a massa experimental do peptídeo (ou dos peptídeos eluídos, no caso de *interlink*), o *score* atribuído à determinada identificação, o erro ppm entre a sequência teórica e o dado experimental, a quantidade de picos espectrais identificados na cadeia α e β , a sequência peptídica propriamente dita, a(s) proteína(s) que é(são) inferida(s) por essa identificação, a posição do(s) *link*(s) na(s) respectiva(s) sequência(s) proteica(s), dentre outros detalhes. O usuário poderá filtrar, por cada campo, a fim de refinar o relatório, podendo também, visualizar o espectro (secção 4.2.5.4) para poder manualmente validar o resultado proposto. Os filtros aplicados no relatório dinâmico são imediatamente refletidos nas outras partes da interface gráfica, como por exemplo, no visualizador em barras, no circular e até mesmo no mapa de calor de interação. A Figura 4.16 ilustra o relatório dinâmico de um resultado.



Scan Number	Experimental H+H	Type Peptide Link	Primary Score	ppm	Peaks Matched α	Peaks Matched β	Peptide Sequence	Protein 1	Protein 2	Position of XL-Residue 1	Position of XL-Residue 2
14382	7333.546080	Inter	6.990635	9.099	45	23	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	56	31
		Inter	6.893110	9.099	44	23	TDEGALLS(156.0786)	PROTEIN1	PROTEIN1	56	31
		Inter	6.889640	9.099	44	23	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	60	31
		Inter	6.889299	9.099	44	23	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	55	31
		Inter	6.792005	9.099	44	22	TDEGALLS(156.0786)	PROTEIN1	PROTEIN1	56	31
14907	5470.726390	Inter	6.673788	0.276	52	12	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	60	20
		Inter	6.062760	0.276	46	12	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	56	20
		Inter	5.960673	0.276	45	12	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	55	20
		Inter	5.949355	0.276	45	12	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	63	20
		Inter	5.829061	0.276	52	4	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	60	26
13696	5646.827465	Inter	6.110379	0.943	53	6	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	60	103
		Inter	5.394750	0.943	46	6	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	55	103
		Inter	5.381976	0.943	46	6	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	63	103
14043	5470.729442	Inter	5.982055	0.834	46	11	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	60	20
		Inter	5.572049	0.834	42	11	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	56	20
		Inter	5.470187	0.834	41	11	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	55	20
		Inter	5.255907	0.834	39	11	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	63	20
		Inter	5.232204	0.834	46	4	TDEGALLSSILAKTASNIIIVSAADSOGMEGHEYMDR	PROTEIN1	PROTEIN1	60	26

Figura 4.16 – Relatório dinâmico de resultados. Através dele é possível analisar cada identificação detalhadamente.

4.2.5.2 Mapa de calor das interações proteicas

O mapa de calor, é uma outra maneira de visualizar os resultados; aqui as regiões proteicas são coloridas de acordo com o número de identificações de *cross-linkers* atribuídos em cada região. Quanto maior o número de identificações, mais intenso é a coloração vermelha. Na Figura 4.17 pode-se observar que o *link* K(241) –

resíduo lisina na posição 241 da *PROTEINI*, e K(257) – resíduo lisina na posição 257 da mesma proteína, foi identificado em seis espectros diferentes. Por outro lado, o *link* K(225) – K(241) é caracterizado por quarenta espectros distintos. O usuário poderá ter acesso a cada espectro, simplesmente clicando na célula desejada, logo, o número do espectro, assim como o *score* obtido e a referida sequência peptídica serão exibidos em um lista.

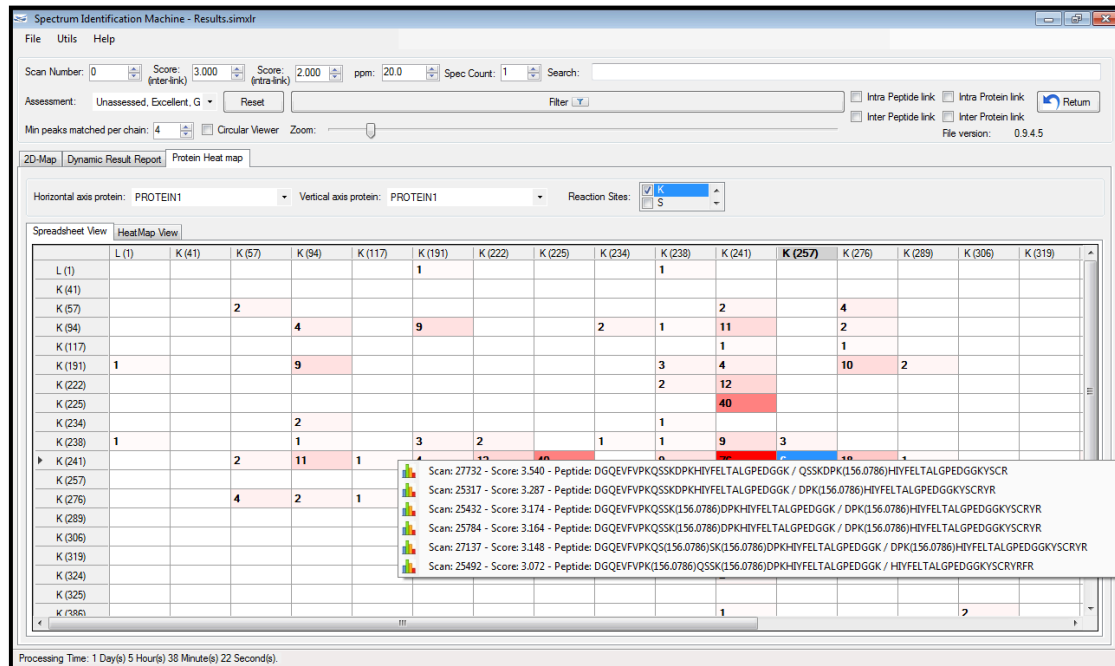


Figura 4.17 – Visualizador do mapa de calor dinâmico. Com ele é possível observar quantas identificações cada *link* obteve, representado por células diferentes. É possível saber também, cada identificação obtida para o *link* apenas clicando sobre a célula desejada.

A Figura 4.18 demonstra um mapa de calor refletindo o número de espectros apenas pela intensidade de cores.

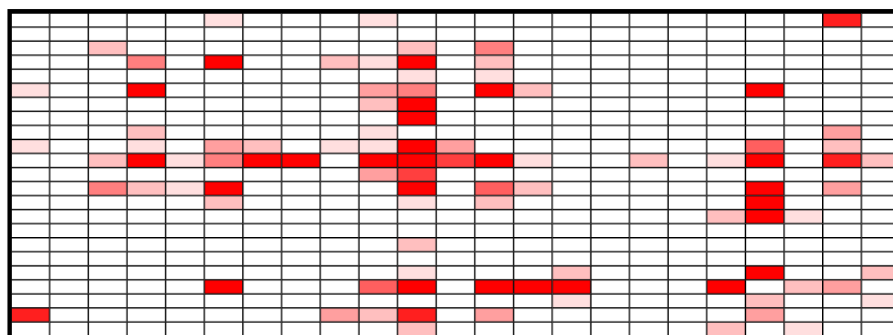


Figura 4.18 – Mapa de calor obtido através das interações entre duas proteínas. É possível verificar, através das intensidades de cores em cada célula do mapa, em quais regiões foram obtidas um maior número de identificações para cada *link* observado.

4.2.5.3 Visualização dos *links* em estruturas terciárias ou quaternárias

Objetivando completar a última etapa do fluxo de trabalho (secção 4.1.6), o SIM-XL permite exportar as informações dos *links* identificados, a fim de que possam ser visualizados na forma tridimensional – 3D (Figura 4.10). Para isso, *scripts* para o *software PyMOL* (“<https://www.pymol.org/>”) são automaticamente gerados através da interface gráfica (Figura 4.19). O SIM-XL permite customizar parâmetros de exibição referente ao modelo proteico.

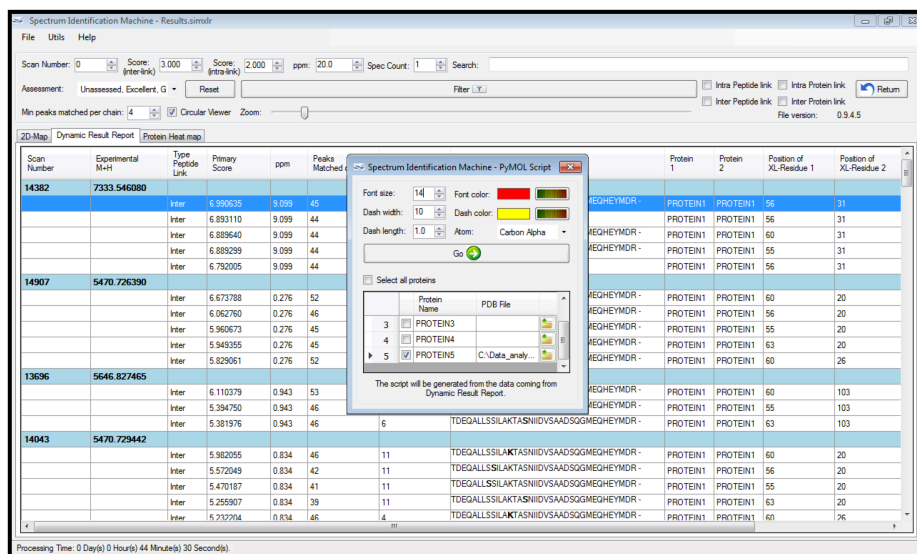


Figura 4.19 – Geração do *script PyMOL*. Com ele é possível criar modelos 3D de estruturas terciárias de proteínas ou quaternárias, no caso de complexos proteicos.

4.2.5.4 Visualizador dinâmico de espectros

Para avaliar as identificações, é possível visualizar a sobreposição do espectro experimental contra o teórico de forma dinâmica a partir do visualizador dinâmico. Ele também permite alterar os parâmetros de sobreposição (secção 1.4) como, por exemplo, a sequência do peptídeo teórico ou até mesmo a tolerância ppm usada durante o pareamento dos espectros. É possível também, ampliar uma região de interesse do espectro para uma avaliação mais detalhada, e caso algum pico espectral foi identificado erroneamente, é possível desmarcá-lo e, logo após, realizar uma nova comparação do espectro para obter um *score* atualizado. A região superior do visualizador exibe um regressor RANSAC (secção 4.1.5.1) mostrando o erro ppm entre os picos experimentais e teóricos, e também a anotação peptídica, mostrando os resíduos que foram identificados (secção 4.1.5). Finalmente, o visualizador permite ao usuário atribuir um grau de qualidade na identificação, no campo *Assessment*, facilitando uma filtragem *a posteriori*, de acordo com esta avaliação pessoal.

É possível também, comparar listas de íons-fragmento de um determinado espectro experimental com uma sequência peptídica (ou duas sequências, no caso do *interlink*). O visualizador dinâmico é capaz de prover um *score* para a referida comparação, permitindo melhorar a avaliação da qualidade do espectro e/ou da(s) sequência(s) fornecida(s).

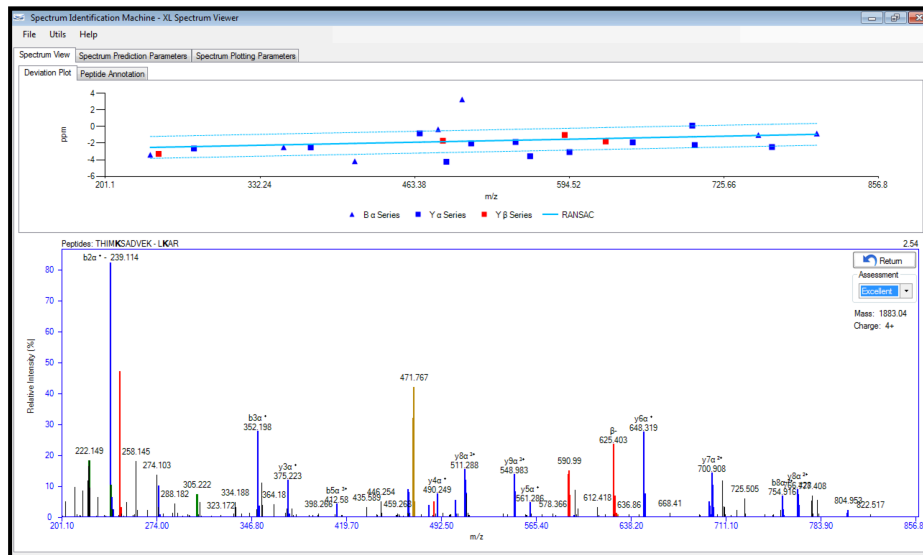


Figura 4.20 – Visualizador de espectro dinâmico. É possível customizar o espectro observado a fim de obter um melhor resultado.

4.3 Repositório de armazenamento de dados – *PRIDE*

A deposição de dados experimentais de espectrometria de massas em repositórios públicos tem se tornado uma prática cada vez mais adotada. Esses repositórios permitem uma maior transparência nas publicações pois outros grupos poderão reproduzir os resultados reportados. Um dos repositórios mais utilizados para estes fins é o *Proteomics Identifications database – PRIDE* (VIZCAÍNO et al., 2013). Ele foi desenvolvido e é mantido pela equipe EMBL-EBI, contendo mais de 2.400 projetos e mais de 42.700 arquivos brutos armazenados.

O SIM-XL é a primeira ferramenta de busca voltada à *cross-linker* compatível com o *PRIDE*, ou seja, o *software* hora apresentado é capaz de gerar arquivos no formato *.mzIdentML (JONES et al., 2012), tornando os resultados de *cross-linkers*, compatíveis com o repositório *online*. Ressaltamos que para isso, tivemos que trabalhar juntamente com a equipe do EMBL para estabelecimento das especificações e regras de padronização de resultados de XL-MS para o *PRIDE*.

4.4 O SIM-XL e o *PatternLab for Proteomics*

O *PatternLab for Proteomics* – PL é um integrado ambiente computacional, que contém ferramentas para realizar análises de proteômica quantitativa, tais como quantificação livre de marcação, ou *label-free*, que utiliza técnicas de *spectral counting* (CARVALHO et al., 2008; LIU et al., 2004) e/ou cromatograma de íons extraídos – XIC (NEILSON et al., 2011), e quantificação relativa e absoluta através de marcadores caracterizados como íons isóbaros – iTRAQ (BONDARENKO et al., 2002; ZHANG et al., 2013) e TMT; análises de proteínas diferencialmente expressas; análises ANOVA (MILONE, 2004); produção de Diagramas de *Venn* para a verificação de proteínas exclusivas em uma amostra; análises do *Gene Ontology*, as quais ajudam no entendimento da significância biológica dos dados; Búzios, o qual permite agrupar perfis proteômicos similares (AQUINO et al., 2014); e também o *XD Scoring system*, que permite avaliar a confiança em sítios de fosforização (FISCHER et al., 2015); um serviço na nuvem, ou *cloud service*, o qual concede fazer uma predição quantitativa em larga escala de domínios proteicos (LEPREVOST et al., 2013); entre outros módulos relacionados à proteômica computacional (Figura 4.21).

O PL também contém módulos para formatar banco de sequências proteicas oriundos de repositório públicos, tais como *Uniprot* (secção 1.4), incluindo sequências *decoys* – ou falsas, o qual permite com que um pós processamento dos resultados através de uma análise estatística seja realizada após a busca por peptídeos utilizando a abordagem PSM tenha sido concluída. Nele é possível também realizar um estudo dirigido à similaridade em dados de genes que ainda não foram sequenciados, utilizando a abordagem *de novo sequencing* (secção 1.4), através do módulo *PepExplorer* (LEPREVOST et al., 2014). Enfim, o *Patternlab for Proteomics* é um amplo ambiente computacional que integra vários campos da proteômica em um único *software*, facilitando a usabilidade e a análise dos dados proteômicos, o que nos permitiu ter aceito para publicação um protocolo em uma das revistas de maiores prestígios na ciência, a *Nature Protocols* (Anexos I – Artigos publicados).

Para divulgação e sendo o principal meio de *download* do *software*, foi desenvolvido um *website* (<http://patternlabforproteomics.org>) contendo as principais informações referente ao PL4.0, assim como o *link* para entrar no fórum de discussão do ambiente computacional proteômica, para que dúvidas pudesse ser solucionadas.

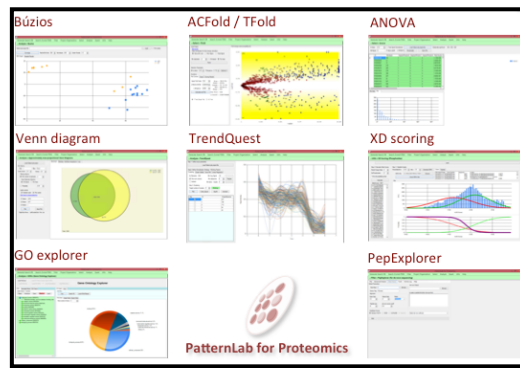


Figura 4.21 – *PatternLab for Proteomics* é um ambiente computacional composto por vários módulos para análise de proteômica qualitativa e quantitativa.

O SIM-XL torna-se um novo módulo estruturante desse ambiente computacional para proteômica que é amplamente difundido internacionalmente. Durante a elaboração da ferramenta de busca, vários módulos do PL foram acoplados e até mesmo aperfeiçoados. Portanto, a elaboração do SIM-XL indiretamente resultou em um impacto na qualidade do *PatternLab*, mesmo em módulos de proteômica quantitativa ou qualitativa que, aparentemente, estariam completamente desvinculados de análises de XL-MS.

5 Resultados

Para comprovar a robustez do SIM-XL optamos em compará-lo com ferramentas de busca de *cross-linking* amplamente utilizadas no âmbito acadêmico. Nossa ferramenta apresentou-se ser mais sensível (conforme pode ser observado na Figura 5.1) e rápido (Figura 5.2) que o Crux e o pLink após experimentação em um conjunto de dados oriundo da proteína de choque térmico HSP90. Comparado ao Crux, nossa ferramenta de busca obteve 50% mais identificações, no modo mais otimizado, e um aumento na performance de 10.000%. Similarmente, a comparação com o pLink, comprovou novamente maior número de identificações em um tempo reduzido de 1.800%.

Analisando um conjunto de dados composto por 1.788 espectros experimentais provenientes de uma amostra, cujo objetivo é estabelecer o modelo estrutural da proteína de choque térmico de alto peso molecular – HSP90, 973 continham pelo menos um íon diagnóstico. Dentre os primeiros 50 espectros analisados pelo SIM-XL no modo normal, ou seja, no modo em que não foram utilizados os íons diagnósticos para selecionar os espectros experimentais, nem foi reduzido dinamicamente o banco de dados, três resultados foram originados de pares de peptídeos *decoys*, ou seja, falsos-positivos, duas das quais apresentaram íons diagnósticos. Neste caso, um peptídeo foi identificado corretamente (ou o peptídeo α ou o β), entretanto o seu par não. Em relação a todas as outras identificações não-*decoys*, quatro não tiveram pelo menos um íon diagnóstico; nós também notamos que dentre todas essas identificações, existe para todo peptídeo covalentemente ligado pelo menos uma versão do peptídeo linear com a modificação pós-traducional *dead-end*, e que somente duas das três identificações de peptídeos *decoys* não possuíam a versão do peptídeo linear com *mono-link*.

Para gerar os dados, resumidamente, utilizou-se o *cross-linker* DSS, o qual foi dissolvido em dimetilformamida (DMF, *Thermo Scientific*), em uma concentração de 27,1 mM. O DSS foi adicionado à proteína HSP90 na extremidade c-terminal a uma relação de 1:50 e incubado com a amostra por 2h em temperatura ambiente. A reação de *cross-linking* foi finalizada com bicarbonato de amônio 100 mM. A redução e a alquilação dos resíduos de cisteína foram realizadas usando ditioneitol e iodacetamida durante 30 min à 60°C e em temperatura ambiente, respectivamente. A amostra foi digerida com tripsina a uma proporção de 1:50 durante 16h à 37°C. Os peptídeos

foram fracionados usando um *Oasis HLB cartridge* (Waters Corp.) e eluídos a uma concentração diferente de acetonitrila, e analisados usando um espectrômetro de massas *Q-Exactive* da Thermo® equipado com uma nano cromatografia líquida em coluna de fase reversa e uma fonte nano *electrospray* (nESI – Thermo, San Jose – CA – EUA).

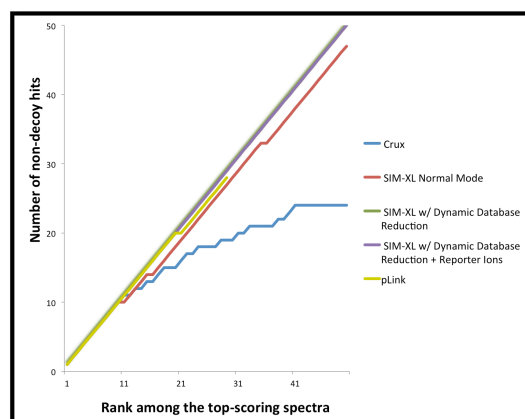


Figura 5.1 – Identificação de espectros de acordo com cada ferramenta analisada. O Crux foi a ferramenta que apresentou a menor eficiência, identificando corretamente apenas 50% dos espectros apresentados; logo em seguida, o pLink não apresentou resultados errôneos, entretanto, só conseguiu identificar 29 dos 50 espectros apresentados. A ferramenta desenvolvida nesse trabalho, o SIM-XL, conseguiu identificar 100% dos espectros apresentados no modo mais otimizado.

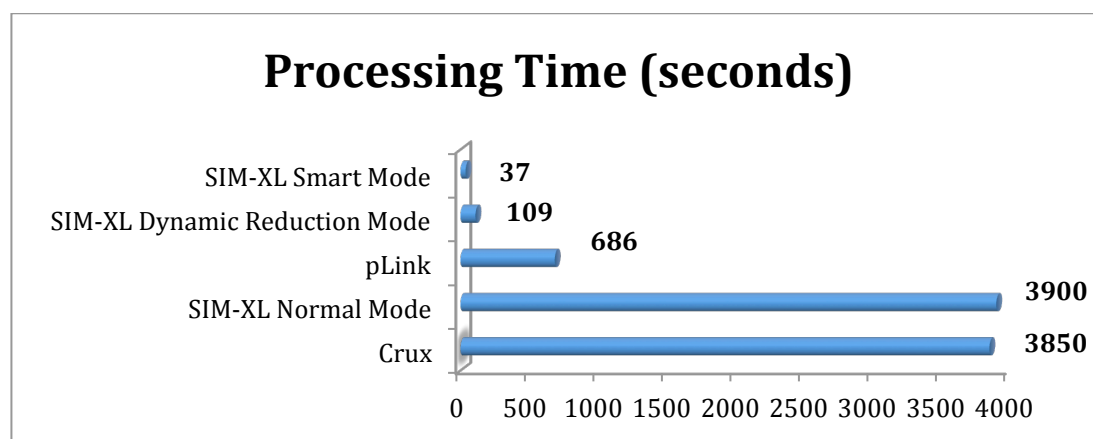


Figura 5.2 – Tempo de processamento, em segundos, das ferramentas de busca após analisar um conjunto de dados relacionados à proteína HSP90.

Uma interface gráfica de fácil utilização, assim como um anotador de espectros capaz de permitir avaliação manual dos resultados, culminaram em uma ferramenta de busca para XL-MS de rápida difusão no meio científico; no momento da escrita desta tese, batemos a marca de 1.000 *downloads* durante 10 meses (Figura 5.3), em mais de cinquenta países diferentes – Figura 5.4 (dados obtidos do *Google Analytics*, 02 de novembro de 2015). Tal divulgação fez com que o SIM-XL fosse

testado também com outros tipos de *cross-linkers*, como *zero-length* (secção 1.3) e a identificação de pontes dissulfeto, mostrando também ser bem sensível nesses casos.

Todos os resultados obtidos e acima mencionados, assim como o programa instalador do SIM-XL estão disponível para *download* em <http://patternlabforproteomics.org/sim-xl>. O site disponibiliza vídeo de demonstração do *software* em uso. Já o código fonte do *software* está disponível em https://bitbucket.org/diogobor/sim_xl.

Specific access	832	832			
	Porcentagem do total: 22,15% (3.756)		Porcentagem do total: 22,15% (3.756)		
1. United States	243	29,21%	13. Austria	13	1,56%
2. Brazil	154	18,51%	14. Spain	10	1,20%
3. Germany	61	7,33%	15. South Africa	9	1,08%
4. Netherlands	49	5,89%	16. Belgium	8	0,96%
5. United Kingdom	45	5,41%	17. South Korea	8	0,96%
6. Switzerland	41	4,93%	18. Poland	8	0,96%
7. China	28	3,37%	19. Denmark	7	0,84%
8. Australia	26	3,12%	20. India	5	0,60%
9. France	22	2,64%	21. Greece	4	0,48%
10. Japan	19	2,28%	22. Israel	4	0,48%
11. Italy	16	1,92%	23. Taiwan	4	0,48%
12. Canada	15	1,80%	24. Indonesia	3	0,36%
			25. Norway	3	0,36%
			26. Sweden	3	0,36%

Figura 5.3 – Exemplos de países que realizaram *download* do SIM-XL, totalizando mais de 800 até o dia 2 de novembro de 2015.

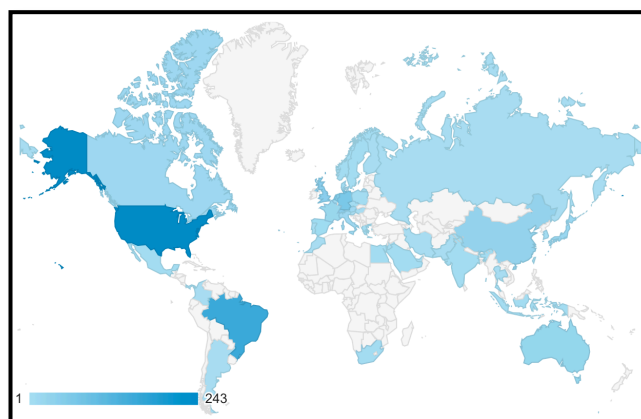


Figura 5.4 – Mapa-múndi representando os países os quais foram realizados *download* do SIM-XL. Quanto mais escuro o país é exibido, maior é o número de *download* feitos.

6 Discussão e Conclusões

Neste trabalho, desenvolveu-se um novo algoritmo de identificação de peptídeos covalentemente ligados, chamado de *Spectrum Identification Machine for Cross-linked peptides* (SIM-XL), capaz de analisar dados gerados através de *cross-linkers* comumente usados, como o DSS e o BS³, assim como futuros agentes de ligação cruzada. O SIM-XL emprega um fluxo de trabalho, dividido em seis etapas: identificar peptídeos lineares com e sem *dead-end*, gerar um novo banco de dados com os peptídeos identificados previamente, filtrar espectros experimentais que são derivados de peptídeos covalentemente ligados, fazer a identificação dos peptídeos com ligação cruzada, avaliar os resultados obtidos e propor os modelos estruturais das proteínas identificadas. Tal fluxo é capaz de otimizar a performance no tempo de processamento dos dados e também na sensibilidade da identificação, fazendo com que o SIM-XL fosse bastante sensível e rápido quando analisado um conjunto de dados para selecionar um modelo estrutural da proteína HSP90, conforme pode ser observado na secção 1. Para verificar a robustez do software uma comparação foi realizada com as ferramentas Crux e pLink, e foi possível notar uma diminuição considerável no tempo de busca (de 1h e 5min para apenas 37 segundos) quando utilizados os íons marcadores, os quais caracterizam que o espectro é derivado de um peptídeo covalentemente ligado, e também usando o modo dinâmico de redução do banco de dados, o qual leva em consideração uma busca prévia por peptídeos do tipo “0”, ou *mono-links*, caracterizando que uma outra versão daqueles peptídeos estará presente na amostra.

O SIM-XL oferece uma interface dinâmica, facilitando o manuseio e a interpretação dos dados, sendo o primeiro *software* a disponibilizar resultado na forma de um mapa de interação dinâmica, mostrando de forma rápida e eficiente, as regiões onde as proteínas se interagem. Duas formas de apresentação do mapa são apresentados: em barra e circular, este último facilitando a visualização dos *links* identificados quando muitas proteínas são identificadas. Uma outra forma de visualizar os dados também é possível, através do mapa de calor, onde mostra, através da intensidade das cores, as regiões da proteína ou do complexo onde se teve o maior número de interações. Com essa ferramenta computacional, também pode-se exportar os dados para serem armazenados em um dos repositórios mais utilizados na área de proteômica, o *Proteomics Identifications – PRIDE*, sendo assim, a primeira

ferramenta de busca de *cross-linking* compatível com dados depositados nesse repositório providos de peptídeos covalentemente ligados. Ele também está integrado ao *PatternLab for Proteomics*, o qual apresenta uma pipeline de módulos relacionados à proteômica computacional.

7 Produção científica e/ou colaborações

O desenvolvimento de uma ferramenta de busca capaz de identificar peptídeos covalentemente ligados durante o período de doutorado apresentado nesta tese, resultou em comunicações em congressos nacionais e internacionais, e na publicação de dois artigos como primeiro autor em revistas de alto prestígio na área de proteômica (Anexos I – Artigos publicados) e um terceiro artigo de primeira autoria encontra-se em elaboração. Tal artigo, refere-se ao visualizador circular do *software*, o qual mostra a facilidade de interpretação quando o número de proteínas identificadas é muito grande

Além disso, o *Spectrum Identification Machine for Cross-linked peptides* proporcionou colaborações científicas nacionais e internacionais, fazendo com que seu desenvolvimento fosse cada vez mais aprimorado, tornando-se assim, reconhecido mundialmente.

7.1 Colaborações em andamento

O presente *software* está sendo utilizado por grupos na área de *cross-linking* em diversas partes no Brasil e no mundo, conforme exemplificado:

- ✓ Laboratório de Espectrometria de Massas Dalton, na Unicamp, São Paulo, com o grupo do Professor Dr. Fabio Cesar Gozzo. Nesta colaboração, o grupo especializado em química vem desenvolvendo novos *cross-linkers* que realizam a ligação cruzada com aminoácidos ácidos (ácido glutâmico e aspártico). Observamos um padrão de fragmentação diferenciado com este procedimento, portanto, estamos criando *scores* especializados para esse tipo de análise.
- ✓ Laboratório de Toxinologia, na Fiocruz, Rio de Janeiro, com a pesquisadora Dra. Ana Gisele da Costa Neves Ferreira. Nesta colaboração, a pesquisadora e seus colaboradores objetivam entender detalhes estruturais do complexo toxina-antitoxina. Mais especificamente dos inibidores proteicos com atividades anti-hemorrágica (DM43, inibidor de metaloproteinases) e antimiotóxica (DM64, inibidor de fosfolipases A₂), incluindo o mapeamento das regiões de interação com as toxinas de venenos de serpentes.
- ✓ Laboratório de Medicina do Departamento de Patologia da Universidade de Cincinnati, Ohio, EUA, grupo liderado pelo pesquisador Dr. Sean Davidson, e colaboração com o pesquisador Dr. John Melchior. Esta colaboração é

voltada ao estudo das interações proteicas da apolipoproteína – (apo)A-I, que é uma das lipoproteínas do tipo de alta densidade – *high density lipoprotein* (HDL), mais comuns nos seres humanos, correspondendo mais de 70% da massa proteica, e estando diretamente ligada ao transporte de lipídeos em meios líquidos, uma vez que elas não têm afinidade com o plasma sanguíneo. As lipoproteínas são frequentemente relacionadas ao colesterol HDL em relação aos benefícios cardioprotetores propostos. Como, por exemplo, a (apo)A-I pode ser considerada como o principal ligante de interação entre o transportador de ATP A1 e o fígado, e periferias, o que propicia a maior parte das HDL em circulação no organismo (BODZIOCH et al., 1999). Um dos problemas fundamentais desse tipo de estudo é que as lipoproteínas tendem a homodimerizar, tornando-se difícil a identificação das interações. Para superar essa dificuldade, o grupo utilizou uma marcação isotópica da lipoproteína, marcando-a com o nitrogênio pesado, o N15, e deixando outra versão da mesma proteína sem marcação (WALKER et al., 2014). Dessa forma, aprimoramos o algoritmo do SIM-XL o qual permite caracterizar a interação entre os homodímeros, ou seja, entre uma proteína mais pesada e a mesma, em sua versão leve. O grupo de Cincinnati já vem reportando resultados positivos com o SIM-XL, resultando na publicação de um novo artigo. (Anexos I – Artigos publicados).

- ✓ Instituto de Biofísica Química, no Instituto Max Plank, Göttingen, Alemanha, pelo pesquisador Dr. Olexandr Dybkov. Nesta colaboração, o pesquisador está utilizando o nosso *software* para analisar um conjunto envolvendo 62 proteínas de *Homo sapiens*. O colaborador vem nos fornecendo importantes *feedbacks* perante o uso do SIM-XL na análise de amostras complexas e dicas de como otimizar a ferramenta para este cenário.

7.1.1 *PatternLab for Proteomics v.4.0*

Uma importante colaboração que realizamos durante o meu doutorado foi a atuação próxima com o ambiente computacional *PatternLab for Proteomics* (secção 4.4) na versão 4.0 – PL4.0, aprimorando algoritmos, juntamente com o Dr. Paulo C Carvalho, de outros módulos do *software* em sua versão 4.0, como por exemplo, o módulo de quantificação relativa que considera a marcação química de isótopos, o iTRAQ. Este módulo, denominado *Isobaric Analyzer*, permite ao usuário realizar a

quantificação de proteínas entre duas amostras diferentes (controle e nativa) a fim de saber se existe alguma proteína mais diferencialmente expressa em uma amostra do que em outra.

Outro módulo que houve contribuição foi na conversão de espectros gerados em formatos diferentes, a fim de facilitar a leitura em distintos módulos do PL4.0. Esse módulo permite a extração de espectros de massa com o objetivo de observar somente ou MS1 ou MS2 ou MS3, sendo independente o formato original do arquivo.

Finalmente, observo que o projeto SIM-XL está inserido no *PatternLab*, sendo caracterizado como um processo “simbiótico” pois, como a ferramenta de busca faz uso de módulos do PL, diversos módulos foram aprimorados, beneficiando assim ambos os *softwares*.

7.2 Artigos publicados em coautoria

Com o objetivo de aprimorar meus conhecimentos na área da proteômica, durante o doutorado ajudei na análise de vários experimentos proteômicos, o que culminou em publicações científicas em revistas de alto impacto, tais como:

- ✓ Paulo C Carvalho*, Diogo B Lima*, Felipe V Leprevost, Marlon M Dias, Juliana S G Fischer, Priscila F Aquino, James J Moresco, John R Yates III, Valmir C Barbosa, “**Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0**”, *Nature Protocols* (11), 102-117 (2016).
- ✓ Priscila Ferreira Aquino*, Diogo Borges Lima*, Juliana de Saldanha da Gama Fischer, Rafael Donadélli Melani, Fabio C S Nogueira, Sidney R S Chalub, Elzalina R. Soares, Valmir C Barbosa, Gilberto B. Domont, Paulo C. Carvalho, “**Exploring the proteomic landscape of a gastric cancer biopsy with the Shotgun Imaging Analyzer**”, *Journal of Proteome Research* (13), pp 314-320, 2014.
- ✓ Felipe V. Leprevost, Richard H. Valente, Diogo Borges, Jonas Perales, Rafael Melani, John R. Yates III, Valmir C. Barbosa, Magno Junqueira, and Paulo C. Carvalho, “**PepExplorer: a similarity-driven tool for analyzing de novo sequencing results**”, *Molecular & Cellular Proteomics* 13, 2480–2489, 2014.

- ✓ Lima, Daniel C; Duarte, Fábio T ; Medeiros, Viviane ; Lima, Diogo B ; Carvalho, Paulo ; Bonatto, Diego ; Batistuzzo de Medeiros, Silvia R, “**The influence of iron on the proteomic profile of *Chomobacterium violaceum***”, *BMC Microbiology*, v. 14, p. 267, 2014.
- ✓ Daniela F. S. Chaves, Paulo C. Carvalho, Diogo B. Lima, Humberto Nicastro , Fabio M Lorenzetti , Mario S. Filho , Sandro M Hirabara , Paulo H.M. Alves , James J Moresco , John R. Yates , Antonio H. Lancha, “**Comparative proteomic analysis of the aging soleus and extensor digitorum longus rat muscles using TMT labeling and mass spectrometry**”, *Journal of Proteome Research* 12 (10), pp 4532–4546, 2013.
- ✓ Felipe da Veiga Leprevost, Diogo Borges Lima, Juliana Crestani, Yasset Perez-Riverol, Nilson Zanchin, Valmir C. Barbosa, Paulo Costa Carvalho, “**Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for Proteomics**”, *Journal of Proteomics* (89): 179-182, 2013.

8 Perspectivas

Como perspectivas, o aprimoramento do *score* do SIM-XL é uma das prioridades, refinando a forma de como a métrica verifica a verossimilhança entre os espectros teórico e experimental, aplicando uma pós filtragem estatística a fim de facilitar a interpretação dos resultados. Outra perspectiva é melhorar a identificação de proteínas com estruturas de homodímeros, as quais possuem duas subunidades idênticas (secção 7.1). A possibilidade da ferramenta de busca identificar peptídeos covalentemente ligados que reagiram com agentes de ligação cruzada do tipo ácido, os quais possuem como resíduo específico, aminoácidos ácidos (ácido glutâmico e ácido aspártico), também é uma das prioridades a serem desenvolvidas e implementadas no SIM-XL. Esse tipo de reação é bastante desafiadora, uma vez que favorece a geração de fragmentos internos, e atualmente não há *softwares* disponíveis que tenham a capacidade de identificar espectros compostos por esse tipo de fragmento. A capacidade de poder trabalhar com novos ALC ampliará, ainda mais, as possibilidades experimentais, aumentando o poder preditivo de estudos com interação proteína-proteína e a determinação de estruturas proteicas a partir da espectrometria de massas.

9 Referências bibliográficas

- [1] AEBERSOLD, R.; MANN, M. Mass spectrometry-based proteomics. **Nature**, v. 422, n. 6928, p. 198–207. doi: 10.1038/nature01511, 2003.
- [2] ALBER, F.; FÖRSTER, F.; KORKIN, D.; TOPF, M.; SALI, A. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. **Annual Review of Biochemistry**, v. 77, n. 1, p. 443–477. doi: 10.1146/annurev.biochem.77.060407.135530, 2008.
- [3] APPEL, R. Comparing the MVC and MVVM patterns along with their respective ViewModels. Retrieved November 2, 2015, from <http://rachelappel.com/comparing-the-mvc-and-mvvm-patterns-along-with-their-respective-viewmodels>.
- [4] AQUINO, P. F.; LIMA, D. B.; DE SALDANHA DA GAMA FISCHER, J.; et al. Exploring the proteomic landscape of a gastric cancer biopsy with the shotgun imaging analyzer. **Journal of proteome research**, v. 13, n. 1, p. 314–320. doi: 10.1021/pr400919k, 2014.
- [5] ARCHAKOV, A. I.; GOVORUN, V. M.; DUBANOV, A. V.; et al. Protein-protein interactions as a target for drugs in proteomics. **PROTEOMICS**, v. 3, n. 4, p. 380–391. doi: 10.1002/pmic.200390053, 2003.
- [6] BAI, H.; MA, W.; LIU, S.; LAI, L. Dynamic property is a key determinant for protein–protein interactions. **Proteins: Structure, Function, and Bioinformatics**, v. 70, n. 4, p. 1323–1331. doi: 10.1002/prot.21625, 2008.
- [7] BENESCH, J. L. P.; RUOTOLO, B. T.; SIMMONS, D. A.; ROBINSON, C. V. Protein Complexes in the Gas Phase: Technology for Structural Genomics and Proteomics. **Chemical Reviews**, v. 107, n. 8, p. 3544–3567. doi: 10.1021/cr068289b, 2007.
- [8] BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; et al. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242. doi: 10.1093/nar/28.1.235, 2000.
- [9] BODZIOCH, M.; ORSÓ, E.; KLUCKEN, J.; et al. The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. **Nature Genetics**, v. 22, n. 4, p. 347–351. doi: 10.1038/11914, 1999.

- [10] BOEHR, D. D.; WRIGHT, P. E. How Do Proteins Interact? **Science**, v. 320, n. 5882, p. 1429–1430. doi: 10.1126/science.1158818, 2008.
- [11] BONDARENKO, P. V.; CHELIUS, D.; SHALER, T. A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. **Analytical Chemistry**, v. 74, n. 18, p. 4741–4749, 2002.
- [12] BORCH, J.; JØRGENSEN, T. J.; ROEPSTORFF, P. Mass spectrometric analysis of protein interactions. **Current Opinion in Chemical Biology**, Mechanisms / Analytical techniques., v. 9, n. 5, p. 509–516. doi: 10.1016/j.cbpa.2005.08.013, 2005.
- [13] BORGES, D.; B. LIMA, D.; B. DE LIMA, T.; et al. Using SIM-XL to identify and annotate cross-linked peptides analyzed by mass spectrometry. **Protocol Exchange**. doi: 10.1038/protex.2015.015, 2015.
- [14] BORGES, D.; PEREZ-RIVEROL, Y.; NOGUEIRA, F. C. S.; et al. Effectively addressing complex proteomic search spaces with peptide spectrum matching. **Bioinformatics (Oxford, England)**, v. 29, n. 10, p. 1343–1344. doi: 10.1093/bioinformatics/btt106, 2013.
- [15] CALLAWAY, E. The revolution will not be crystallized: a new method sweeps through structural biology. **Nature**, v. 525, n. 7568, p. 172–174. doi: 10.1038/525172a, 2015.
- [16] CAÑAS, B.; LÓPEZ-FERRER, D.; RAMOS-FERNÁNDEZ, A.; CAMAFEITA, E.; CALVO, E. Mass spectrometry technologies for proteomics. **Briefings in Functional Genomics & Proteomics**, v. 4, n. 4, p. 295–320. doi: 10.1093/bfpg/eli002, 2006.
- [17] CANFIELD, R. E. Peptides Derived from Tryptic Digestion of Egg White Lysozyme. **Journal of Biological Chemistry**, v. 238, n. 8, p. 2691–2697. Retrieved November 2, 2015, , 1963.
- [18] CARVALHO, P. C.; BARBOSA, V. C. **Um Ambiente Computacional para Proteômica**. Doctorate Thesis, Rio de Janeiro: COPPE/UFRJ. Retrieved from http://www3.cos.ufrj.br/index.php?option=com_publicacao&task=visualizar&id=2119, 2010, March 31.
- [19] CARVALHO, P. C.; HEWEL, J.; BARBOSA, V. C.; YATES, J. R., 3rd. Identifying differences in protein expression levels by spectral counting and feature selection. **Genetics and molecular research: GMR**, v. 7, n. 2, p. 342–356, 2008.

- [20] CHEN, F.; GÜLBAKAN, B.; WEIDMANN, S.; et al. Applying mass spectrometry to study non-covalent biomolecule complexes. **Mass Spectrometry Reviews**, p. n/a–n/a. doi: 10.1002/mas.21462, 2015.
- [21] CHEN, Z. A.; JAWHARI, A.; FISCHER, L.; et al. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. **The EMBO journal**, v. 29, n. 4, p. 717–726. doi: 10.1038/emboj.2009.401, 2010.
- [22] CHU, F.; SHAN, S.; MOUSTAKAS, D. T.; et al. Unraveling the interface of signal recognition particle and its receptor by using chemical cross-linking and tandem mass spectrometry. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, n. 47, p. 16454–16459. doi: 10.1073/pnas.0407456101, 2004.
- [23] CLARKE, R.; SHAJAHAN, A. N.; WANG, Y.; et al. Endoplasmic reticulum stress, the unfolded protein response, and gene network modeling in antiestrogen resistant breast cancer. **Hormone molecular biology and clinical investigation**, v. 5, n. 1, p. 35–44. doi: 10.1515/hmbci.2010.073, 2011.
- [24] DANCHIN, A.; MÉDIGUE, C.; GASCUEL, O.; SOLDANO, H.; HÉNAUT, A. From data banks to data bases. **Research in Microbiology**, v. 142, n. 7-8, p. 913–916, 1991.
- [25] DROIT, A.; POIRIER, G. G.; HUNTER, J. M. Experimental and bioinformatic approaches for interrogating protein–protein interactions to determine protein function. **Journal of Molecular Endocrinology**, v. 34, n. 2, p. 263–280. doi: 10.1677/jme.1.01693, 2005.
- [26] DSS (disuccinimidyl suberate). **DSS - Thermo Fisher Scientific**. Retrieved October 30, 2015, from <https://www.thermofisher.com/order/catalog/product/21655>.
- [27] DYSON, H. J.; WRIGHT, P. E. Intrinsically unstructured proteins and their functions. **Nature Reviews Molecular Cell Biology**, v. 6, n. 3, p. 197–208. doi: 10.1038/nrm1589, 2005.
- [28] EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride). **EDC - Thermo Fisher Scientific**. Retrieved November 4, 2015, from <https://www.thermofisher.com/order/catalog/product/22980>.
- [29] EIDHAMMER, I.; FLIKKA, K.; MARTENS, L.; MIKALSEN, S.-O. Protein, Proteome, and Proteomics. **Computational Methods for Mass Spectrometry Proteomics**. p.1–29. John Wiley & Sons, Ltd. Retrieved October 30, 2015, from <http://onlinelibrary.wiley.com/doi/10.1002/9780470724309.ch1/summary>, 2007.

- [30] ENG, J. K.; JAHAN, T. A.; HOOPMANN, M. R. Comet: an open-source MS/MS sequence database search tool. **Proteomics**, v. 13, n. 1, p. 22–24. doi: 10.1002/pmic.201200439, 2013.
- [31] FENN, J. B.; MANN, M.; MENG, C. K.; WONG, S. F.; WHITEHOUSE, C. M. Electrospray ionization for mass spectrometry of large biomolecules. **Science**, v. 246, n. 4926, p. 64–71. doi: 10.1126/science.2675315, 1989.
- [32] FERREIRA, C. R.; SARAIVA, S. A.; GARCIA, J. S.; et al. Princípios e aplicações da espectrometria de massas em produção animal. **Anais do II Simpósio de Biologia Molecular Aplicada à Produção Animal**, p. 109–136. Embrapa Pecuária Sudeste - São Carlos - SP, 2009, June.
- [33] FISCHER, J. DE S. DA G.; DOS SANTOS, M. D. M.; MARCHINI, F. K.; et al. A scoring model for phosphopeptide site localization and its impact on the question of whether to use MSA. **Journal of Proteomics**. doi: 10.1016/j.jprot.2015.01.008, 2015.
- [34] FISCHLER, M. A.; BOLLES, R. C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. **Commun. ACM**, v. 24, n. 6, p. 381–395. doi: 10.1145/358669.358692, 1981.
- [35] GABALDÓN, T.; RAINEY, D.; HUYNEN, M. A. Tracing the Evolution of a Large Protein Complex in the Eukaryotes, NADH:Ubiquinone Oxidoreductase (Complex I). **Journal of Molecular Biology**, v. 348, n. 4, p. 857–870. doi: 10.1016/j.jmb.2005.02.067, 2005.
- [36] GÖTZE, M.; PETTELKAU, J.; SCHAKS, S.; et al. StavroX--a software for analyzing crosslinked products in protein interaction studies. **Journal of the American Society for Mass Spectrometry**, v. 23, n. 1, p. 76–87. doi: 10.1007/s13361-011-0261-2, 2012.
- [37] HECK, A. J. R.; VAN DEN HEUVEL, R. H. H. Investigation of intact protein complexes by mass spectrometry. **Mass Spectrometry Reviews**, v. 23, n. 5, p. 368–389. doi: 10.1002/mas.10081, 2004.
- [38] VAN DEN HEUVEL, R. H. H.; VAN DUIJN, E.; MAZON, H.; et al. Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. **Analytical Chemistry**, v. 78, n. 21, p. 7473–7483. doi: 10.1021/ac061039a, 2006.

- [39] HILLENKAMP, F.; KARAS, M. Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. **Methods in Enzymology**, v. 193, p. 280–295, 1990.
- [40] HILLENKAMP, F.; KARAS, M.; BEAVIS, R. C.; CHAIT, B. T. Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. **Analytical Chemistry**, v. 63, n. 24, p. 1193A–1203A. doi: 10.1021/ac00024a716, 1991.
- [41] <https://www.pymol.org/>. **The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC**. Retrieved November 3, 2015, from <https://www.pymol.org/>.
- [42] <https://www.rosettacommons.org/>. **Rosetta Commons - The hub for Rosetta modeling software**. Retrieved November 3, 2015, from <https://www.rosettacommons.org/>.
- [43] IGLESIAS, A. H.; SANTOS, L. F. A.; GOZZO, F. C. Identification of cross-linked peptides by high-resolution precursor ion scan. **Analytical chemistry**, v. 82, n. 3, p. 909–916. doi: 10.1021/ac902051q, 2010.
- [44] JAMES, P. Protein identification in the post-genome era: the rapid rise of proteomics. **Quarterly Reviews of Biophysics**, v. 30, n. 4, p. 279–331, 1997.
- [45] JONES, A. R.; EISENACHER, M.; MAYER, G.; et al. The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. **Molecular & Cellular Proteomics**, v. 11, n. 7, p. M111.014381–M111.014381. doi: 10.1074/mcp.M111.014381, 2012.
- [46] JUHASZ, P.; COSTELLO, C. E.; BIEMANN, K. Matrix-assisted laser desorption ionization mass spectrometry with 2-(4-hydroxyphenylazo)benzoic acid matrix. **Journal of the American Society for Mass Spectrometry**, v. 4, n. 5, p. 399–409. doi: 10.1016/1044-0305(93)85005-I, 1993.
- [47] KAO, A.; CHIU, C.; VELLUCCI, D.; et al. Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. **Molecular & cellular proteomics: MCP**, v. 10, n. 1, p. M110.002212. doi: 10.1074/mcp.M110.002212, 2011.
- [48] KEBARLE, P.; VERKERK, U. H. Electrospray: From ions in solution to ions in the gas phase, what we know now. **Mass Spectrometry Reviews**, v. 28, n. 6, p. 898–917. doi: 10.1002/mas.20247, 2009.
- [49] KESKIN, O.; GURSOY, A.; MA, B.; NUSSINOV, R. Principles of Protein–Protein Interactions: What are the Preferred Ways For Proteins To Interact? **Chemical Reviews**, v. 108, n. 4, p. 1225–1244. doi: 10.1021/cr040409x, 2008.

- [50] KOENIG, T.; MENZE, B. H.; KIRCHNER, M.; et al. Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. **Journal of Proteome Research**, v. 7, n. 9, p. 3708–3717. doi: 10.1021/pr700859x, 2008.
- [51] LACROIX, M.; ROSSI, V.; GABORIAUD, C.; et al. Structure and assembly of the catalytic region of human complement protease C1r: a three-dimensional model based on chemical cross-linking and homology modeling. **Biochemistry**, v. 36, n. 21, p. 6270–6282. doi: 10.1021/bi962719i, 1997.
- [52] LEITNER, A.; WALZTHOENI, T.; AEBERSOLD, R. Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. **Nature Protocols**, v. 9, n. 1, p. 120–137. doi: 10.1038/nprot.2013.168, 2014.
- [53] LEITNER, A.; WALZTHOENI, T.; KAHRAMAN, A.; et al. Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. **Molecular & Cellular Proteomics**, v. 9, n. 8, p. 1634–1649. doi: 10.1074/mcp.R000001-MCP201, 2010.
- [54] LEPREVOST, F. V.; LIMA, D. B.; CRESTANI, J.; et al. Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for Proteomics. **Journal of proteomics**, v. 89, p. 179–182. doi: 10.1016/j.jprot.2013.06.013, 2013.
- [55] LEPREVOST, F. V.; VALENTE, R. H.; BORGES, D. L.; et al. PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. **Molecular & Cellular Proteomics**. doi: 10.1074/mcp.M113.037002, 2014.
- [56] LIMA, D. B.; DE LIMA, T. B.; BALBUENA, T. S.; et al. SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis. **Journal of Proteomics**. doi: 10.1016/j.jprot.2015.01.013, 2015.
- [57] LIPFERT, J.; DONIACH, S. Small-Angle X-Ray Scattering from RNA, Proteins, and Protein Complexes. **Annual Review of Biophysics and Biomolecular Structure**, v. 36, n. 1, p. 307–327. doi: 10.1146/annurev.biophys.36.040306.132655, 2007.
- [58] LIU, H.; SADYGOV, R. G.; YATES, J. R., 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. **Analytical chemistry**, v. 76, n. 14, p. 4193–4201. doi: 10.1021/ac0498563, 2004.

- [59] LORENZEN, K.; OLIA, A. S.; UETRECHT, C.; CINGOLANI, G.; HECK, A. J. R. Determination of stoichiometry and conformational changes in the first step of the P22 tail assembly. **Journal of Molecular Biology**, v. 379, n. 2, p. 385–396. doi: 10.1016/j.jmb.2008.02.017, 2008.
- [60] MARTENS, L.; CHAMBERS, M.; STURM, M.; et al. mzML--a Community Standard for Mass Spectrometry Data. **Molecular & Cellular Proteomics**, v. 10, n. 1, p. R110.000133–R110.000133. doi: 10.1074/mcp.R110.000133, 2011.
- [61] MASAMICHI, Y.; B. FEN, J. Electrospray Ion Source. Another Variation on the Free-Jet Theme. Retrieved August 20, 2015, from https://masspec.scripps.edu/mshistory/timeline/time_pdf/1984_YamashitaM_another.pdf, 1983, April 19.
- [62] MCILWAIN, S.; TAMURA, K.; KERTESZ-FARKAS, A.; et al. Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis. **Journal of Proteome Research**, v. 13, n. 10, p. 4488–4491. doi: 10.1021/pr500741y, 2014.
- [63] MERKLEY, E. D.; CORT, J. R.; ADKINS, J. N. Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze. **Journal of Structural and Functional Genomics**, v. 14, n. 3, p. 77–90. doi: 10.1007/s10969-013-9160-z, 2013.
- [64] MILONE, G. **Estatística: geral e aplicada**. São Paulo: Pioneira Thomson Learning, 2004.
- [65] NEILSON, K. A.; ALI, N. A.; MURALIDHARAN, S.; et al. Less label, more free: Approaches in label-free quantitative mass spectrometry. **PROTEOMICS**, v. 11, n. 4, p. 535–553. doi: 10.1002/pmic.201000553, 2011.
- [66] RAO, V. S.; SRINIVAS, K.; SUJINI, G. N.; et al. Protein-Protein Interaction Detection: Methods and Analysis, Protein-Protein Interaction Detection: Methods and Analysis. **International Journal of Proteomics, International Journal of Proteomics**, v. 2014, 2014, p. e147648. doi: 10.1155/2014/147648, 10.1155/2014/147648, 2014.
- [67] ROBINSON, C. V.; SALI, A.; BAUMEISTER, W. The molecular sociology of the cell. **Nature**, v. 450, n. 7172, p. 973–982. doi: 10.1038/nature06523, 2007.
- [68] ROEPSTORFF, P.; FOHLMAN, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. **Biomedical Mass Spectrometry**, v. 11, n. 11, p. 601. doi: 10.1002/bms.1200111109, 1984.
- [69] ROSE, R. J.; VERGER, D.; DAVITER, T.; et al. Unraveling the molecular basis of subunit specificity in P pilus assembly by mass spectrometry. **Proceedings of**

the National Academy of Sciences of the United States of America, v. 105, n. 35, p. 12873–12878. doi: 10.1073/pnas.0802177105, 2008.

[70] Rules of De Novo Sequencing. . Tutorial, . Retrieved September 6, 2015, from <http://ionsource.com/tutorial/DeNovo/rules.htm>, 2015, September 2.

[71] RUSS, A. P.; LAMPEL, S. The druggable genome: an update. **Drug Discovery Today**, v. 10, n. 23-24, p. 1607–1610. doi: 10.1016/S1359-6446(05)03666-4, 2005.

[72] SILVA, R. A. G. D.; HILLIARD, G. M.; FANG, J.; MACHA, S.; DAVIDSON, W. S. A three-dimensional molecular model of lipid-free apolipoprotein A-I determined by cross-linking/mass spectrometry and sequence threading. **Biochemistry**, v. 44, n. 8, p. 2759–2769. doi: 10.1021/bi047717+, 2005.

[73] SINZ, A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. **Journal of mass spectrometry: JMS**, v. 38, n. 12, p. 1225–1237. doi: 10.1002/jms.559, 2003.

[74] SINZ, A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. **Mass Spectrometry Reviews**, v. 25, n. 4, p. 663–682. doi: 10.1002/mas.20082, 2006.

[75] SINZ, A. The advancement of chemical cross-linking and mass spectrometry for structural proteomics: from single proteins to protein interaction networks. **Expert Review of Proteomics**, v. 11, n. 6, p. 733–743. doi: 10.1586/14789450.2014.960852, 2014.

[76] SINZ, A.; ARLT, C.; CHOREV, D.; SHARON, M. Chemical cross-linking and native mass spectrometry: A fruitful combination for structural biology. **Protein Science**, v. 24, n. 8, p. 1193–1209. doi: 10.1002/pro.2696, 2015.

[77] STELZL, U.; WORM, U.; LALOWSKI, M.; et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. **Cell**, v. 122, n. 6, p. 957–968. doi: 10.1016/j.cell.2005.08.029, 2005.

[78] STRONG, M.; SAWAYA, M. R.; WANG, S.; et al. Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 21, p. 8060–8065. doi: 10.1073/pnas.0602606103, 2006.

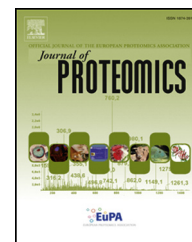
[79] Sulfo-SMCC (sulfosuccinimidyl 4-(N-maleimidomethyl)cyclohexane-1-carboxylate). Retrieved October 30, 2015, from <https://www.thermofisher.com/order/catalog/product/22322>.

- [80] SVERGUN, D. I.; KOCH, M. H. J. Small-angle scattering studies of biological macromolecules in solution. **Reports on Progress in Physics**, v. 66, n. 10, p. 1735. doi: 10.1088/0034-4885/66/10/R05, 2003.
- [81] VAZQUEZ, A.; FLAMMINI, A.; MARITAN, A.; VESPIGNANI, A. Global protein function prediction from protein-protein interaction networks. **Nature Biotechnology**, v. 21, n. 6, p. 697–700. doi: 10.1038/nbt825, 2003.
- [82] VIZCAÍNO, J. A.; CÔTÉ, R. G.; CSORDAS, A.; et al. The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. **Nucleic Acids Research**, v. 41, n. Database issue, p. D1063–1069. doi: 10.1093/nar/gks1262, 2013.
- [83] WALKER, R. G.; DENG, X.; MELCHIOR, J. T.; et al. The Structure of Human Apolipoprotein A-IV as Revealed by Stable Isotope-assisted Cross-linking, Molecular Dynamics, and Small Angle X-ray Scattering. **Journal of Biological Chemistry**, v. 289, n. 9, p. 5596–5608. doi: 10.1074/jbc.M113.541037, 2014.
- [84] WILKINS, M. R.; PASQUALI, C.; APPEL, R. D.; et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. **Bio/Technology (Nature Publishing Company)**, v. 14, n. 1, p. 61–65, 1996.
- [85] WONG, S. S. **Chemistry of protein conjugation and cross-linking**. Boca Raton, Fla. [u.a.]: CRC Press, 1993.
- [86] YANG, B.; WU, Y.-J.; ZHU, M.; et al. Identification of cross-linked peptides from complex samples. **Nature Methods**, v. 9, n. 9, p. 904–906. doi: 10.1038/nmeth.2099, 2012.
- [87] YATES, J. R.; RUSE, C. I.; NAKORCHEVSKY, A. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. **Annual Review of Biomedical Engineering**, v. 11, n. 1, p. 49–79. doi: 10.1146/annurev-bioeng-061008-124934, 2009.
- [88] ZHANG, Y.; FONSLow, B. R.; SHAN, B.; BAEK, M.-C.; YATES, J. R. Protein analysis by shotgun/bottom-up proteomics. **Chemical Reviews**, v. 113, n. 4, p. 2343–2394. doi: 10.1021/cr3003533, 2013.
- [89] ZUBAREV, R. A.; MAKAROV, A. Orbitrap Mass Spectrometry. **Analytical Chemistry**, v. 85, n. 11, p. 5288–5296. doi: 10.1021/ac4001223, 2013.

10 Anexos I – Artigos publicados

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Technical note

SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis

Diogo B. Lima^{a,*}, Tatiani B. de Lima^b, Tiago S. Balbuena^c, Ana Gisele C. Neves-Ferreira^d, Valmir C. Barbosa^e, Fábio C. Gozzo^{b,*}, Paulo C. Carvalho^{a,*}

^aLaboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

^bDalton Mass Spectrometry Laboratory, University of Campinas, São Paulo, Brazil

^cCollege of Agricultural and Veterinary Sciences, State University of São Paulo, Jaboticabal, São Paulo, Brazil

^dLaboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil

^eSystems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

ARTICLE INFO

Article history:

Received 17 December 2014

Accepted 21 January 2015

Keywords:

Cross-linked

Cross-linking

Protein–protein

ABSTRACT

Chemical cross-linking has emerged as a powerful approach for the structural characterization of proteins and protein complexes. However, the correct identification of covalently linked (cross-linked or XL) peptides analyzed by tandem mass spectrometry is still an open challenge. Here we present SIM-XL, a software tool that can analyze data generated through commonly used cross-linkers (e.g., BS3/DSS). Our software introduces a new paradigm for search-space reduction, which ultimately accounts for its increase in speed and sensitivity. Moreover, our search engine is the first to capitalize on reporter ions for selecting tandem mass spectra derived from cross-linked peptides. It also makes available a 2D interaction map and a spectrum-annotation tool unmatched by any of its kind. We show SIM-XL to be more sensitive and faster than a competing tool when analyzing a data set obtained from the human HSP90. The software is freely available for academic use at <http://patternlabforproteomics.org/sim-xl>. A video demonstrating the tool is available at <http://patternlabforproteomics.org/sim-xl/video>. SIM-XL is the first tool to support XL data in the mzIdentML format; all data are thus available from the ProteomeXchange consortium (identifier PXD001677).

© 2015 Elsevier B.V. All rights reserved.

Recently, chemical cross-linking coupled to high-resolution mass spectrometry (XL-MS) emerged as a powerful strategy to broaden the toolset for protein structural characterization and for determining protein–protein interactions. In this approach, the side chains of amino acids in proteins and/or their complexes are covalently linked by reactions with cross-linkers. After enzymatic digestion of the cross-linked

protein(s), cross-linked peptides can be identified by tandem mass spectrometry, generating spatial constraints between amino acid residues. In other words, the distance between two interacting partners (e.g., amino acids of the same protein or different ones or even lipids, RNA, DNA, and carbohydrates) can be inferred through the establishment of the covalent bond, therefore allowing for low-resolution characterization.

* Corresponding authors.

E-mail addresses: diogobor@gmail.com (D.B. Lima), fabio@iqm.unicamp.br (F.C. Gozzo), paulo@pcarvalho.com (P.C. Carvalho).

Ultimately, these distance restraints enable a variety of structural information to be obtained, unraveling important information for understanding protein folding, complex topology and interaction regions [1,2].

Although identifying unmodified peptides (i.e., linear peptides) by mass spectrometry is a rather solved problem in proteomics, reliably identifying pairs of covalently linked peptides (i.e., interpeptide / type 2 cross-links and intrapeptide / type 1 cross-links) is still a bottleneck. To date, there are only a few reference engines for XL search, most prominently Crux [3], CrossWork [4], StavroX[5], and pLink[6]. Differently than in classical proteomics, where proteins are identified from a large redundancy of peptides, generally allowing a 1% false-discovery rate (FDR), we argue that XL-MS studies should avoid false-positives at all costs, as the structural information brought by each cross-link is not only fundamental but also unique. We therefore advocate that the FDR control of classical proteomics is not sufficient in the context of cross-linking: besides having each identification associated with a stringent family-wise error rate estimate or empirically derived score, a personal assessment must be carried out. This is because a single wrong XL identification is enough to create a conflicting protein model or to incorrectly suggest an interaction between proteins. Yet the problem of identifying XL peptides is far more challenging than those of conventional proteomics, as the search space for cross-linked peptides grows quadratically with the number of peptides, which naturally decreases sensitivity and selectivity in the classical search engine approach [7]. Moreover, the population of XL spectra in an LC/MS/MS run is minute when compared to those originating from linear peptides and from type 0 cross-links (i.e., peptides containing dead-end modifications). Consequently, XL identification tools need to be very sensitive and selective to provide means for the user to easily interpret and verify each identification. In our hands, existing tools presented false-positives among their top-scoring hits and were computationally costly (data not shown). Additionally, they provided limited or no resources at all for viewing, editing, and manually validating XL peptide identifications, which is a fundamental and time-consuming step in any experiment addressing XL-MS.

Here we present the Spectrum Identification Machine for Cross-Linked Peptides (SIM-XL), a fast and sensitive XL search engine that is part of the PatternLab for proteomics environment [8]. The SIM-XL software was programmed in C# with .NET Framework 4.5. The software requires a computer with Windows 7 or later, and at least 4 GB of RAM. To perform a search, the user begins by specifying the tandem mass spectrum file, the FASTA sequence database file, as well as parameters related to sample preparation (e.g., modifications) and mass spectrometry features (e.g., ppm). A detailed description of each SIM-XL parameter is available in its integrated manual, which is accessible through the Help menu, Read Me option. The current version is compatible with the Mascot Generic Format [9], MS2 [10], and mzML[11] and can work directly with Thermo .RAW files if the freely available MSFileReader is locally installed.

Among its novelties, we point out three: (I) SIM-XL builds on a new paradigm for search-space reduction. As previously mentioned, the larger the search space (i.e., the set of possibilities of theoretical peptides or, in this case, combinations of peptides

originating from a database matching the experimental precursor mass), the lower the sensitivity of the search engine [12]. To address the quadratic growth arising from cross-linked peptide candidates, our search engine employs a dynamic database reduction heuristic to eliminate possibilities by considering only combinations that contain at least one linear peptide identified with a dead-end modification. (II) SIM-XL search engine takes advantage of reporter ions [13], i.e., fingerprints of mass spectral peaks found almost exclusively in tandem mass spectra derived from cross-linked peptides. By searching only tandem mass spectra with these reporter ions, the chance of false-positive identifications is decreased and the search speed increases considerably. We note that feature I and II are optional and therefore can be switched on or off. (III) Our search engine provides a user-friendly Graphical User Interface (GUI) that allows the user to assess each identification interactively through a spectrum viewer and annotation tool.

As described, SIM-XL can reduce the search space when working in dynamic database reduction mode. To do this, it begins by wrapping the Comet [14] search engine to perform a preliminary search aiming to identify peptide spectrum matches (PSMs) of linear peptides with a user-configurable XCorr cutoff (default value: 1.5). A secondary database is then dynamically generated with all possible pairs containing a linear peptide with a dead-end and a peptide from the identified proteins having a reactive site on the sequence. When this mode is not activated, by contrast, all pairs of peptides containing a reactive site in the sequence database are considered. SIM-XL makes use of reporter ions by only considering tandem mass spectra that contain reporter ions from cross-linked peptides [13]. The idea of using reporter ions for different strategies has been previously reported [7,15–17]. This significantly decreases the number of spectra to be searched and thus improves on both selectivity and processing time, especially on large data sets. As previously reported, these so-called reporter fragment ions are specific to Lys-Lys cross-linked or dead-end modified peptides and consist of a rearranged lysine side chain and the spacer arm of the linker. SIM-XL can work with any set of diagnostic ions that are specified in its GUI or XML parameter file. Yet we note that to take advantage of reporter-ion filtering, MS/MS acquisition should start at least at m/z of the lowest mass reporter ion (in the case of DSS/BS3, m/z 220). Although this is usually not a problem for TOF instruments, special attention should be paid when acquiring data using Orbitrap analyzers, as the m/z range is more restricted. For cases such as these, mass spectra from the same precursor can optionally be acquired in different ranges of m/z and our tool will automatically generate consensus (merged) spectra.

SIM-XL uses multi-threading to take advantage of multiple hardware cores and therefore significantly increase the search speed. These speedups become evident especially in those cases in which (i) no dead-end modifications are specified, so the software has to work on the full search space; (ii) MS/MS acquisition does not contain reporter ions, so no reporter-ion spectrum filtering is possible; or (iii) the number of proteins in the database is large, so the search space is huge, even using the two filtering options described previously.

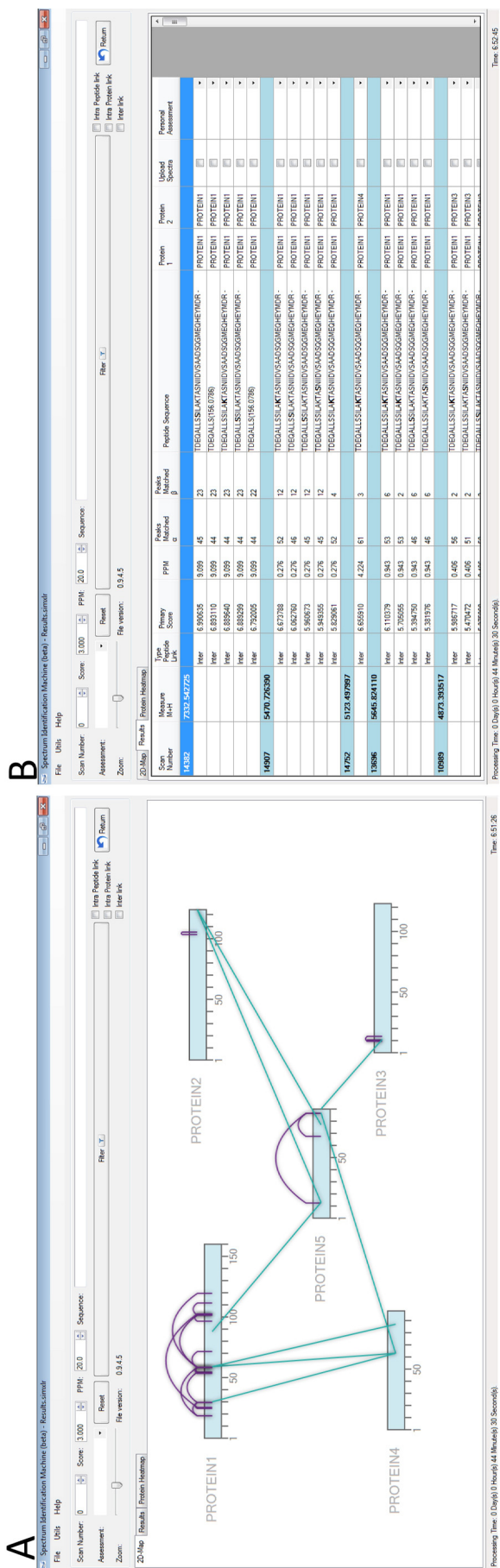


Fig. 1 – Panel A shows the 2D interaction map for a data set from a protein complex consisting of five proteins (data undisclosed). By clicking on the Results tab, the dynamic report is displayed. Mass spectra can be accessed by either clicking on the 2D-map link or on any dynamic report result. Dynamic cutoff scores can be applied and combined with personal assessments (i.e., excellent, good, medium, fair, or poor).

Once the search engine finishes, SIM-XL presents its results in three interconnected modes: the 2D interaction map and the dynamic report (Fig. 2) and a “Heat map” of a pairwise comparison (not shown). The former provides a graphical representation of all XL links among the protein(s) and the latter allows the user to sort identifications by several criteria (e.g., primary score and ppm) and to make an assessment for each spectrum. This assessment can be saved in SIM-XL’s dynamic report to help in keeping track of which spectra were already evaluated and approved to be considered, say, when determining a protein’s structure or inferring protein–protein interactions. Fig. 1 shows a screenshot of SIM-XL’s main GUI exemplifying its 2D interaction map and the dynamic report.

We note that in the dynamic report, buttons are made available to allow for the upload of high-quality annotated XL-MS spectra to the online database we are developing to support the creation of even more effective machine learning approaches for identifying cross-linked peptide species originating from different cross-linkers, mass spectrometers, etc. When a spectrum is uploaded, only information pertaining to that single spectrum, including which peptides were cross-linked, is sent to our server. As the number of XL mass spectra is generally low for an experiment, we advocate that libraries such as in this initiative can become fundamental for the development of future, more sensitive tools.

SIM-XL’s spectrum viewer and 2D interaction map are its high points, constituting unique features that greatly simplify the assessment of identification candidates, each of which can be easily visualized by double-clicking on the identification provided in the dynamic search engine report or in the graphical representation in the 2D map. The spectrum viewer allows the user to view the annotated ions and to zoom in on a region of interest in the mass spectrum. Its importance to a SIM-XL user resides in that it allows for the manual validation of all assignments given by the software and, importantly, for the easy verification of other assignment possibilities for the same mass spectrum, thus supporting unbiased judgments, independent of SIM-XL’s scoring heuristic through an immediate comparison assisted by SIM-XL’s theoretical spectrum predictor. We strongly encourage viewing further details and functionalities in our online supplementary video available at <http://patternlabforproteomics.org/sim-xl/video>.

We demonstrate the effectiveness of SIM-XL by analyzing a data set aiming to aid in establishing a structural model for the Human HSP90; the data set was generated as previously described [13]. Briefly, disuccinimidyl suberate (DSS) cross-linker was dissolved in dimethylformamide (DMF, Thermo Scientific) at a stock concentration of 27.1 mM. DSS was added to the human C-terminal of HSP90 at a 1:50 (protein: DSS) ratio and incubated with the sample for 2 h at room temperature. Cross-linking reaction was quenched with ammonium bicarbonate 100 mM. Reduction and alkylation of cysteine residues were performed using dithiothreitol and iodacetamide during 30 min at 60 °C and at room temperature, respectively. The sample was digested with trypsin (Promega) at 1:50 for 16 h at 37 °C. The peptides were fractionated using an Oasis HLB cartridge (Waters Corp.) and eluted with different concentration of acetonitrile and analyses were performed using a Thermo Q-Exactive mass

spectrometer equipped with a nano-electrospray source coupled to a nano EasyLC (Thermo, San Jose – CA).

The search engines used were Crux v. 2.0 and SIM-XL 1.0. All searches were performed using carbamidomethylation of cysteine as fixed modification; for SIM-XL, the variable modifications were a dead-end DSS of 156.0786 Da and a DSS cross-linker mass modification of 138.0681 Da; the remaining parameters were defaults. The precursor and fragment ion-mass tolerances were of 20 ppm. The sequence database comprised the sequence of HSP90 plus those from five decoy sequences. Benchmarking was performed on a MacPro with Intel Xeon X5670 processors.

The searching times were 1 h 4 min 10 s, 1 h 5 min, 1 min 49 s, and 37 s, respectively, for Crux, SIM-XL in normal mode (i.e., with features I and II off), SIM-XL with dynamic database reduction activated, and SIM-XL with both dynamic database reduction and the use of reporter ions activated. Plots of the cumulative number of non-decoy hits among the 50 top-scoring spectra for these searches are found in Fig. 2.

Our data set consisted of 1,788 tandem mass spectra, of which 973 contained at least one XL reporter ion. Among the top-50 mass spectra reported by SIM-XL with both dynamic database reduction and reporter ion modes turned off, three XL originated from an HSP90-decoy peptide pair, two of which presented reporter ions; in this case, an HSP90 peptide could actually be there, but having its counterpart wrongly attributed. As for the remaining non-decoy identifications, all but four did not have at least one reporter-ion peak. We also note that

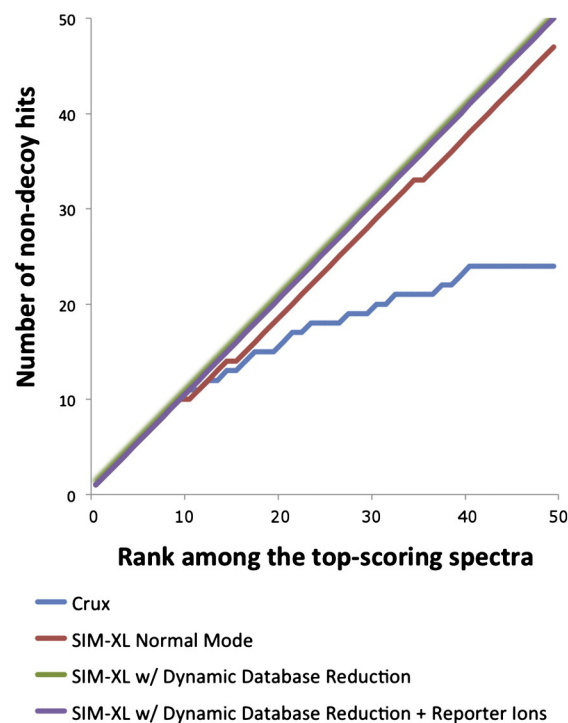


Fig. 2 – Plots of the cumulative number of non-decoy hits for SIM-XL operating in different modes and for Crux, considering in all cases the 50 top-scoring mass spectra. The “SIM-XL w/ Dynamic Database Reduction” and “SIM-XL w/ Dynamic Database Reduction + Reporter Ions” lines coincide.

among all non-decoy identifications appearing in normal mode, at least one of the cross-linked peptides had their linear peptide version identified with a dead-end. Two of the three HSP90-decoy duets did not have dead-end counterparts.

We recommend using SIM-XL with both the dynamic database reduction and the reporter ion modes activated. These can drastically decrease the chances of a false-positive and significantly increase search speed; for the task at hand, this resulted in reducing the search time from 1 h 5 min to 37 s. We believe these two features underscore SIM-XL as a promising tool for addressing next generation challenges such as *in vivo* cross-linking [19].

Finally, we note that SIM-XL is the first XL tool capable of exporting results in the forthcoming mzIdentML 1.2 format [18], established by the Proteomics Standards Initiative (PSI). This has enabled us to perform the first complete submission of an XL data set to the ProteomeXchange consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository [20] (data set identifier PXD001677, DOI 10.6019/PXD001677). Consequently, all our data are readily available to the scientific community. The remaining files, which include the search results, parameter files, and the sequence database, are available at the project's website (<http://patternlabforproteomics.org/sim-xl>).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank FAPESP, FAPERJ, CAPES, Universal CNPq, Microsoft Research—Microsoft Azure Research Award, Programa Estratégico de Apoio à Pesquisa em Saúde (PAPES), and Fundação Oswaldo Cruz for financial support. We thank the PRIDE Team for working together with us to enable SIM-XL to support the next version of mzIdentML. The authors declare no competing financial interest.

REFERENCES

- [1] Preston GW, Radford SE, Ashcroft AE, Wilson AJ. Covalent cross-linking within supramolecular peptide structures. *Anal Chem* Aug. 2012;84(15):6790–7.
- [2] Merkley ED, Cort JR, Adkins JN. Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze. *J Struct Funct Genomics* Sep. 2013;14(3):77–90.
- [3] McIlwain S, Draghicescu P, Singh P, Goodlett DR, Noble WS. Detecting cross-linked peptides by searching against a database of cross-linked peptide pairs. *J Proteome Res* May 2010;9(5):2488–95.
- [4] Rasmussen MI, Refsgaard JC, Peng L, Houen G, Højrup P. CrossWork: software-assisted identification of cross-linked peptides. *J Proteomics* Sep. 2011;74(10):1871–83.
- [5] Götz M, Pettelkau J, Schaks S, Bosse K, Ihling CH, Krauth F, et al. StavroX—a software for analyzing crosslinked products in protein interaction studies. *J Am Soc Mass Spectrom* Jan. 2012;23(1):76–87.
- [6] Yang B, Wu Y-J, Zhu M, Fan S-B, Lin J, Zhang K, et al. Identification of cross-linked peptides from complex samples. *Nat Methods* Jul. 2012;9(9):904–6.
- [7] Borges D, Perez-Riverol Y, Nogueira FCS, Domont GB, Noda J, da Veiga Leprevost F, et al. Effectively addressing complex proteomic search spaces with peptide spectrum matching. *Bioinforma Oxf Engl* May 2013;29(10):1343–4.
- [8] Carvalho PC, Fischer JSG, Xu T, Yates III JR, Barbosa VC. PatternLab: from mass spectra to label-free differential shotgun proteomics". In: Andreas Baxevanis AI Board, editor. *Curr. Protoc. Bioinforma.*, vol. Chapter 13; Dec. 2012. p. Unit13.19.
- [9] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* Dec. 1999; 20(18):3551–67.
- [10] McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, et al. MS1, MS2, and SQT—three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom* RCM 2004;18(18):2162–8.
- [11] Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* Jan. 2011;10(1) [R110.000133–R110.000133].
- [12] Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* Oct. 2010; 73(11):2092–123.
- [13] Iglesias AH, Santos LFA, Gozzo FC. Identification of cross-linked peptides by high-resolution precursor ion scan. *Anal Chem* Feb. 2010;82(3):909–16.
- [14] Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics* Jan. 2013; 13(1):22–4.
- [15] Perez-Riverol Y, Sánchez A, Noda J, Borges D, Carvalho PC, Wang R, et al. HI-bone: a scoring system for identifying phenylisothiocyanate-derivatized peptides based on precursor mass and high intensity fragment ions. *Anal Chem* Apr. 2013;85(7):3515–20.
- [16] Tang X, Munske GR, Siems WF, Bruce JE. Mass spectrometry identifiable cross-linking strategy for studying protein–protein interactions. *Anal Chem* Jan. 2005;77(1): 311–8.
- [17] Perez-Riverol Y, Sánchez A, Ramos Y, Schmidt A, Müller M, Betancourt L, et al. In silico analysis of accurate proteomics, complemented by selective isolation of peptides. *J Proteomics* Sep. 2011;74(10):2071–82.
- [18] Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* Jul. 2012;11(7) [M111.014381–M111.014381].
- [19] Weisbrod CR, Chavez JD, Eng JK, Yang L, Zheng C, Bruce JE. In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy. *J Proteome Res* Apr. 2013;12(4):1569–79.
- [20] Vizcaíno JA, Côté RG, Csordas A, Dienes JA, Fabregat A, Foster JM, et al. The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* Jan. 2013;41(Database issue):D1063–9.

PROTOCOL EXCHANGE | COMMUNITY CONTRIBUTED Using SIM-XL to
identify and
annotate cross-linked peptides analyzed by
mass spectrometry

Diogo B. Lima, Tatiani B. de lima, Tiago S. Balbuena, Ana Gisele C. Neves-Ferreira,
Valmir C. Barbosa, Fabio C. Gozzo & Paulo C. Carvalho

Fiocruz - PR,RJ / IQ - Unicamp / COPPE - UFRJ / Unesp - SP

Abstract

State-of-the-art structural mass spectrometry-based proteomics is often accomplished by using cross-linkers to covalently bind two or more amino-acid groups. This strategy is complementary to classical structural biology and therefore broadens the toolset for analyzing protein and protein-protein complex structures. One of the greatest challenges in identifying cross-linked peptides in complex protein mixtures is computationally dealing with the large search space, which grows quadratically with each peptide included in the sequence database. The **Spectrum Identification Machine for Cross-Linked Peptides (SIM-XL)** software uses an algorithm that overcomes this limitation by capitalizing on experimental features that allow it to effectively address the massive combinatorial problem at hand, and presents the results in a user-friendly manner. Thus, SIM-XL is recommended for studies dealing with protein structure and protein-protein interaction in either simple or complex protein mixtures. SIM-XL also allows the sharing of results through PRIDE by exporting them in the mzIdentML format.

Subject terms: Computational biology Protein analysis Proteomics

Structural biology Biochemistry

Keywords: cross-linked peptide identification cross-linked peptides

cross-linking search engine mass spectrometry

search tool computational biology structural proteomics

Introduction

One of the goals of systems biology is to determine how a system works, beginning at the molecular-level characterization of proteins and their interactions up to the level of

cellular physiological pathways¹. In a broad sense, the determination of protein structures and protein-protein interactions has an immediate impact on many biological and biotechnological fields, including medicinal chemistry, immunology, and molecular medicine to name a few. For example, the determination of a protein's three-dimensional structure allows us to better understand its biological function and consequently opens the doors to the design of new drugs or even provides new insights into how to completely engineer new proteins to fulfill specific biological functions². This qualifies the understanding of cell biology, at the atomic level, as a key to answering a number of important biological questions and thus providing a roadmap for numerous biotechnological applications such as the design of new drugs. The current gold-standard methods for determining a protein's structure are X-ray diffraction³ and nuclear magnetic resonance (NMR)⁴. However, the majority of proteins and their complexes are not amenable to these strategies, as they either do not crystallize, require large amounts of high-purity protein, or the system is just too large to be analyzed⁵. These facts make evident the need for developing novel structural approaches that are applicable to a larger number of systems.

Chemical cross-linking coupled to high-resolution mass spectrometry (XL-MS) has become a key method to broaden the toolset for protein structural characterization and in determining protein-protein interactions. During sample preparation, synthetic cross-linkers are included; these covalently link to the side chains of amino acids in proteins and/or their complexes. The sample is then digested, ultimately allowing the cross-linked peptides to be identified by analyzing the mixture using tandem mass spectrometry. The covalently linked peptides carry a very unique piece of information, i.e., distance constraints, which allows for further elucidation of tertiary structures and protein-protein interactions^{6,7}.

In this protocol we describe the key steps for using the **Spectrum Identification Machine for Cross-linked Peptides (SIM-XL)**, a fast and sensitive XL search engine that is part of the PatternLab for proteomics environment⁸, to analyze tandem mass spectrometry data derived from cross-linked peptides. A video demonstrating SIM-XL v1.0 in action is available at <http://patternlabforproteomics.org/sim-xl/video>.

Equipment

Hardware

- A computer with at least a 4 GB RAM and 2 computing cores is recommended.

Software

- Windows 7 (32 or 64 bits) or later.
- For reading Thermo RAW files, the MSFileReader must be installed.
- The .NET framework 4.5, which will be automatically updated by SIM-XL if necessary.
- The SIM-XL software is available at <http://patternlabforproteomics.org/sim-xl/>.

Data files

- SIM-XL v1.0 is compatible with data files in the formats mzML 1.0, MS2⁹, and Mascot Generic Format (MGF)¹⁰, and can work directly with Thermo RAW files if the freely available MSFileReader is locally installed.
- SIM-XL exports results to its own format (i.e., *.simxlr) and in the mzIdentIML 1.2 draft format that is currently under development by the HUPO Proteomics Standards Initiative to support the identification of cross-linked peptides. We note this is the first search engine that enables complete submissions of XL-MS data to PRIDE¹¹, being therefore is compatible with the PRIDE Inspector software.

Procedure

1. Software installation:

Download SIM-XL by clicking on the *Download* button at <http://patternlabforproteomics.org/sim-xl>.

2. Workflow

The following workflow demonstrates how to perform a search using the SIM-XL search engine.

2.1. Execute the *Spectrum Identification Machine for Cross-linked Peptides* (**Figure 1**)

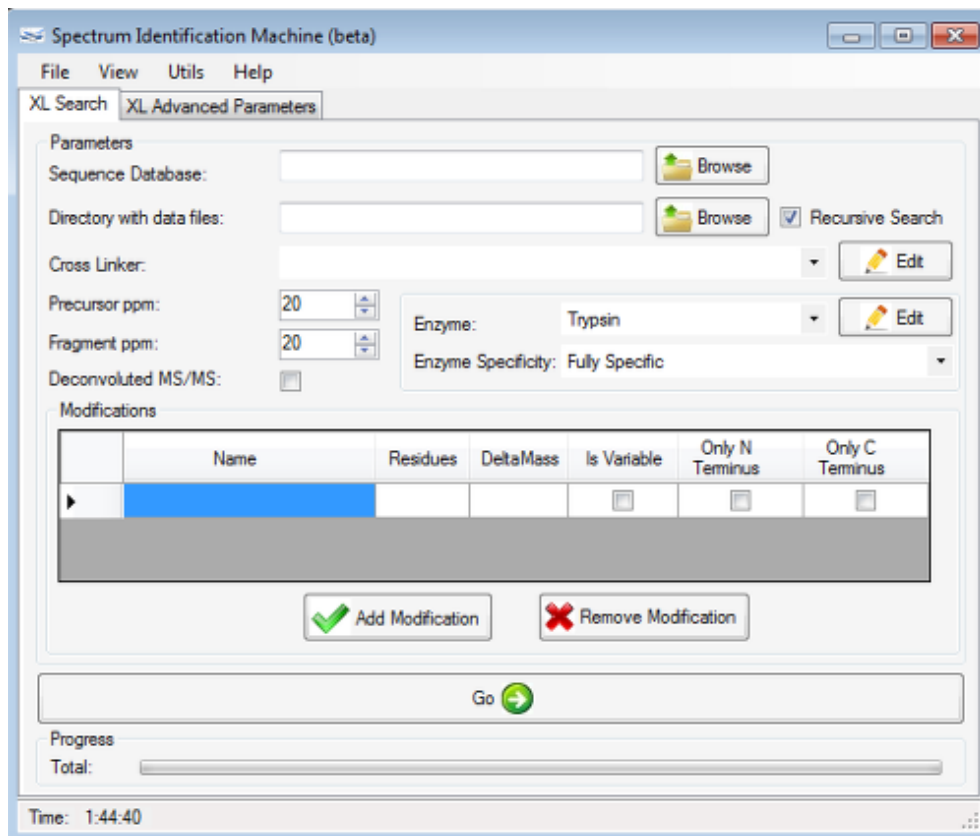


Figure 1: Graphical User Interface for the main window of SIM-XL.

2.2. Specify the protein sequence database file by clicking on the *Browse* button. The file extension has to be *.FASTA, *.T-R, or *.T. We refer the reader to **Basic Protocol 1: Preparing a sequence database to be searched by Prolucid or the academic Sequest⁸** for more on how to generate target-decoy (*.T-R) databases compatible with PatternLab.

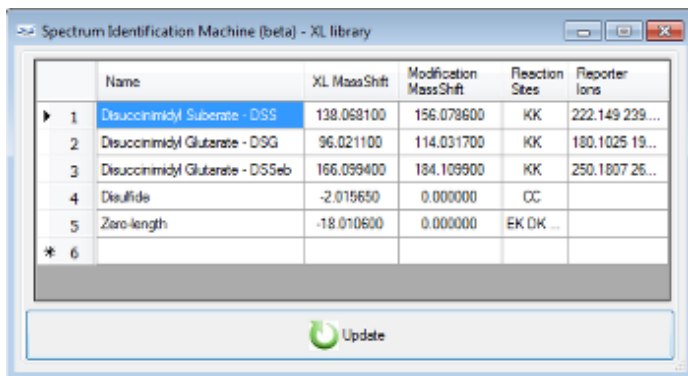
2.3. Specify a directory containing XL-MS data in any of the following formats: mzML 1.1.0, MS2, MGF, or Thermo RAW. If *Recursive Search* is checked, all subdirectories will be searched.

2.4. Select one of the pre-registered cross-linkers in the drop-down list. By default, there are five cross-linkers registered:

- Disuccinimidyl Suberate (DSS);
- Disuccinimidyl Glutarate (DSG);
- Disuccinimidyl Glutarate (DSSeb);
- Disulfide; and
- Zero-length.

Optionally including a new cross-linker in the XL library

2.4.1. To register a new cross-linker, click on the *Edit* button beside the *Cross-linker* field and the *XL library* window will pop up (**Figure 2**).



	Name	XL Mass Shift	Modification Mass Shift	Reaction Sites	Reporter Ions
▶ 1	Disuccinimidyl Suberate - DSS	138.068100	156.078600	KK	222.149 239...
2	Disuccinimidyl Glutarate - DSG	96.021100	114.031700	KK	180.1025 19...
3	Disuccinimidyl Glutarate - DSGeb	166.039400	184.109900	KK	250.1807 26...
4	Disulfide	-2.015650	0.000000	CC	
5	Zero-length	-18.010600	0.000000	EK DK ...	
* 6					

Figure 2: XL library. A cross-linker can be inserted or removed in this window.

2.4.2. Fill out the fields *XL Name*, *XL Mass Shift* (reaction XL mass) in Daltons, *Reaction Sites*, and optionally, the *Modification Mass Shift* and *Reporter Ions* masses.

Parameter descriptions

2.4.2.1. *XL Name* is a user-defined identifier for the cross-linker.

2.4.2.2. *XL Mass Shift* is the net mass of the cross-linker that will be added to the peptide masses.

2.4.2.3. *Reaction Sites* are all the combinations of amino acids that react with the cross-linker. If the cross-linker reacts with the N-terminus, the keyword *N-TERM* should be included. For example, the entry for the DSS cross-linker should be *KK KS SS KN-TERM SN-TERM*. Similarly, C-terminal reactivity can be specified using the keyword *C-TERM*. Note that reaction sites must be separated by a single space. Also, when N-TERM is specified SIM-XL performs an unspecific digestion on the first 20 amino acids of all database entries, starting with Methionine; this is done to address the possibility of signal peptides, frequently present in public databases.

The optional *Modification Mass Shift* field defines an artificial modification caused by the cross-linker. For example, DSS/BS3 can react with a single lysine residue, generating the so-called dead-end modification. If this field is filled out and the *Dynamic DB Reduction* box (see **item 2.10.5**) is checked, SIM-XL will use Comet¹² to identify the modified peptides and use this information to only consider theoretical cross-linked peptides that had at least one chain identified with the modification. Consequently, this reduces the search space and is thus indicated when searching complex samples. If the *Modification Mass Shift* field is left blank, SIM-XL will always consider all possible combinations of cross-linked peptides.

The optional *Reporter Ions* field defines the *m/z* of fragments that are specific to cross-linked peptides¹³. If these *m/z* values are given, SIM-XL will only search MS/MS spectra containing at least one of the corresponding fragments, thus speeding up the search; otherwise, SIM-XL will search all spectra.

2.4.3. To finish the new cross-linker inclusion, click on the *Update* button and then on the *OK* button. This will make the new cross-linker available within the library and usable for searching.

2.4.4. In order to delete a cross-linker entry from the library, the whole line must be selected (by clicking on the row's header cell) and then the DEL key must be pressed, followed by a click on the *Update* button and another on the *OK* button.

2.5 Select an enzyme from the drop-down enzyme list. By default, Trypsin and Lys-C are registered.

Optionally including a new enzyme in the Enzyme library

2.5.1. To include a new enzyme, click on the *Edit* button beside the *Enzyme* field; the enzyme library window will pop-up (**Figure 3**). In an empty line, complete the corresponding fields with the enzyme's name and a regular expression encoding the enzymatic cleavage.

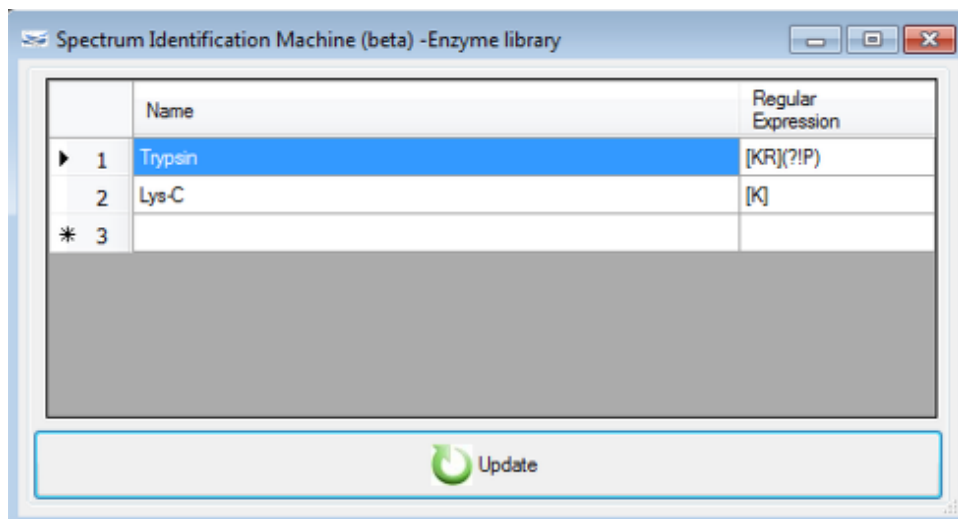


Figure 3: Enzyme library. An enzyme can be inserted or removed in this window. A regular expression is required to specify the cleavage sites of a new enzyme. For example, the regular expressions for Trypsin and Lys-C are '[KR](?!P)' and '[K]', respectively. For more on building regular expressions we refer the reader to <http://www.regular-expressions.info/>.

2.5.2. In case one wishes to remove an enzyme from the library, the whole line must be selected (by clicking on the row's header cell) and then the DEL key must be pressed, followed by a click on the *Update* button and another on the *OK* button.

2.6. Choose *Enzyme Specificity* from the drop-down list; the options are: *Semi-Specific* or *Fully Specific*. The latter refers to peptides originating from a complete digestion (i.e., with enzyme cleavage sites at both the C- and the N-terminus). *Semi-specific* means that the

constraint of having a cleavage site at one end is lifted. For example, in the sequence *R.APBCK.A*, where “.” denotes the occurrence of cleavage, selecting *Semi-Specific* will make SIM-XL consider *A*, *AP*, *APB*, *APBC*, *K*, *CK*, *BCK*, *PBCK*, and *APBCK*. Otherwise (i.e., if *Fully Specific* is selected), the search space is limited to *APBCK*.

2.7. Specify the *Precursor* and *Fragment ppm* tolerances.

2.8. Check the *Deconvoluted MS/MS* option if the spectra in the data files are deconvoluted (i.e., decharged and de-isotoped). We refer the reader to YADA as a tool for deconvoluting mass spectra¹⁴.

Considering modifications

2.9. To include a modification from the *Modification library*, select it from the drop-down list and then click on the *OK* button found in the *Modifications* group box (**Figure 4**).



Figure 4: Select a pre-defined modification or add a new one. To edit a pre-defined modification, click on the *Edit* button.

Optionally including new modifications in the Modification library

2.9.1. To include a new modification or edit an existing one, click on the *Edit* button.

2.9.1.1. A new window will open. To include a new modification, fill out the fields *Name*, monoisotopic *Mass Shift*, and *Amino acid(s)* to can carry the modification.

2.9.1.2. To delete a modification from the library, select the whole line by clicking on the row's header cell, then press the DEL key, click on the *Update* button, and then on the *OK* button.

2.9.2. The user should indicate whether the modification is a variable one and whether it applies to the C-term and/or the N-term by checking the corresponding boxes. For example, if not all Methionines in the sample are expected to be oxidized, the modification should be checked as *variable*; however, for modifications that are expected in all occurrences of the amino acid, such as, say, carbamidomethylation of cysteine, the *variable* option should remain unchecked.

2.9.3. To remove a modification, select the desired one, then click on the *Remove Modification* button and confirm the exclusion.

2.10. The *XL Advanced Parameters* tab allows access to various parameters that are not usually required to be changed for XL-MS analyses. In this tab, the parameters are divided into three groups: *SIM-XL Parameters*, *Dynamic DB Reduction* (which uses Comet¹² for performing a preliminary search), and *_Common Parameters* (indicating that the parameters belong to both the SIM-XL and the Comet search engines).

SIM-XL Parameters

2.10.1. *Number of Isotopic Possibilities*: The precursor mass stored in raw data files may not correspond to the monoisotopic peak. This option allows the software to find the correct monoisotopic peak, which is required to identify the molecule but at the cost of opening up the search space. So, for example, for a peptide with a monoisotopic mass of 4000 Da, the most intense peak in the isotopic envelope is M+2 (that is, 4002 Da), which will most likely be selected as the precursor mass. If the number of possibilities is set to three, SIM-XL will search this MS/MS spectrum considering the precursor masses 4002, 4001, and 4000, plus or minus the given ppm tolerance. In this example, the correct monoisotopic precursor mass is 4000 and thus can be correctly identified by SIM-XL. If a high number of isotopic possibilities is set, the search space will increase accordingly and impact SIM-XL's sensitivity negatively.

2.10.2. *Minimum AA Residues per chain*: Minimum number of amino acids a peptide should have to be considered a candidate for cross-linking.

2.10.3. *Maximum results to report*: Number of top-scoring XL candidates to be reported for each queried spectrum.

2.10.4. *Intra-link charge*: Maximum charge of precursor ions to be searched in an intra-molecular cross-link candidate. All candidates are also considered for the inter-molecular searches.

Dynamic DB Reduction parameters

2.10.5. *Enable*: If enabled, SIM-XL will run Comet¹² to perform a preliminary search to identify peptides containing the modification specified in the cross-linker definition as modification mass shift (see **item 2.4.2**). These identifications are used to generate a theoretical cross-linked peptides database in which all entries contain at least one chain previously identified with the modification.

2.10.6. *XCorr Threshold*: Minimum Comet XCorr value for identifying peptides containing

the modification mass shift specified in **item 2.4.2**.

2.10.7. *Minimum number of peptides*: Minimum number of peptides required to include a protein to be later used during the generation of the theoretical cross-linked peptide sequence database.

Common Parameters

2.10.8. *Maximum missed cleavages*: Maximum number of missed cleavages allowed during the theoretical digestion of the sequence database.

2.10.9. *Minimum and Maximum MH*: Minimum and maximum masses of singly-charged peptide ions to be searched.

2.10.10. *Peaks Matched cutoff*: Minimum number of matching fragments between the theoretical and experimental spectra. Identifications not satisfying this constraint will be discarded.

2.10.11 *Merge High Resolution Spectra*: Enabling this option will let the search engine merge two or more high-resolution spectra that likely belong to the same precursor. The motivation is that several MS/MS spectra may be acquired from the same precursor during its elution peak; the merged spectrum will have a better signal-to-noise ratio than the individual spectra.

2.10.11.1 *Chromatogram Tolerance (seconds)*: Maximum time difference, in seconds, between tandem mass spectra having the same precursor mass to be merged.

Note: The *Log* field reports notes on the progress of the search.

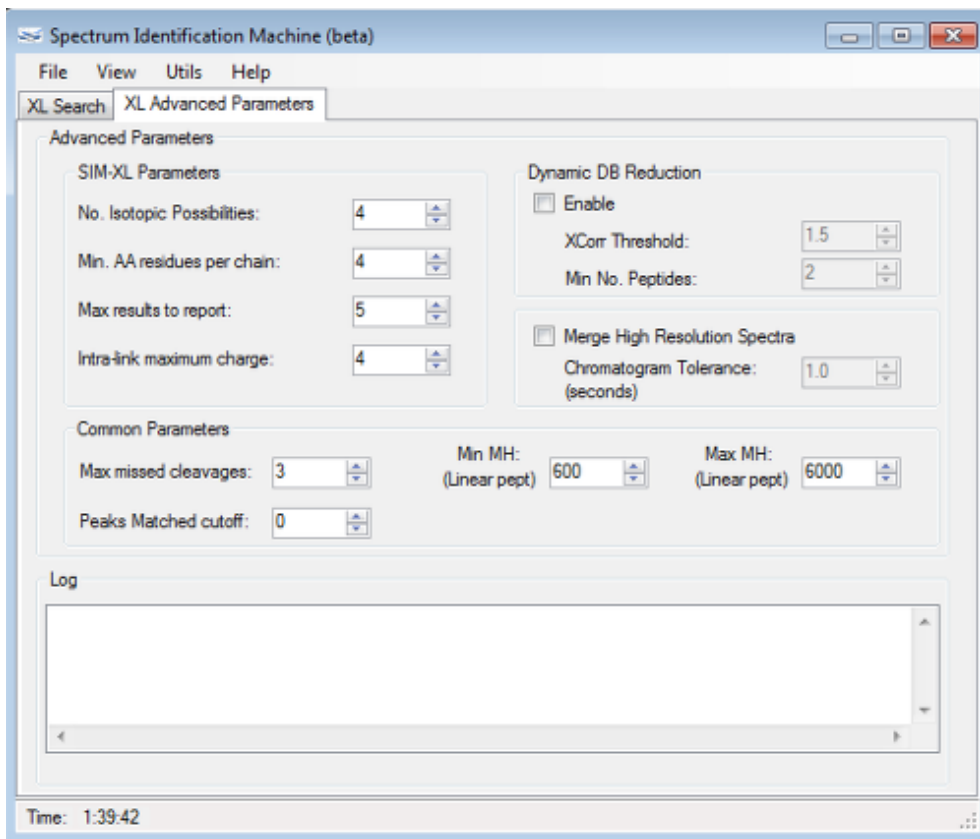


Figure 5: XL Advanced Parameters tab.

2.11 Once all parameters have been set, we strongly recommend saving them for future searches. This is accomplished by selecting Save SIM-XL Params from the File menu or pressing CTRL + S (**Figure 6**).

2.12. To load a previously saved search parameter file, select Load SIM-XL Params from the File menu or press ALT + L (**Figure 6**).

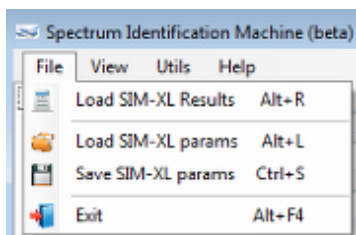


Figure 6: Save or Load SIM-XL params.

2.13. To begin searching, click on the GO button in the *XL Search* tab.

3. Exploring the results

Note: At this point we recommend saving the results by selecting Save results from the File menu or pressing CTRL + S (**Figure 7**).

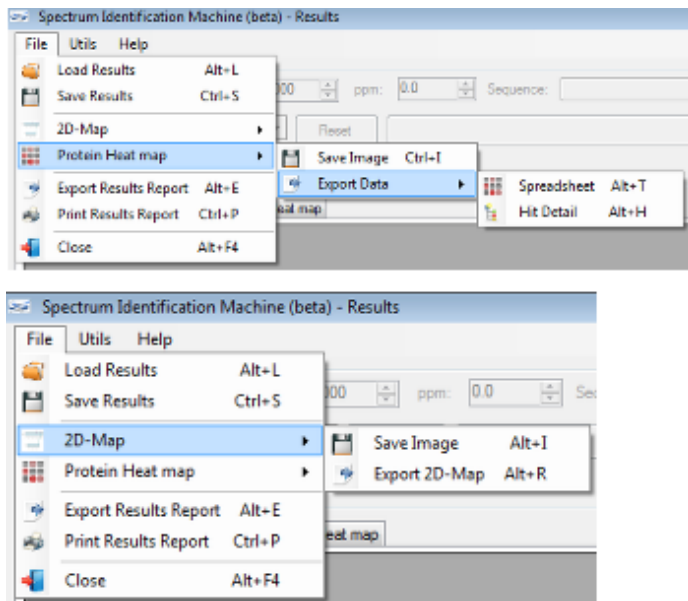


Figure 7: Results Browser's File menu. Here the user can access many features, such as loading or saving the search results, exporting the 2D-Map to an image or PDF file, exporting the Protein Heat map to an image or Excel© (XLS) file, and printing or exporting the results to a spreadsheet with all protein interactions and their hit details.

3.1. 2D-Map

3.1.1. The 2D-Map is an interactive map showing all the cross-links identified with a score above the cutoff value given in the *Score* field. In this map (**Figure 8**), each protein is represented as a rectangle of size directly proportional to its sequence length, with residue numbers marking the ticks at the bottom. The protein's ID is shown outside the rectangle, on the left. Each intra-protein cross-link is represented as a red arc. Likewise, inter-protein cross-links are given in blue straight lines. The position of each linker corresponds to the amino-acid number. The user can customize the view by left-clicking on the rectangles and dragging them around, as well as zoom in or out using the zoom bar. By letting the mouse pointer hover over the cross-linker representation, a window will pop up showing linker details such as the linking amino acids and their positions.

3.1.2. By right-clicking on a cross-linker representation, a pop-up window will be displayed showing all the identified cross-links, with corresponding scan numbers, scores, and charge states. The user can then left-click on a desired identification to access the spectrum with the *Spectrum Viewer* (see **item 4**).

3.1.3. The 2D-Map can be exported as a PNG, TIFF, or JPG image, or as a PDF file: On the

File menu, select 2D-Map, then Save Image (ALT + I), or select Export 2D-Map (ALT + R) (Figure 7).

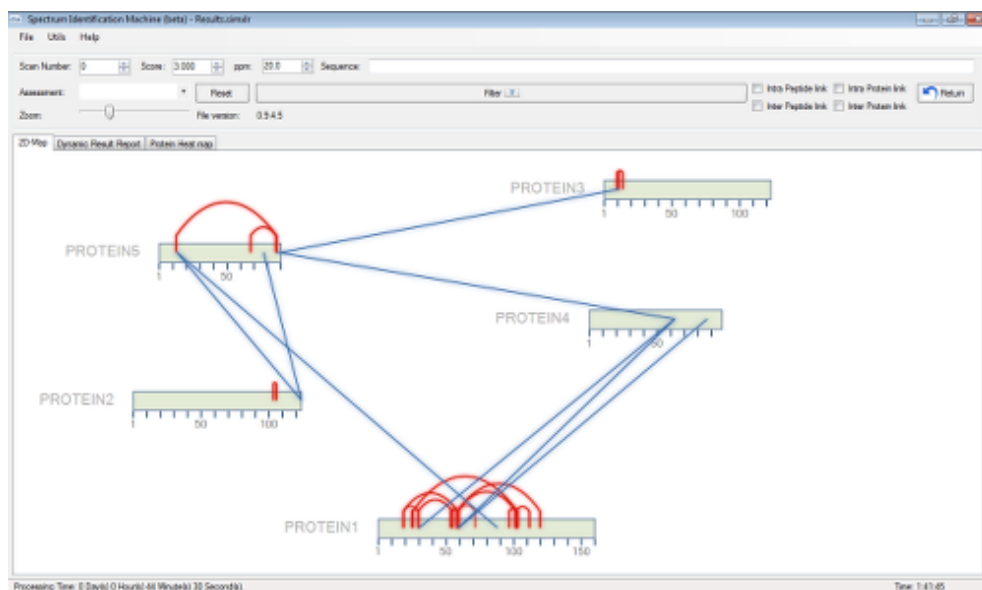


Figure 8: Protein-protein interaction map (2D-Map).

3.2. Loading results

3.2.1. SIM-XL can load results in the SIM-XL (*.simxlr) or mzIdentML 1.2 draft file formats. This can be accomplished in several ways, the easiest one being to double-click on a SIM-XL results file. If the *Result Browser* window is open, select Load Results from the File menu or press ALT + S, as seen in **Figure 7**. Otherwise, if the main window is open, select Load SIM-XL Results from the File menu (or press ALT + R), as seen in **Figure 6**.

We note that mzIdentML results can only be loaded within SIM-XL by accessing the Load Results option, which will open another windows where the user can specify both required files, the *Result file* (*.mzIdentML) and the *Data file* (e.g. *.mzML, *.MGF, *.MS2, or *.RAW).

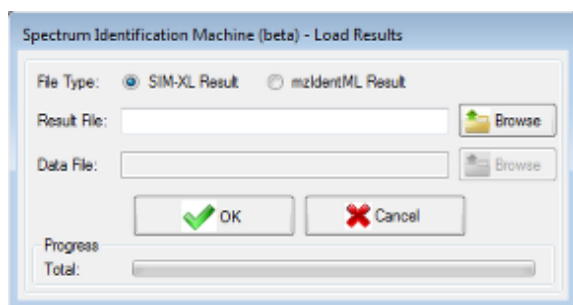


Figure 9: Input file window. SIM-XL accepts the mzIdentML format, in addition to its own format (simxlr).

3.3. Dynamic Result Report

3.3.1. A dynamic report is made available by clicking on the *Dynamic Result Report* tab

(Figure 10). The user can sort/search the results according to user-specified criteria. By double-clicking on an entry, the *Spectrum Viewer* will open, enabling access to the spectrum for the identification in that entry (see item 4).

Scan Number	Measure Value	Type Peptide Link	Primary Score	ppm	Peptide Matched α	Peptide Matched β	Peptide Sequence	Protein 1	Protein 2	Upload Spectra	Personal Assessment
14382	7332.542775	Intra	5.990620	3.089	45	23	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	6.893110	3.089	44	23	TDEDALLSISKSTGQ	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	6.890640	3.089	44	23	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	6.893290	3.089	44	23	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	6.792005	3.089	44	22	TDEDALLSISKSTGQ	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
14387	5470.726390	Intra	6.672780	0.276	52	12	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	6.062760	0.276	46	12	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	5.962672	0.276	45	12	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	5.949395	0.276	45	12	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	5.829061	0.276	52	4	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
14752	5123.497997	Intra	6.859310	4.224	61	3	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN4	<input type="checkbox"/>	<input type="checkbox"/>
13636	5640.824110	Intra	6.113279	0.943	53	6	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	5.705605	0.943	53	2	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	5.394790	0.943	46	6	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	5.381976	0.943	46	6	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
10989	4873.392617	Intra	5.906717	0.406	56	2	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN0	<input type="checkbox"/>	<input type="checkbox"/>
		Intra	5.470472	0.406	51	2	TDEDALLSSILAKTASNIQVSAADSGGMEGHEYMDR	PROTEIN1	PROTEIN0	<input type="checkbox"/>	<input type="checkbox"/>

Figure 10: Dynamic Result Report.

3.3.1.1. Filtering results

3.3.1.1.1. *ScanNumber*: In case this field is not empty, only spectra whose scan numbers match that of this field will be displayed.

3.3.1.1.2. *Score*: Only results containing identification scores greater than or equal to this value will be displayed.

3.3.1.1.3. *ppm*: Only results containing a ppm less than or equal to this value will be displayed.

3.3.1.1.4. *Sequence*: Only results from peptides containing the sequence input to this field will be displayed. The user can further specify whether only results from intra-link or inter-link peptides/proteins are to be displayed.

3.3.1.1.5. *Assessment*: Only results containing a *Personal Assessment* equal to this value will be displayed.

All criteria specified in these fields will be reflected in all tabs (2D-Map, Dynamic Result Report, and Protein Heat map).

Fields in the Dynamic Result Report

3.3.2. The *Peptide Sequence* field shows the search result's identified peptides. The amino acids interacting with the cross-linker are shown in a bold typeface. Double-clicking on this field makes the *Spectrum Viewer* (see **item 4**) pop up, enabling the user to assess the spectrum that resulted in the respective identification.

3.3.3. The *Protein 1* and *Protein 2* columns display the protein(s) that contain the identified peptide sequence(s). The remaining proteins having conserved regions containing the peptide(s) can be assessed by double-clicking on one of these columns (**Figure 11**).

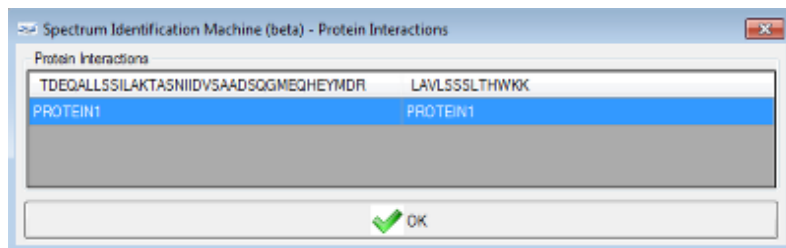


Figure 11: Proteins containing identified peptides window. The header displays the identified peptides: TDEQALLSSILAKTASNIIDVSAADSQGMQHEYMDR and LAVLSSSLTHWKK. Below, the protein that contains these sequence(s) (PROTEIN1) is listed.

3.3.4. The *Upload Spectra* column is part of a global effort for improving cross-linker identification scoring functions. By checking beside the desired spectra and then selecting Send Spectra to Server from the Utils menu (or pressing ALT + S), the user will donate the identifications and spectra for further research on the topic.

3.3.5. The *Personal Assessment* column allows the user to input a personal assessment on the quality of each identification. This is accomplished by selecting from the drop-down list one of the five choices ranging from Excellent to Poor.

3.3.6. At this point, we recommend saving the results once again so that the personal assessments can be included. This is done by selecting Save Results from the File menu or pressing CTRL + S.

3.4. Protein Heat map

3.4.1. The Protein Heat map (**Figure 12**) displays regions where inter-protein cross-linkers were identified. To generate such a map, two proteins must be selected by using the horizontal-axis and the vertical-axis drop-down lists. The heat map can be limited to desired amino acids by selecting them in the *Reaction Site* field.

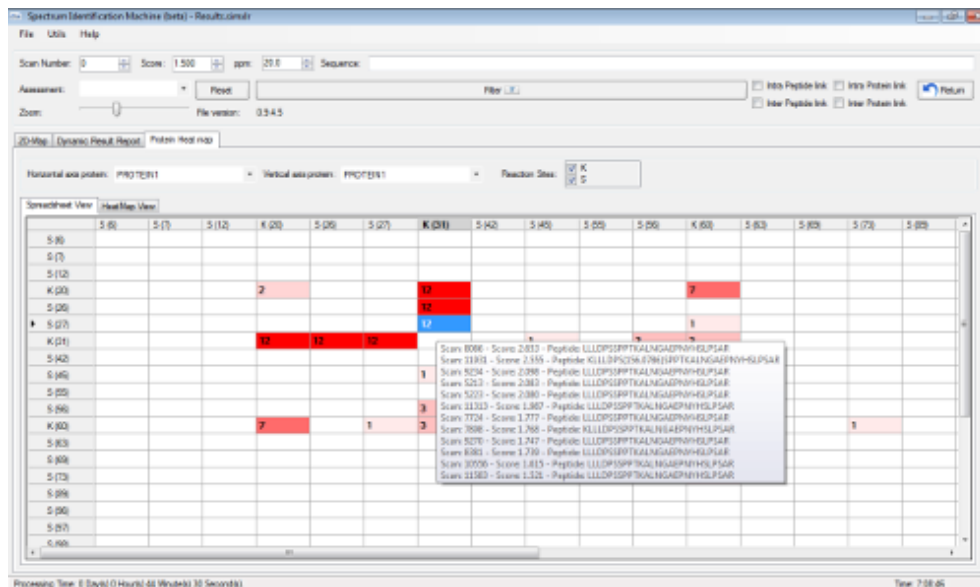


Figure 12: A Protein Heat map showing the interaction regions defined by cross-linkers. The red scale is associated with the number of identified XL spectra. By clicking on a cell, all identifications supporting that interaction will be displayed.

3.4.2. The Protein Heat map can be exported as an image or a spreadsheet file containing the information about the interactions. To export the map as an image, select Save Image, then Protein Heat map, from the File menu (or press CTRL + I); to export it as a spreadsheet, select Spreadsheet from the Export Data menu (or press ALT + T), or Hit Details (i.e., information of the identifications contained in each cell) (ALT + H) (**Figure 7**).

3.5. Utils menu

3.5.1. *Report Fusion*: This option allows merging several SIM-XL results into a single report. To accomplish this, select Report Fusion from the Utils menu (or press ALT + F) and select all files to be joined.

3.5.2. *Custom Report Results*: This option allows the addition or removal of columns in the Dynamic Result Report. For this, select Custom Report Results from the Utils menu (or press ALT + C). Following that, a new window containing all columns that can be included or removed will be displayed. After checking beside the features of interest, click on the OK button.

4. Spectrum Viewer

4.1. The *Spectrum Viewer* (**Figure 13**) displays an annotated XL mass spectrum. The *Spectrum View* tab allows the user to browse the spectrum, zoom in and out, and easily view which peak was attributed to the corresponding fragment. To zoom in, click and drag

the mouse over the desired m/z range (**Figure 14**). To zoom out, double left-click.

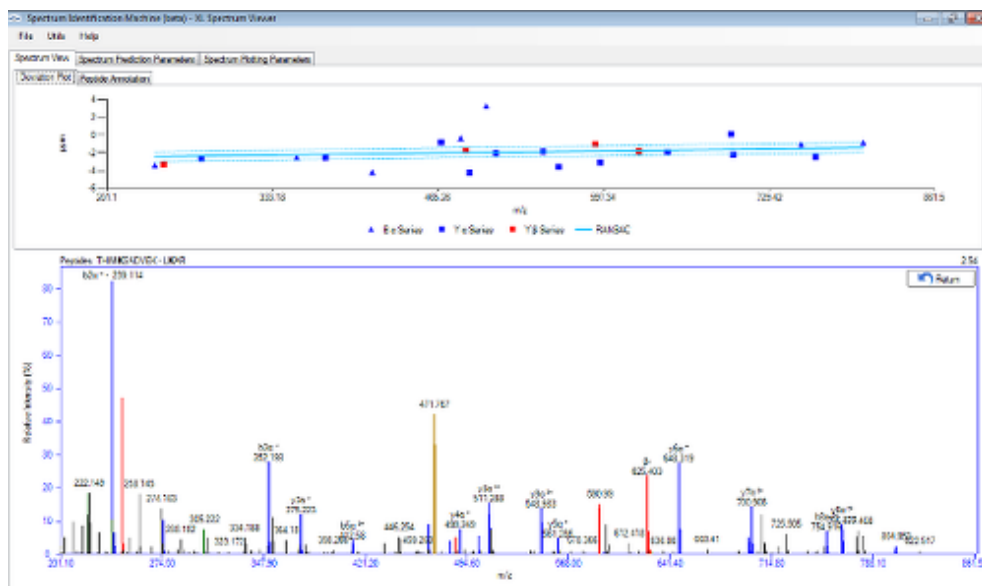


Figure 13: XL Spectrum Viewer. The *Spectrum View* tab allows the user to browse the spectrum, zoom in and out, as well as easily view which peaks were attributed to which series. A ppm deviation plot is available above the annotated mass spectrum.

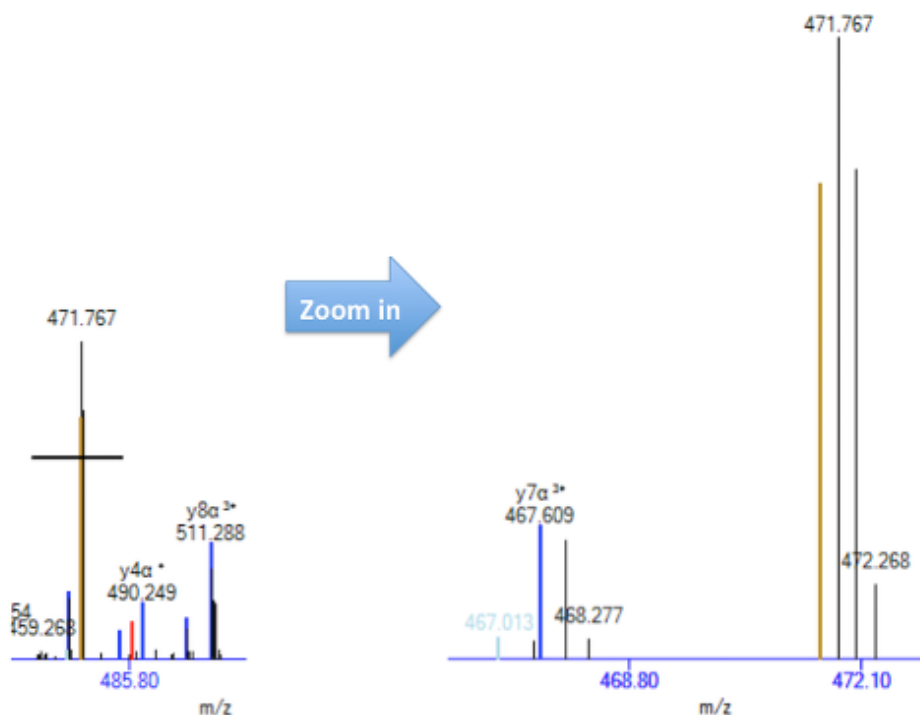


Figure 14: Zoom-in on a specific m/z range of the XL mass spectrum.

4.2. A ppm deviation plot is available above the annotated mass spectrum (**Figure 13**). This plot displays the deviation between the theoretical and experimental peaks, in ppm, along the m/z range. The continuous blue line represents the *Random Sample Consensus*

(RANSAC), which is a linear regression of matched peaks. The blue dotted lines represent three standard deviations from the regression line. To save this plot, right-click on the image and choose between *Copy to clipboard* or *Save*.

4.3. The *Peptide Annotation* tab (**Figure 15**) shows a fragmentation diagram of the cross-linked peptide. The plot can be saved by right-clicking on the image and choosing between *Copy to clipboard* or *Save*.



Figures 15: The *Peptide Annotation* tab displays the fragmentation diagram of the cross-linked peptides.

4.4. The *Spectrum Prediction Parameters* tab (**Figure 16**) provides a table showing all theoretical fragments and their assignments when matched. Matching criteria are shown in the panel on the left. Changes in these parameters, followed by pressing the *Plot* button, will update the assignments. These features allow the user to verify, for example, the effects of changing the cross-linker position or even, say, to evaluate the impact on the score of oxidizing a methionine or even changing the sequence of the matched peptide(s).

4.4.1. *Peptide Sequence 1 and 2*: These are the sequences of α and β chains. For intra-links, fill out the *Peptide Sequence 1* field only. Any modification mass shift must be enclosed in parentheses after the modified amino acid. For example: oxidation of methionine should be input as 'M(15.9949)'.

4.4.2. *Position XL 1 and 2*: These are the positions of the cross-linked residues in both chains. For intra-links, both fields correspond to the cross-linking positions in the α chain.

4.4.3. *Deconvoluted MS/MS*: Check this option if the spectrum is deconvoluted (see **item 2.8**).

4.4.4. *XL Mass*: The cross-linker mass shift (see **item 2.4.2.2**).

4.4.5. *ppm*: The tolerance of each spectrum peak match.

4.4.6. *Precursor Charge and Precursor Mass*: The charge and mass of the precursor ion.

4.4.7. *Ion Series*: Check the series to be considered by the *Spectrum Viewer*. For inter-links, both α and β series should be checked.

4.4.8. Click on the *Plot* button to update the *Spectrum Viewer*.

4.4.9. The *Load Example* button loads an example spectrum.

Peptide Sequence 1: THINKSADVEK
 Peptide Sequence 2: LKAR
 Position XL 1: 5 Position XL 2: 2 Matching peaks / Experimental peaks: 27 / 221 = 12.2%

Deconvoluted MS/MS:
 XL Mass: 138.06810
 Reaction Site: KK
 ppm: 20
 Precursor Charge: 4
 Precursor Mass: 1883.04123

Ion Series
 A C X Z
 B Y α β
 Only Matched Y
 Neutral Loss (H₂O and/or NH₃)

Charge	m/z	Series	FinalAA	Matched	Number
2	804.4511	B_ α	V	True	9
2	766.4298	Y_ α	M	True	8
2	754.9169	B_ α	D	True	8
1	741.4081	A	A	True	7
2	700.9096	Y_ α	K	True	7
2	698.8612	AlphaXL_Chain	α -	True	0
1	648.3204	Y_ α	S	True	6
3	628.3523	Precursor		True	1
1	625.4037	BetaXL_Chain	β -	True	0
3	594.6701	Y_ α	H	True	10
3	590.658	Y_ β	K	True	3
1	561.2884	Y_ α	A	True	5
3	548.9838	Y_ α	I	True	9
3	511.2091	Y_ α	M	True	8

Figure 16: Spectrum Prediction Parameters tab. The user can adjust parameters to check assignments.

4.5. The *Spectrum Plotting Parameters* tab (**Figure 17**) allows the user to enter an individual experimental mass-spectral peak list. The user can also specify the XL reporter ions.

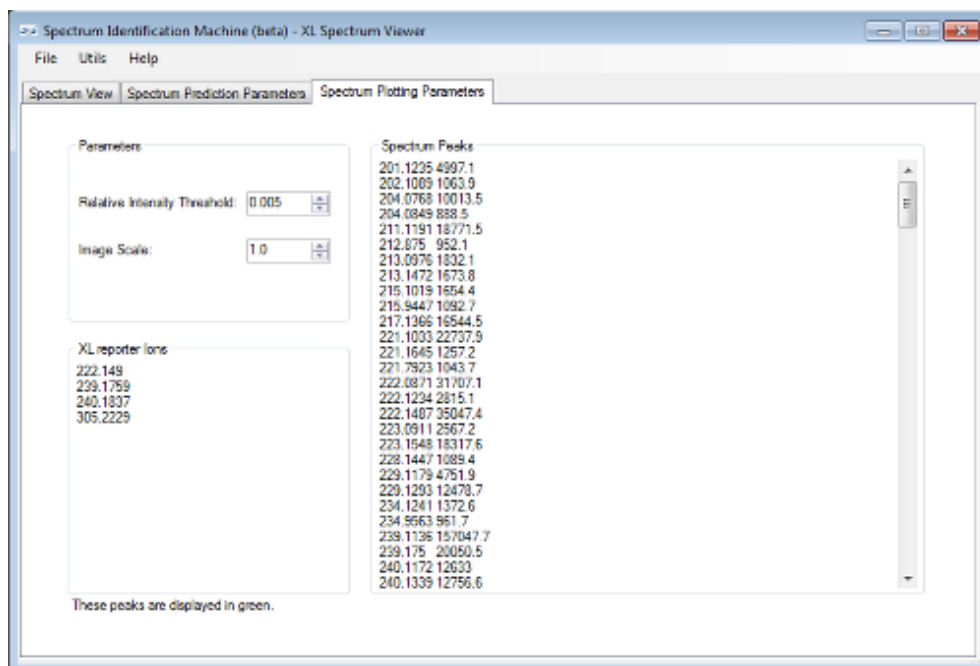


Figure 17: Spectrum Plotting Parameters tab. The user can add individual mass-spectral peak lists to visualize the spectral assignments.

4.6. To save the annotated XL mass spectrum, select Save Spectrum from the File menu or press CTRL + S. An image can be saved by selecting Export Image, then Spectrum Image, from the File menu or pressing CTRL + I. To load the annotated XL mass spectrum, select Load Spectrum from the File menu or press ALT + L.

4.7. The user can customize the XL mass spectrum annotation by selecting Custom Spectrum View from the Utils menu or pressing CTRL + L. A new window will open with several options, as shown in **Figure 18**.

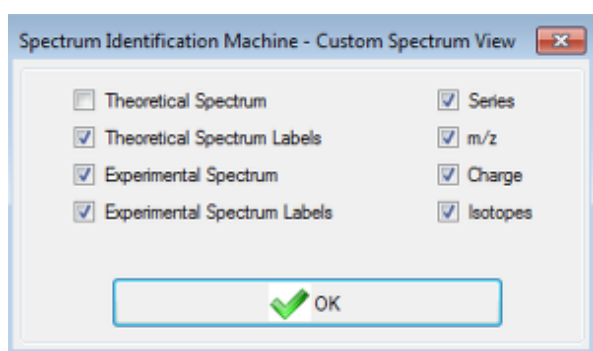


Figure 18: Custom Spectrum View. Spectrum annotations can be customized by checking the option in this menu.

4.8. By selecting Send Spectrum to Server from the Utils menu (or pressing ALT + S), SIM-XL will upload the annotated spectrum to a server and thus contribute to a global effort aiming towards creating more sophisticated cross-linker scoring functions through machine

learning.

References

- [1] A. Ma'ayan, A. D. Rouillard, N. R. Clark, Z. Wang, Q. Duan, and Y. Kou, "Lean Big Data integration in systems biology and systems pharmacology," *Trends Pharmacol. Sci.*, Aug. 2014.
- [2] C. Jost and A. Plückthun, "Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs," *Curr. Opin. Struct. Biol.*, vol. 27C, pp. 102–112, Jul. 2014.
- [3] E. F. Garman, "Developments in x-ray crystallographic structure determination of biological macromolecules," *Science*, vol. 343, no. 6175, pp. 1102–1108, Mar. 2014.
- [4] R. Hänsel, L. M. Luh, I. Corbeski, L. Trantirek, and V. Dötsch, "In-Cell NMR and EPR Spectroscopy of Biomacromolecules," *Angew. Chem. Int. Ed Engl.*, Jul. 2014.
- [5] M. Strong, M. R. Sawaya, S. Wang, M. Phillips, D. Cascio, and D. Eisenberg, "Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 21, pp. 8060–8065, May 2006.
- [6] G. W. Preston, S. E. Radford, A. E. Ashcroft, and A. J. Wilson, "Covalent cross-linking within supramolecular peptide structures," *Anal. Chem.*, vol. 84, no. 15, pp. 6790–6797, Aug. 2012.
- [7] E. D. Merkley, J. R. Cort, and J. N. Adkins, "Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze," *J. Struct. Funct. Genomics*, vol. 14, no. 3, pp. 77–90, Sep. 2013.
- [8] P. C. Carvalho, J. S. G. Fischer, T. Xu, J. R. Yates 3rd, and V. C. Barbosa, "PatternLab: from mass spectra to label-free differential shotgun proteomics," *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis AI*, vol. Chapter 13, p. Unit13.19, Dec. 2012.
- [9] W. H. McDonald, D. L. Tabb, R. G. Sadygov, M. J. MacCoss, J. Venable, J. Graumann, J. R. Johnson, D. Cociorva, and J. R. Yates 3rd, "MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications," *Rapid Commun. Mass Spectrom. RCM*, vol. 18, no. 18, pp. 2162–2168, 2004.

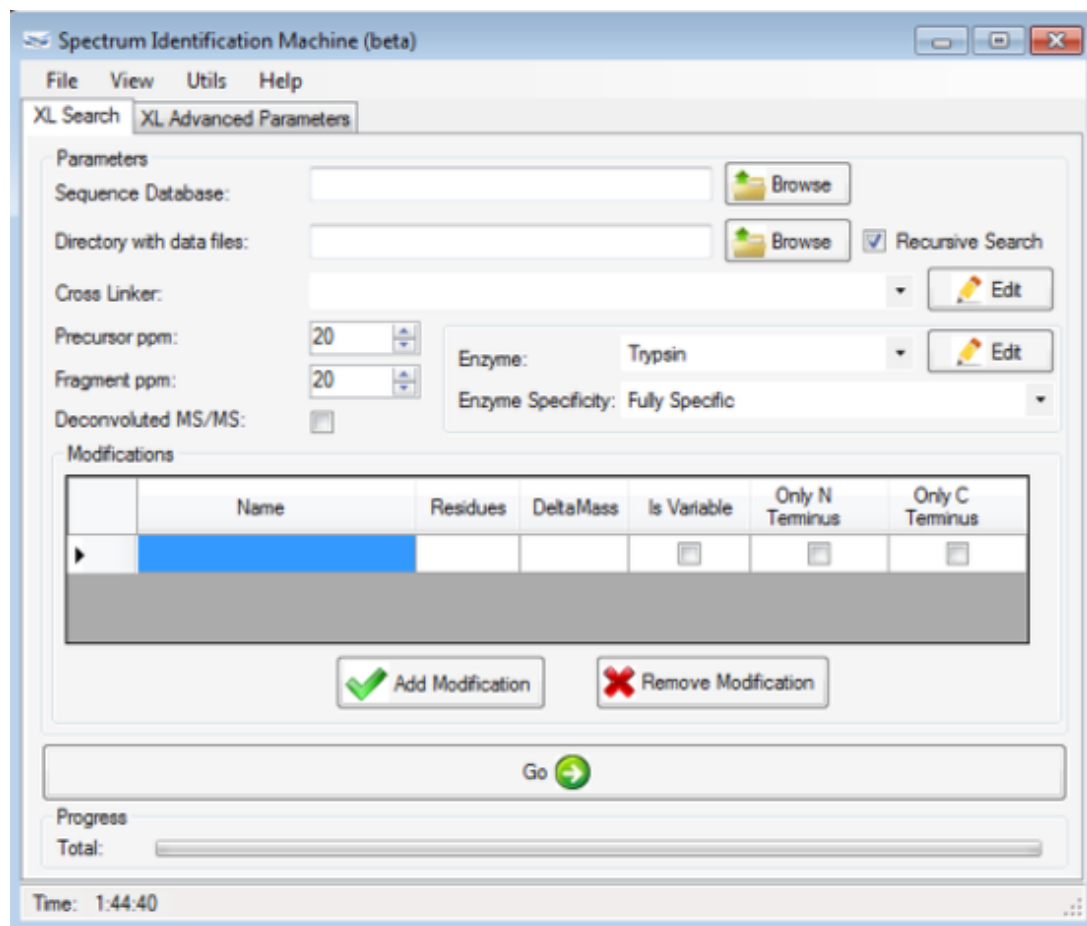
- [10] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, Dec. 1999.
- [11] J. A. Vizcaíno, R. G. Côté, A. Csordas, J. A. Dianes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O'Kelly, A. Schoenegger, D. Ovelleiro, Y. Pérez-Riverol, F. Reisinger, D. Ríos, R. Wang, and H. Hermjakob, "The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D1063–1069, Jan. 2013.
- [12] J. K. Eng, T. A. Jahan, and M. R. Hoopmann, "Comet: an open-source MS/MS sequence database search tool," *Proteomics*, vol. 13, no. 1, pp. 22–24, Jan. 2013.
- [13] A. H. Iglesias, L. F. A. Santos, and F. C. Gozzo, "Identification of cross-linked peptides by high-resolution precursor ion scan," *Anal. Chem.*, vol. 82, no. 3, pp. 909–916, Feb. 2010.
- [14] P. C. Carvalho, T. Xu, X. Han, D. Cociorva, V. C. Barbosa, and J. R. Yates, "YADA: a tool for taking the most out of high-resolution spectra," *Bioinformatics*, vol. 25, no. 20, pp. 2734–2736, Oct. 2009.

Acknowledgements

The authors thank FAPESP, FAPERJ, CAPES, Universal CNPq, Microsoft Research – Microsoft Azure Research Award, Programa Estratégico de Apoio à Pesquisa em Saúde (PAPES), and Fundação Oswaldo Cruz for financial support. They also thank the PRIDE Team for working together with them to enable SIM-XL to support the next version of mzIdentML.

Figures

Figure 1: SIM-XL Main GUI



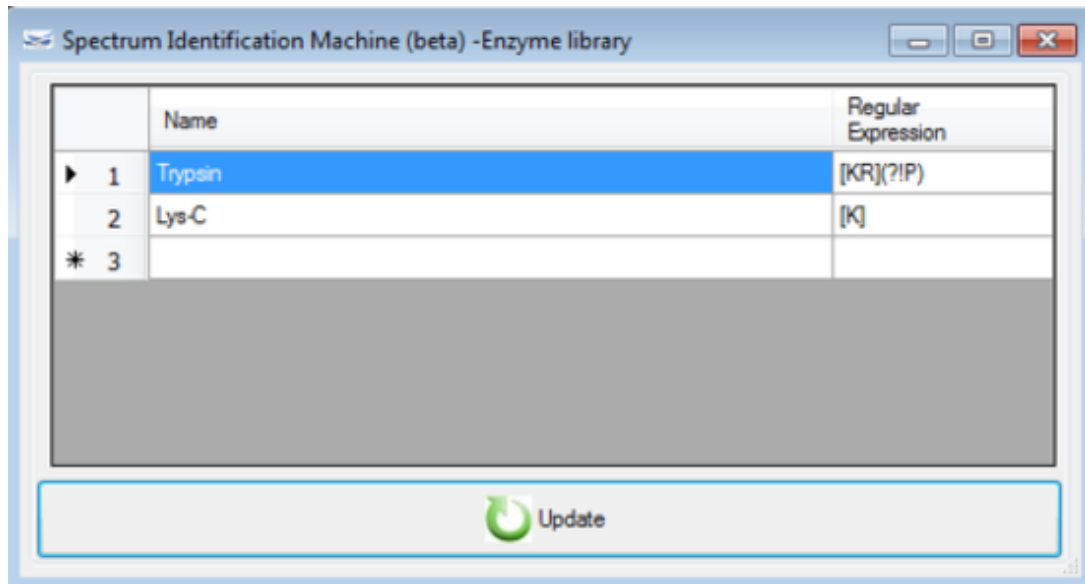
In this window the user will perform the search filling out all fields according to each analysis.

Figure 2: Cross-linker (XL) library

	Name	XL MassShift	Modification MassShift	Reaction Sites	Reporter Ions
▶ 1	Disuccinimidyl Suberate - DSS	138.068100	156.078600	KK	222.149 239...
2	Disuccinimidyl Glutarate - DSG	96.021100	114.031700	KK	180.1025 19...
3	Disuccinimidyl Glutarate - DSSeb	166.099400	184.109900	KK	250.1807 26...
4	Disulfide	-2.015650	0.000000	CC	
5	Zero-length	-18.010600	0.000000	EK DK ...	
* 6					

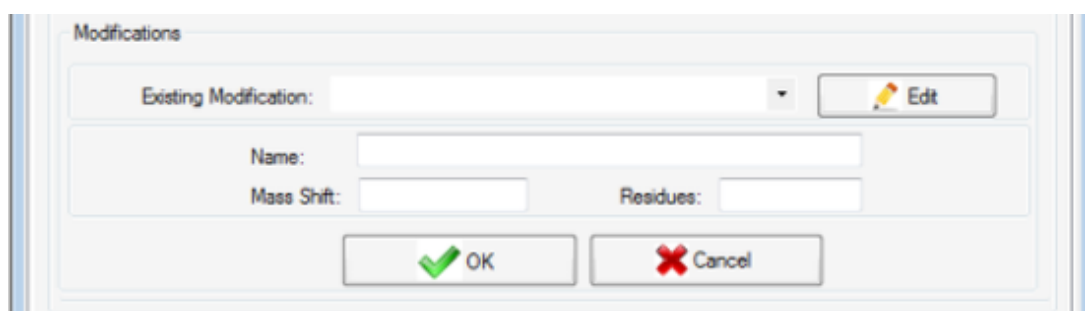
A cross-linker can be inserted or removed in this window.

Figure 3: Enzyme library



An enzyme can be inserted or removed in this window. A regular expression is required to specify the cleavage sites of a new enzyme. For example, the regular expressions for Trypsin and Lys-C are '[KR](?!P)' and '[K]', respectively. For more on building regular expressions we refer the reader to <http://www.regular-expressions.info/>.

Figure 4: Add/Remove Modifications



Select a pre-defined modification or add a new one. To edit a pre-defined modification, click on the Edit button.

Figure 5: SIM-XL Advanced Parameters

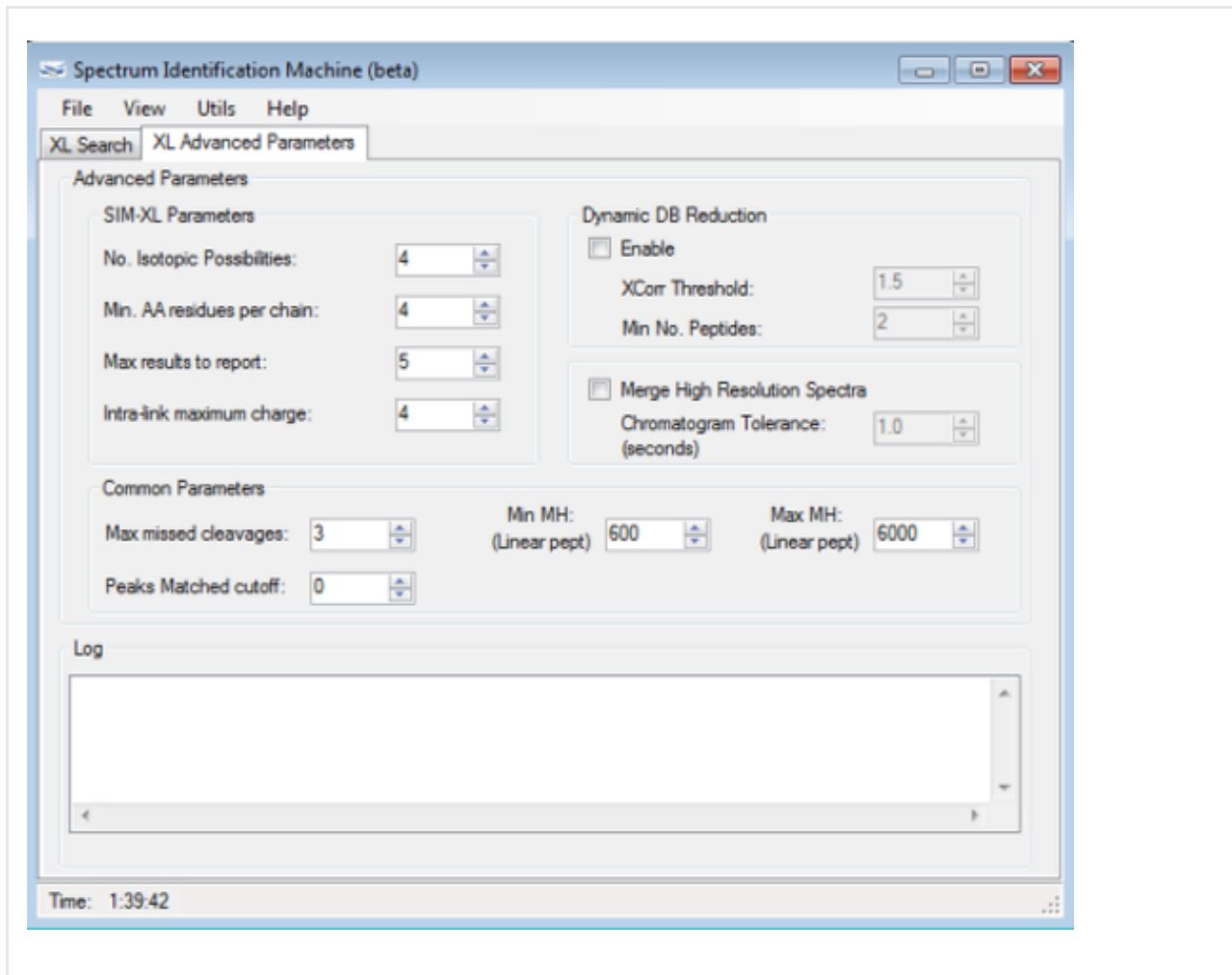


Figure 6: Save or Load SIM-XL parameters

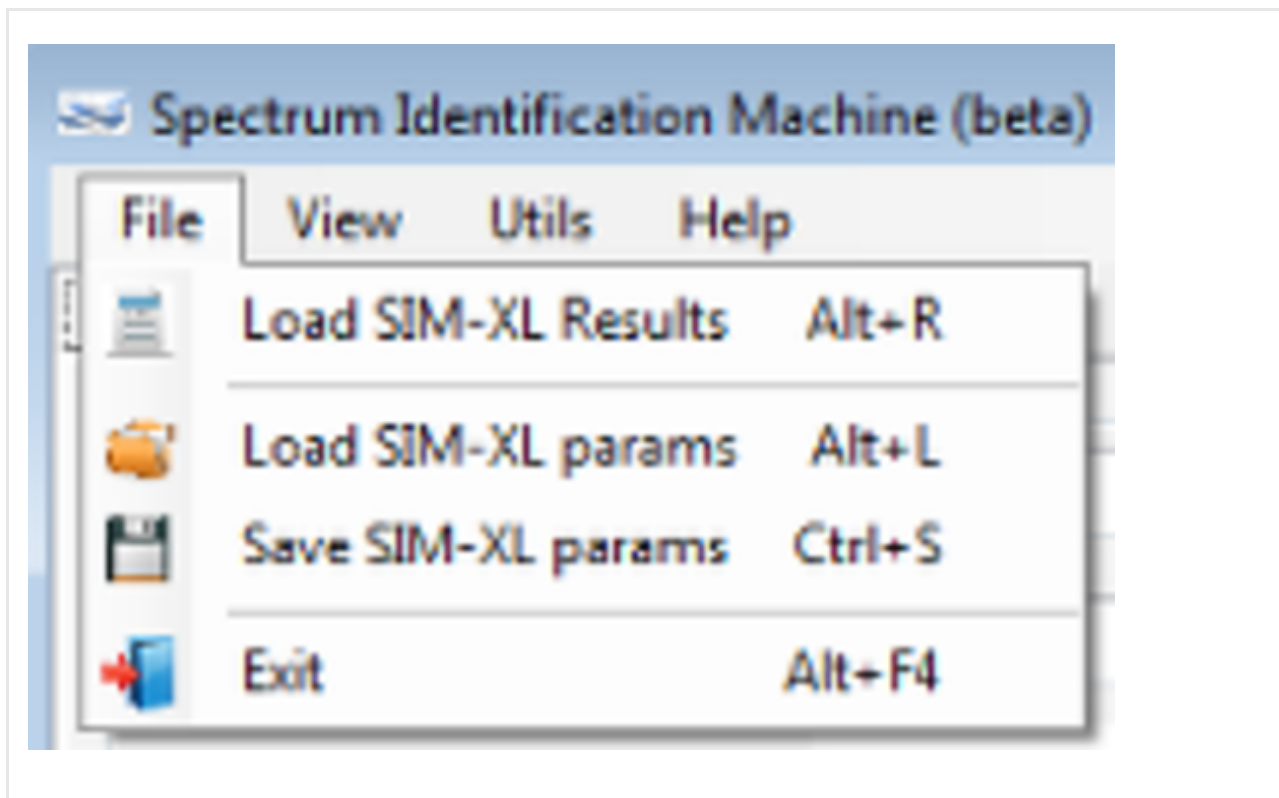
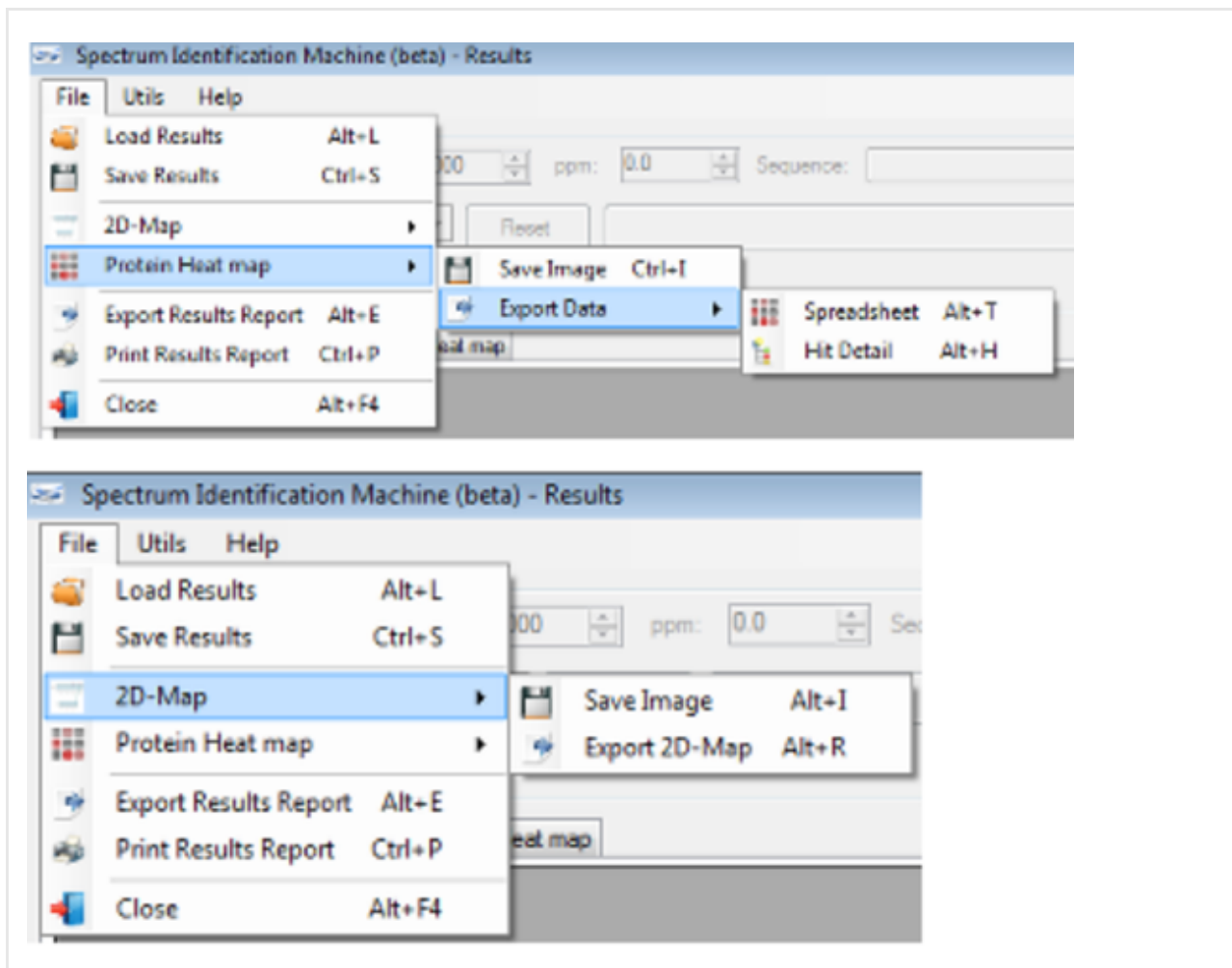


Figure 7: Menu File of results



Here the user can access many features, such as loading or saving the search results, exporting the 2D-Map to an image or PDF file, exporting the Protein Heat map to an image or Excel® (XLS) file, and printing or exporting the results to a spreadsheet with all protein interactions and their hit details.

Figure 8: Protein-protein interactions Map (2D-Map)

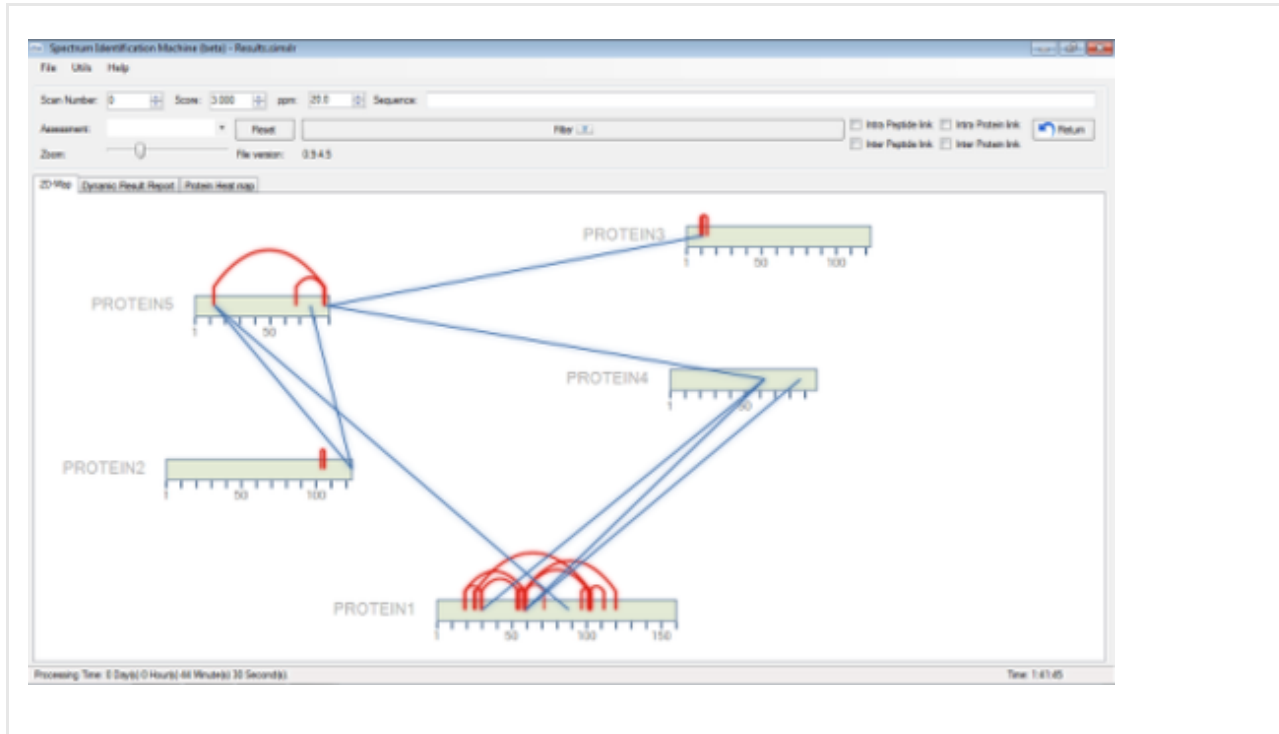


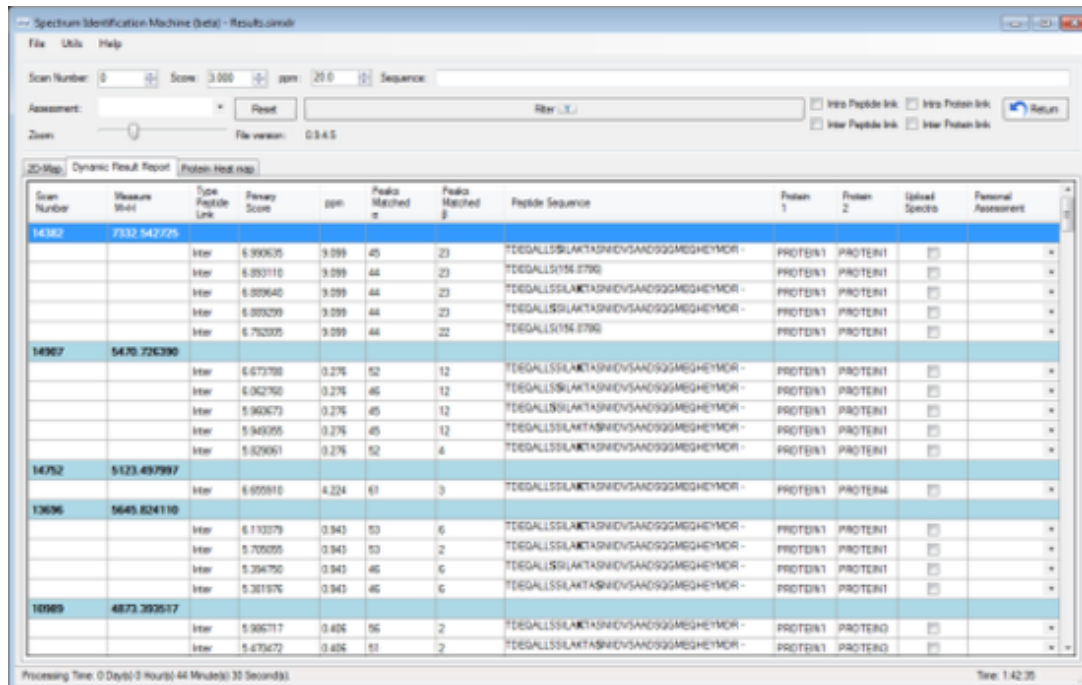
Figure 9: Input file window



SIM-XL accepts the mzIdentML format, in addition to its

own format (simxlr).

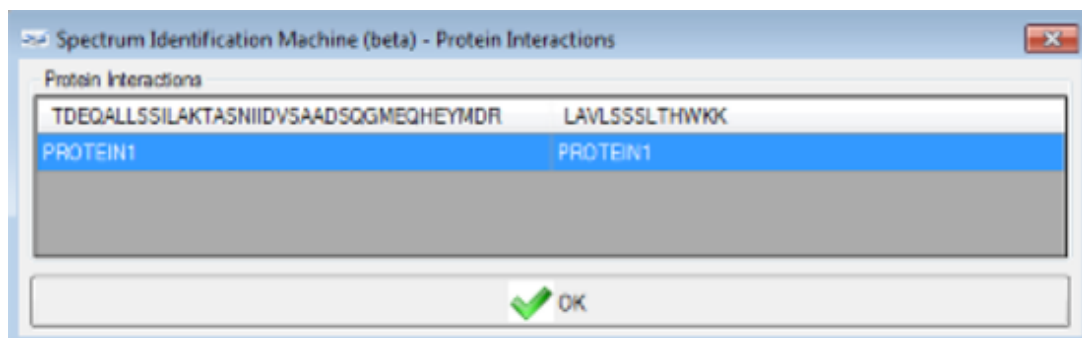
Figure 10: Dynamic Result Report



The screenshot shows the 'Dynamic Result Report' window of the Spectrum Identification Machine (beta). The window displays a table with the following columns: Scan Number, Masses (kDa), Type Peptide Link, Primary Score, ppm, Peptide Matched α, Peptide Matched β, Peptide Sequence, Protein 1, Protein 2, Upload Spectra, and Personal Assessment. The table lists several scan results, including scan numbers 14382, 14907, 14752, 13096, and 10989, each with associated mass values, scores, and peptide sequences. The peptide sequences are variations of TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR and LAVLSSSLTHWKK. The window also shows a 'Processing Time' of 0 Days 0 Hours 44 Minutes 30 Seconds and a 'Time' of 1:42:35.

Scan Number	Masses (kDa)	Type Peptide Link	Primary Score	ppm	Peptide Matched α	Peptide Matched β	Peptide Sequence	Protein 1	Protein 2	Upload Spectra	Personal Assessment
14382	7332.542725	Inter	5.99035	3.099	45	23	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.89110	3.099	44	23	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.89040	3.099	44	23	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.89329	3.099	44	23	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.76205	3.099	44	22	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
14907	5470.726390	Inter	5.67398	0.276	52	12	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.60290	0.276	46	12	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.96073	0.276	45	12	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.94355	0.276	45	12	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.82901	0.276	52	4	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
14752	5123.497997	Inter	6.69910	4.224	61	3	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN4	<input type="checkbox"/>	<input type="checkbox"/>
13096	5645.824110	Inter	5.13079	0.943	53	6	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.70595	0.943	53	2	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.39470	0.943	46	6	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.38197	0.943	46	6	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN1	<input type="checkbox"/>	<input type="checkbox"/>
10989	4873.392517	Inter	5.99517	0.406	56	2	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN3	<input type="checkbox"/>	<input type="checkbox"/>
		Inter	5.47072	0.406	53	2	TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR -	PROTEIN1	PROTEIN3	<input type="checkbox"/>	<input type="checkbox"/>

Figure 11: Protein interactions window

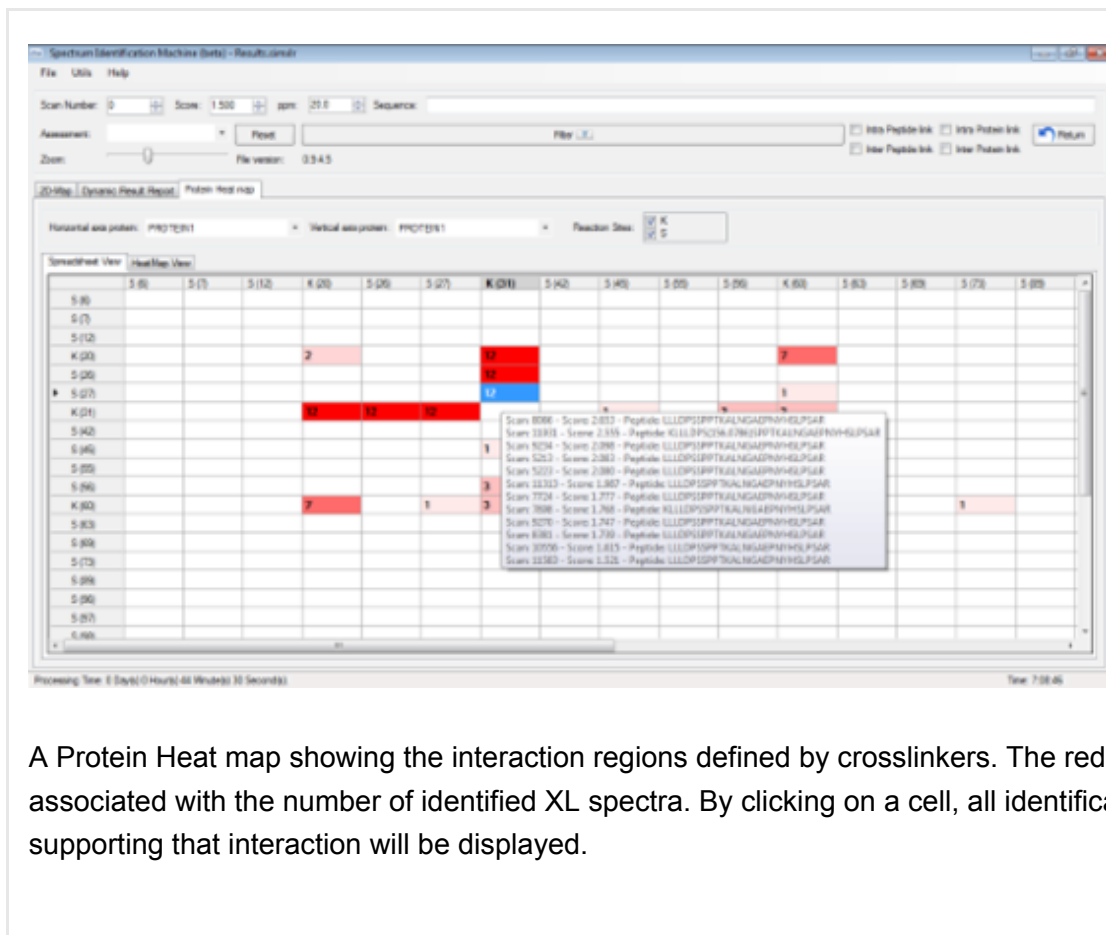


The screenshot shows the 'Protein Interactions' window of the Spectrum Identification Machine (beta). The window displays a table with two columns: the identified peptide sequence and the protein name. The header displays the identified peptides: TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR and LAVLSSSLTHWKK. Below, the protein that contains these sequence(s) (PROTEIN1) is listed. The window also shows a green checkmark and the text 'OK'.

Identified Peptides	Protein
TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR	PROTEIN1
LAVLSSSLTHWKK	PROTEIN1

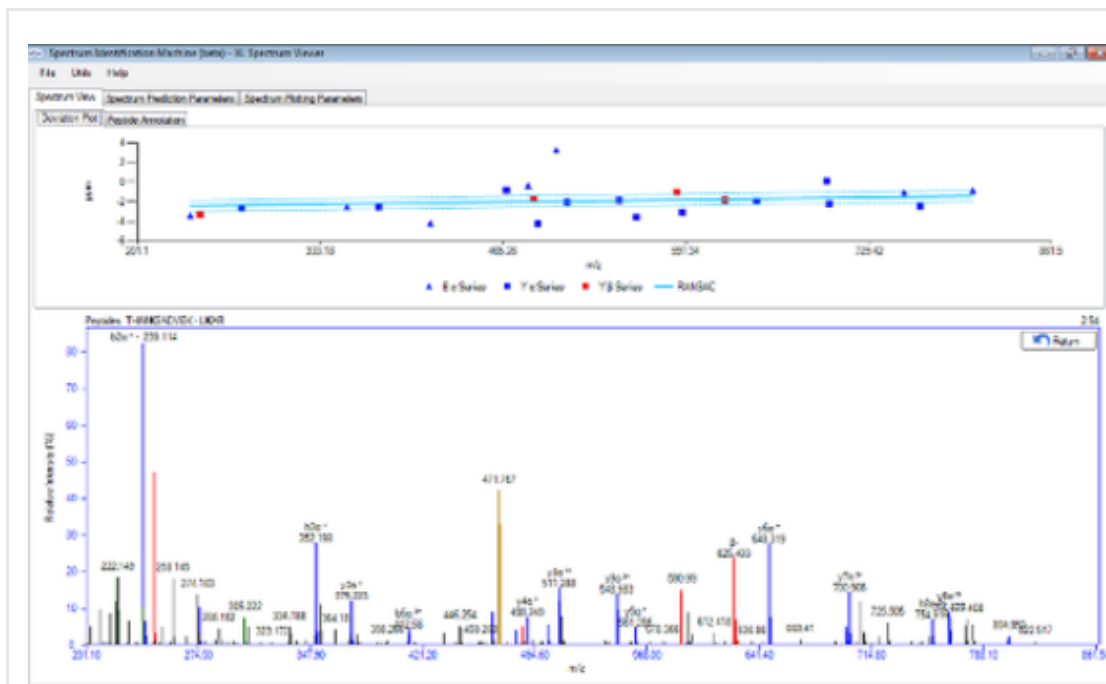
Proteins containing identified peptides window. The header displays the identified peptides: TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR and LAVLSSSLTHWKK. Below, the protein that contains these sequence(s) (PROTEIN1) is listed.

Figure 12: Protein interactions map



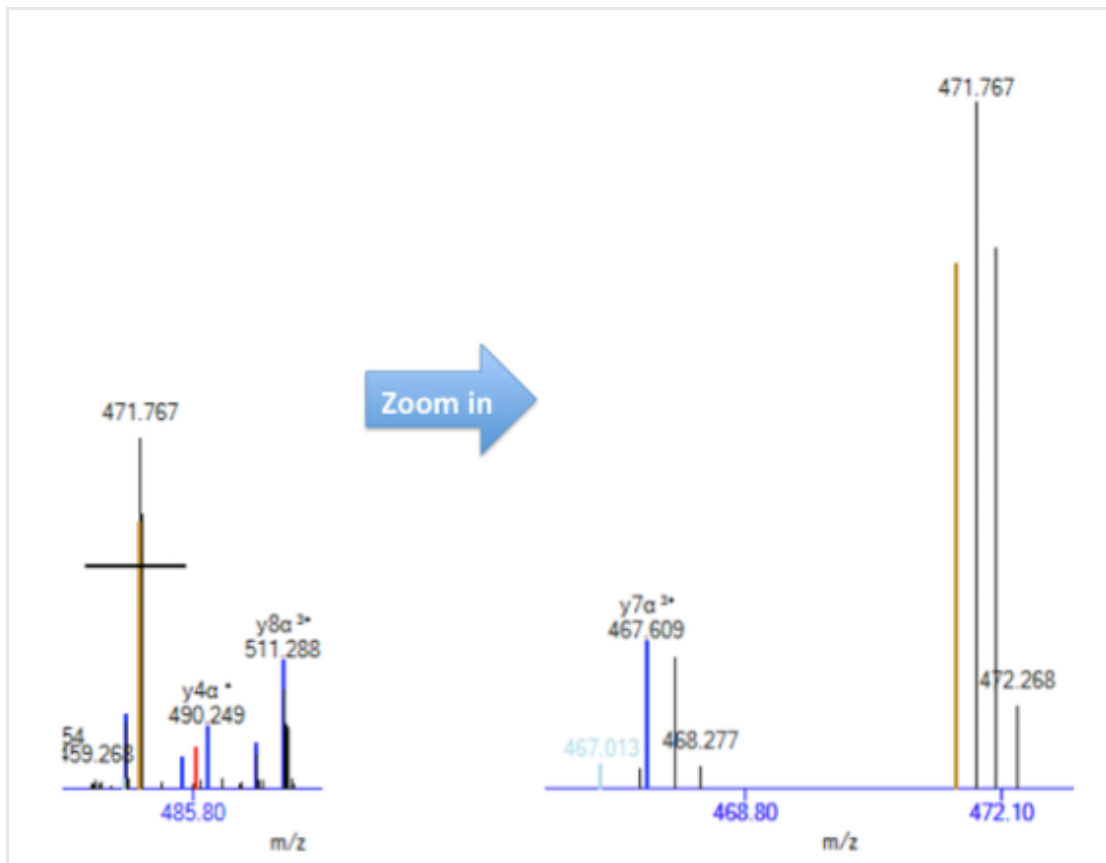
A Protein Heat map showing the interaction regions defined by crosslinkers. The red scale is associated with the number of identified XL spectra. By clicking on a cell, all identifications supporting that interaction will be displayed.

Figure 13: XL Spectrum Viewer



The Spectrum View tab allows the user to browse the spectrum, zoom in and out, as well as easily view which peaks were attributed to which series. A ppm deviation plot is available above the annotated mass spectrum.

Figure 14: Spectrum zoom-in



Zoom-in on a specific area of the XL mass spectrum

Figure 15: Peptide Annotation



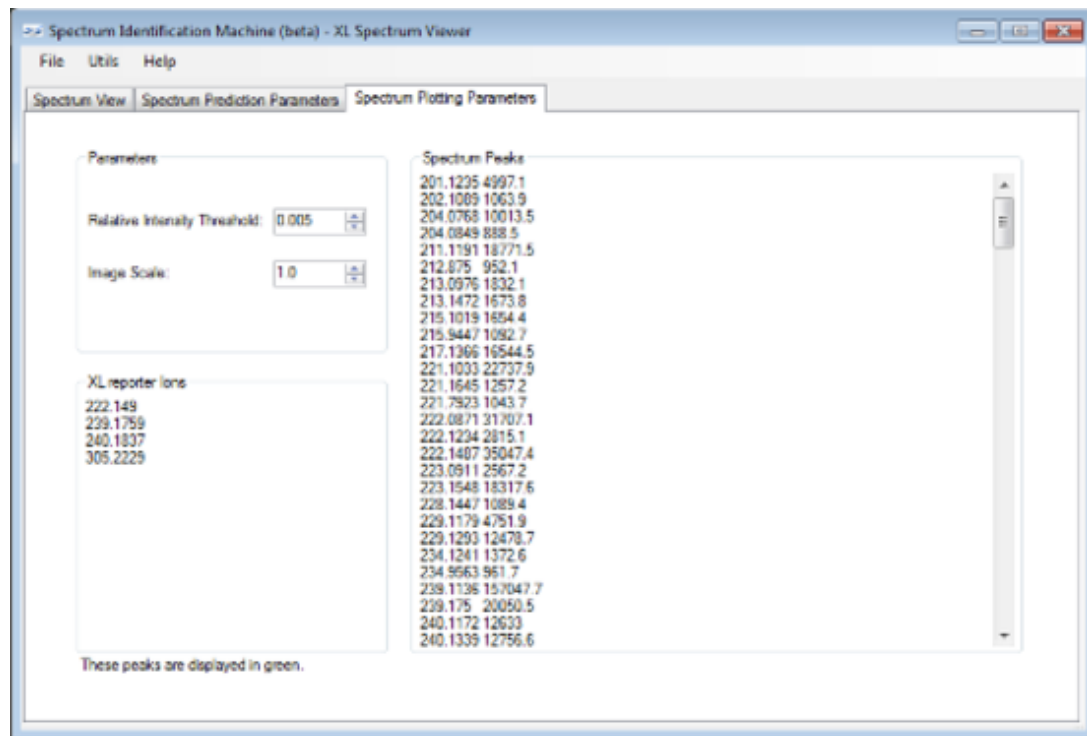
The Peptide Annotation tab displays the fragmentation diagram of the cross-linked peptides.

Figure 16: Spectrum Prediction Parameters tab

Peptide Sequence 1: THIMKSADVEK
 Peptide Sequence 2: LKAR
 Position XL 1: 5 Position XL 2: 2 Matching peaks / Experimental peaks: 27 / 221 = 12.2%

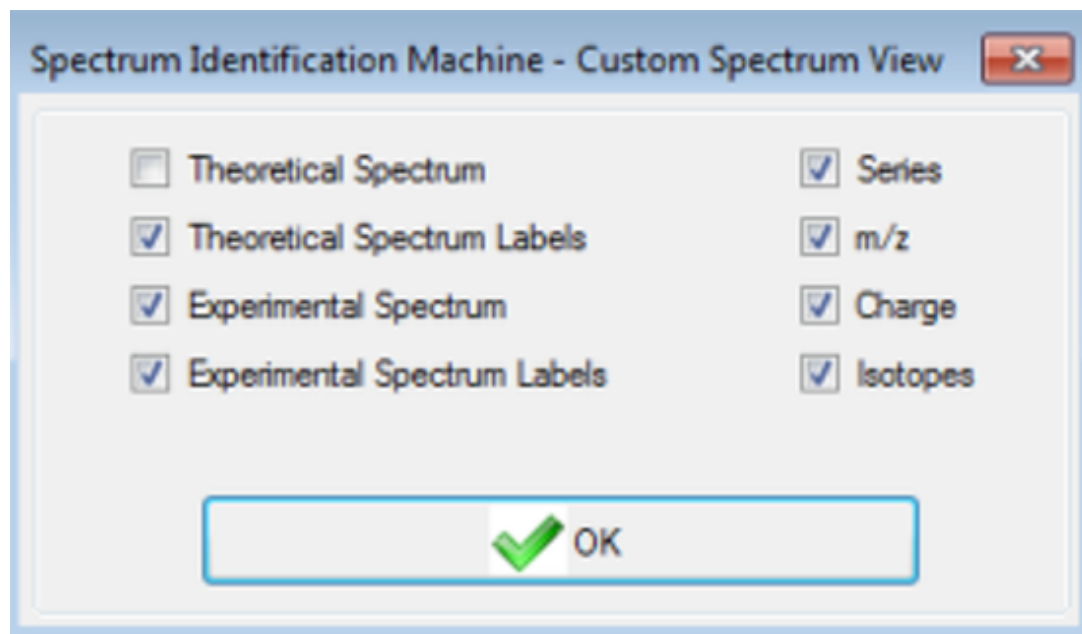
Charge	m/z	Series	FinalAA	Matched	Number
2	804.4511	B_Afa	V	True	9
2	766.4298	Y_Afa	M	True	8
2	754.9169	B_Afa	D	True	8
1	741.4081	A	A	True	7
2	700.9096	Y_Afa	K	True	7
2	690.8612	AlphaXL_Chain	α-	True	0
1	648.3204	Y_Afa	S	True	6
3	628.3523	Precursor		True	1
1	625.4037	BetaXL_Chain	β-	True	0
3	594.6701	Y_Afa	H	True	10
3	590.658	Y_Beta	K	True	3
1	561.2884	Y_Afa	A	True	5
3	548.9638	Y_Afa	I	True	9
3	511.2091	Y_Afa	M	True	8

The user can adjust parameters to check assignments.

Figure 17: Spectrum Plotting Parameters tab

The user can add individual mass-spectral peak lists to visualize the spectral assignments.

Figure 18: Custom Spectrum View



Spectrum annotations can be customized by checking the option in this menu.

Associated Publications

This protocol is related to the following articles:

-

Author information

Affiliations

Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

Diogo B. Lima & Paulo C. Carvalho

Dalton Mass Spectrometry Laboratory, University of Campinas, São Paulo, Brazil

Tatiani B. de lima & Fabio C. Gozzo

College of Agricultural and Veterinary Sciences, State University of São Paulo, Jaboticabal, São Paulo, Brazil

Tiago S. Balbuena

Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil

Ana Gisele C. Neves-Ferreira

**Systems Engineering and Computer Science Program, Federal
University of Rio de Janeiro, Rio de Janeiro, Brazil**

Valmir C. Barbosa

Competing financial interests

The authors declare no conflicting financial interests.

Corresponding author

Correspondence to: Diogo B. Lima (diogobor@gmail.com) Fabio C. Gozzo
(fabio@iqm.unicamp.br) Paulo C. Carvalho (paulo@pcarvalho.com)

Readers' Comments

Comments on this thread are vetted after posting.

Protocol Exchange ISSN 2043-0116

© 2015 Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

partner of AGORA, HINARI, OARE, INASP, CrossRef and COUNTER

An Evaluation of the Crystal Structure of C-terminal Truncated Apolipoprotein A-I in Solution Reveals Structural Dynamics Related to Lipid Binding

John T. Melchior^a, Ryan G. Walker^b, Jamie Morris^a, Martin K. Jones^c, Jere P. Segrest^c, Diogo B. Lima^d, Paulo C. Carvalho^d, Fábio C. Gozzo^e, Mark Castleberry^b, Thomas B. Thompson^b, and W. Sean Davidson^a

From the: ^a Department of Pathology and Laboratory Medicine, University of Cincinnati, Cincinnati, Ohio 45237; ^b Department of Molecular Genetics, Biochemistry and Microbiology, University of Cincinnati, Cincinnati, Ohio 45237; ^c Department of Medicine and Atherosclerosis Research Unit, University of Alabama at Birmingham, Birmingham, Alabama 35294; ^d Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil 81350-010 and the ^e Dalton Mass Spectrometry Laboratory, University of Campinas, São Paulo, Brazil

*Running title: *Solution Structure of Truncated ApoA-I*

To whom correspondence should be addressed: Davidson: Department of Pathology and Laboratory Medicine, University of Cincinnati, 2120 Galbraith Rd., Cincinnati, Ohio 45237-0507 USA, Tel.: (513) 558-3707; Fax: (513) 558-1312; E-mail: Sean.Davidson@UC.edu or Thompson: Department of Molecular Genetics, Biochemistry and Microbiology, University of Cincinnati, Cincinnati OH 45237 USA, Tel.: 513-558-4517; E-mail: Tom.Thompson@uc.edu.

Keywords: apolipoprotein, structural model, structural biology, small-angle X-ray scattering (SAXS), mass spectrometry (MS), oligomerization

ABSTRACT

Apolipoprotein (apo)A-I mediates many of the anti-atherogenic functions attributed to high density lipoprotein (HDL). Unfortunately, efforts toward a high-resolution structure of full-length apoA-I have not been fruitful, though there have been successes with deletion mutants. Recently, a C-terminal truncation (apoA-I^{A185-243}) was crystallized as a dimer. The structure showed two helical bundles connected by a long, curved pair of swapped helical domains. To compare this structure to that existing under solution conditions, we applied small angle X-ray scattering and isotope-assisted chemical cross-linking to apoA-I^{A185-243} in its dimeric and monomeric forms. For the dimer, we found evidence for the shared domains and aspects of the N-terminal bundles, but not the molecular curvature seen in the crystal. We also found that the N-terminal bundles equilibrate between open and closed states. Interestingly, this movement is one of the transitions proposed during lipid binding. The monomer was consistent with a model in which the long shared helix doubles back onto the helical bundle. Combined with the crystal structure, these data offer an important starting point to understand the molecular details of HDL biogenesis.

Apolipoprotein (apo)A-I is the most common protein constituent of human high density lipoprotein (HDL), comprising up to 70% of protein mass. As such, it is often credited with defining the functionality of HDL as related to its proposed cardioprotective benefits. For example, apoA-I dramatically stimulates lecithin:cholesterol acyl transferase (LCAT) which esterifies cholesterol to create a concentration gradient that promotes free cholesterol movement from peripheral cells to HDL (1). ApoA-I also interacts with the ATP binding cassette transporter A1 in the liver and periphery (2-5), an event that produces the bulk of circulating HDL. Given the recent revelations that HDL also contains greater than 90 “minor” proteins (HDL Proteome Watch <http://homepages.uc.edu/~davidswm/HDLproteome.html>), apoA-I is likely an important scaffold that coordinates these factors to affect functions related to lipid metabolism, inflammation, innate immunity and more (6).

Since the first descriptions of amphipathic helices inferred in the sequence of apoA-I in the 1970's (7,8), considerable effort has been put into understanding apoA-I structure in its lipid-free form or when bound to lipid in HDL. With its flexibility and propensity to interact with lipid or itself (oligomerization), it has not yet been possible

to apply traditional high-resolution structural techniques such as nuclear magnetic resonance or X-ray crystallography to full-length, wild-type apoA-I. However, there has been success using deletion mutants. Borhani et al. crystallized human apoA-I lacking the N-terminal 43 amino acids (9). The structure showed a ring-like tetramer with amphipathic helical domains coiled around each other. This was thought to reflect the lipid-bound form and became the basis for the double belt model of apoA-I (10). However, information about the monomeric, lipid-free form remained elusive but highly sought after - even prompting the publication of a falsified structure (11).

Recently, Mei and colleagues published a crystal structure of a mutant lacking the C-terminal 60 amino acids, apoA-I $^{\Delta 185-243}$ (12). It showed a curved dimer in which two four helical bundles were connected by a pair of long antiparallel helices. This strongly resembled a structure of a dimeric apoA-IV mutant that we reported (13). These “helix swapped” models are attractive in that they offer clear predictions of how: *i*) a four helical bundle can transition to the widely accepted double belt orientation upon lipid binding, and *ii*) the dimer can transition to monomeric or trimeric forms (14).

Since apolipoproteins exhibit significant conformational flexibility, we investigated the structure of the truncated apoA-I $^{\Delta 185-243}$ in both its dimeric and monomeric forms in solution to draw comparisons to the crystal structure. We used the lower resolution, but solution-based, techniques of chemical cross-linking and small angle X-ray scattering (SAXS). Our results confirm some of the general features of the Mei crystal structure including the long antiparallel helices participating in the domain swap, but we note important differences in the flexibility of the N-terminal domain and the overall shape of the molecule.

EXPERIMENTAL PROCEDURES

ApoA-I $^{\Delta 185-243}$ Protein Expression and Purification - Our previously reported construct for recombinant, full-length apoA-I (15) was modified with a stop codon (TAG) after Asn-184. The mutant apoA-I $^{\Delta 185-243}$ was then expressed and purified as described (15). pET30 vectors (Novagen) containing mutant apoA-I $^{\Delta 185-243}$ were transformed into BL-21 *E. coli* cells. Cells were grown at 37°C in Luria-Bertani culture media containing kanamycin for selection of pET30 transformants. Protein expression was induced by addition of

isopropyl β -D-thiogalactopyranoside (0.1 mM) followed by shaking at 225 RPM for 2 h at 37°C. Cells were pelleted by centrifugation, supernatant was discarded, and cells were resuspended in binding buffer (5 mM imidazole, 500 mM NaCl, 20 mM Tris-HCl, pH 7.9) and lysed at 4°C by probe sonication. The cell lysate was pelleted and the supernatant was applied to His-bind columns. Fractions containing apoA-I $^{\Delta 185-243}$ were pooled and the His-tag was cleaved by tobacco etch virus (TEV) protease at a mass ratio of 20 (apoA-I):1(TEV) for 2 h at room temperature. The sample was reapplied to the His-bind resin to remove the His-tag and fractions containing apoA-I $^{\Delta 185-243}$ were pooled, concentrated and dialyzed into 10 mM NH₄HCO₃ pH 8.1, lyophilized to dryness, and stored at -80°C until ready for use. The recombinant protein contained an additional glycine at the N-terminus after cleavage of the his-tag by TEV. The sequence was confirmed by the Cincinnati Children’s Hospital Sequencing Core. ¹⁵N-labeled apoA-I $^{\Delta 185-243}$ was generated as previously described (16).

¹⁴N and ¹⁵N versions of apoA-I $^{\Delta 185-243}$ were solubilized individually in 3 M guanidine in Tris-HCl. The ¹⁴N and ¹⁵N apoA-I $^{\Delta 185-243}$ were mixed at a 1:1 molar ratio at 37°C for 1 h with intermediate vortexing. Protein was refolded with exhaustive dialysis against PBS, pH 7.4. Where indicated, apoA-I $^{\Delta 185-243}$ was not subjected to denaturation (i.e. left in solution) to characterize any differences that may occur in the denaturation/reassembly process. All protein preparations were further purified using size exclusion chromatography using a single Superdex 200 gel filtration column (10/300 GL; GE Healthcare) on a ÄKTA FPLC system (GE Healthcare) in PBS to ensure removal of any bacterial protein contamination. Appropriate fractions were pooled and concentrated by ultrafiltration for cross-linking and SAXS. Protein concentration was determined by the Markwell modified Lowry assay (17). Purity was routinely >95% as determined by SDS-PAGE and mass spectrometry.

Cross-linking - All proteins were cross-linked with bis-(sulfosuccinimidyl) suberate (BS³) (Thermo Scientific) as previously described (16). Protein was cross-linked at a molar ratio of 50:1 cross-linker to protein for 12 h at 4°C and quenched by excess Tris-HCl. Cross-linked monomeric and dimeric apoA-I $^{\Delta 185-243}$ were separated by gel filtration chromatography on three Superdex 200

gel filtration columns (10/300 GL; GE Healthcare) in series on a ÄKTA FPLC system (GE Healthcare) in PBS at a flow rate of 0.3 ml/min. 0.25 ml fractions were collected and analyzed by SDS PAGE. Fractions containing the pure monomeric and dimeric apoA-I^{Δ185-243} were pooled and concentrated. Samples undergoing MS analysis were dialyzed into 10 mM NH₄HCO₃ pH 8.1. 50 μg aliquots were digested with sequencing grade trypsin (Promega) at 1:20 mass ratio of trypsin to protein for 16 h at 37°C. Peptides cross-linked with BS³ were lyophilized to dryness and stored at -20°C until MS analysis. Crosslinking experiments were performed on two independent preparations of protein.

Mass Spectrometry and Identification of Cross-linked Peptides - Mass spectrometry analyses were performed as previously described (16). Nano-LC-MS/MS analyses were performed on a TripleTOF[®] 5600+ (AB Sciex, Toronto, Canada) coupled to an Eksigent (Dublin, CA) NanoLC-Ultra[®] nanoflow system. Dried samples were reconstituted in formic acid/H₂O 0.1/99.9 (v/v), and 5 μl (~1-2 μg of digest) was loaded onto C18 IntegraFrit[™] trap column (New Objective, Inc.) at 2 μl/min in FA/H₂O 0.1/99.9 (v/v) for 15 min to desalt and concentrate the samples. For the chromatographic separation, the trap-column was switched to align with the analytical column, Acclaim[®] PepMap100 (Dionex-Thermo Fisher Scientific). Peptides were eluted at 300 nl/min using a varying mobile phase gradient from 95% phase A (FA/H₂O 0.1/99.9, v/v) to 40% phase B (FA/ACN 0.1/99.9 v/v) for 35 min (1% per min), then from 40% B to 85% B in 5 min with re-equilibration. Effluent was introduced to the mass spectrometer using a NANOSpray[®] III Source (AB Sciex, Toronto, Canada). The instrument was operated in positive ion mode for 65 min, where each cycle consisted of one TOF-MS scan (0.25 s accumulation time, in a 350 to 1500 m/z window) followed by 30 information dependent acquisition mode MS/MS-scans on the most intense candidate ions selected from initially performed TOF-MS scan during each cycle. Each product ion scan had an accumulation time of 0.075 s and CE of 43 with an 8 unit scan range. The .wiff files were converted to Mascot generic files using PeakView[®] v1.2.0.3 software (AB Sciex).

Mascot generic files were loaded into the SIM-XL search engine (18) for cross-linked peptides link version 1.1. Briefly, this latest version is optimized for characterizing interaction between

homodimers by allowing the user to specify which proteins(s) in the sequence database have light (e.g., ¹⁴N) and heavy (e.g., ¹⁵N) versions. The carbamidomethylation of the cysteine and the BS³ cross-linker mass modification of 138.0681 at the N-terminus and lysine was considered as fixed. A tolerance of 20 ppm was accepted at the MS1 and MS2 levels. All initial identification of cross-linked peptides required a SIM-XL primary score greater than 1.5. As a single incorrect cross-link identification may lead to an erroneous model, a manual post-validation of the search engine results, at the MS/MS level, was independently performed by two experienced analysts.

Small-Angle X-ray Scattering - SAXS data was collected using the SIBYLS beamline (Berkeley, CA) (19). Cross-linked, monomeric and dimeric apoA-I^{Δ185-243} were separated by gel filtration chromatography as described above. Purified samples were shipped overnight at 4°C for SAXS data collection within 24 hours of isolation to avoid aggregation artifacts. Three concentrations of the purified monomeric and dimeric apoA-I^{Δ185-243} in PBS were sampled at 10°C with four exposure times; 0.5, 1.0, 2.0, and 5.0 s. Scattering profiles from samples suffering radiation damage were discarded. ScÅtter (SIBYLS) and ATSAS program suite (EMBL) were used for data analysis. 20 independent *ab initio* molecular envelope reconstructions were generated using the online DAMMIF server (EMBL-Hamburg) (20). The envelopes were superimposed and averaged using SUPCOMB and DAMAVER (ATSAS, EMBL-Hamburg). The averaged molecular envelope graphics were rendered using UCSF Chimera.

Model Generation and Evaluation 3-D composite models were generated based on the dimeric crystal structure of apoA-I^{Δ185-243} [Protein Data Bank (PDB) entry 3R2P]. The structure was manually manipulated in Pymol guided by experimentally derived cross-links and SAXS data. Modeller v9.14 (21) was used to perform a sequence alignment and generate 100 iterations of each starting model. Models were constrained with identified cross-links with an upper bound of 26.0 ± 0.001 Å from C-α to C-α. The starting model for reported structures were chosen based on the best-fit to experimental SAXS data using FoXS (22,23). An energy minimization was performed on the initial structure using the AllosMod-FoXS web server (23,24) to generate 3000 intermediate conformations consistent with the input structure

using a temperature scan (300K) (16). The final models were presented based on satisfaction of all cross-linking constraints and best-fit to the experimental SAXS scattering profile.

RESULTS

Concept of Isotope Assisted Cross-linking—A problem with using chemical crosslinking to understand the structure of a homodimer is the inability to distinguish between intra- and intermolecular cross-links. In previous work with apoA-IV, we solved this problem by isotopically labeling one polypeptide of the dimer with ^{15}N , leaving the other with naturally occurring ^{14}N (16). The two forms were expressed in bacteria grown with the appropriate nitrogen isotope, denatured in guanidine HCl, mixed together at a 1:1 molar ratio, and then allowed to refold into dimers with mixed isotopic species. We pursued the same strategy with apoA-I $^{\Delta 185-243}$ (see *Methods*). **Fig. 1** illustrates the concept. Intermolecular cross-links between two peptides (A and B) result in four possible mass combinations depending on the isotopic makeup of the two species in a dimer: ^{14}A to ^{14}B , ^{15}A to ^{15}B , ^{14}A to ^{15}B , and ^{15}A to ^{14}B and a MS spectrum containing four sets of peaks (top panel). Alternatively, intramolecular cross-links only have two possibilities, ^{14}A to ^{14}B and ^{15}A to ^{15}B , resulting in two sets of peaks (bottom panel).

Oligomerization Properties of ApoA-I $^{\Delta 185-243}$ —**Fig. 2a** contrasts WT apoA-I (28 kDa) with apoA-I $^{\Delta 185-243}$ (22 kDa). The high purity of both preparations is apparent on the SDS gel and by direct injection MS of a mixture of both ^{14}N apoA-I $^{\Delta 185-243}$ and ^{15}N apoA-I $^{\Delta 185-243}$ (**Fig. 2b**). ^{14}N apoA-I $^{\Delta 185-243}$ exhibited an experimental molecular mass of 21624 Da, nearly identical to its theoretical mass of 21623.96 Da. ^{15}N apoA-I $^{\Delta 185-243}$ had an experimental molecular mass of 21890 Da compared to a theoretical 21894.25 Da. From both direct injection MS and peptide analyses, we determined that our labeling efficiency was $> 95\%$.

We observed that unfolding the protein in guanidine HCl affected the relative monomer/dimer distribution of the truncation mutant (**Fig. 3**). When ^{14}N apoA-I $^{\Delta 185-243}$ was isolated directly from the bacteria without a refolding step, the preparation contained about 60% monomer and 40% dimer by gel filtration chromatography at room temperature. However, denaturation in guanidine and refolding via dialysis resulted in a shift to the monomeric state (~90%)

with only about 10% dimer. This agrees with a previous characterization of this mutant (25). This may result from concentration differences between compartments within the bacteria and bulk solution *in vitro*. Since it is not possible to perform the dual isotope technique with non-refolded protein, we scaled up our separations in order to isolate enough dimer for the current cross-linking studies.

Cross-Linking—We used the homobifunctional cross-linker BS³, an NHS ester with a preference for lysines within its spacer arm length of 12 Å. However, it can also react with serine at a lower frequency (26,27). Mixed ^{14}N and ^{15}N apoA-I $^{\Delta 185-243}$ was cross-linked yielding the monomeric and dimeric species shown in **Fig. 4a**, Lane 1. Monomer and dimer were then isolated by gel filtration chromatography (**Fig. 4b** lanes 1 and 2, respectively). The samples were subjected to tryptic digestion and MS was used to identify the cross-links. In total, 29 cross-linked peptide pairs were identified in both monomeric and dimeric apoA-I $^{\Delta 185-243}$; intramolecular cross-links are listed in **Table 1** and intermolecular links are in **Table 2**. As expected, all cross-links in the monomer sample showed the two mass peak pattern described in **Fig. 1** and exemplified in **Fig. 4c** and **4d** (blue) indicating intramolecular span. The dimer sample contained a mixture of intra and intermolecular cross-links (the four peak pattern). 13 of the identified cross-links were shared between monomer and dimer. Monomeric apoA-I $^{\Delta 185-243}$ had 9 unique cross-links while dimeric apoA-I $^{\Delta 185-243}$ had 7 unique cross-links. This indicates that the dimer and monomer structures are related, but not identical. There were several examples of cross-links that were intramolecular in the monomer, but intermolecular in the dimer, despite linking the same Lys or Ser residues. This is a signature of a domain swap model of oligomerization.

We also compared the cross-links observed in monomeric apoA-I $^{\Delta 185-243}$ to those reported in the same region for WT apoA-I in solution. Including the current study, there have been 35 total cross-links reported in this region in 4 studies (28-30). This study reports 9 previously observed cross-links and 13 unique cross-links not previously found in WT apoA-I. This may suggest a general similarity in the structural fold of the two proteins, but we caution that the large deletion makes direct comparisons with full length apoA-I difficult.

Small Angle X-ray Scattering—To assess the molecular shape of apoA-I $^{\Delta 185-243}$, we performed

SAXS experiments. Stability experiments demonstrated that the dimer slowly dissociated into monomer after purification by size-exclusion chromatography (not shown). Therefore, to insure the molecular homogeneity of the samples during SAXS data collection, we used BS³ to lock the dimer and monomers into their respective states and separated them by gel filtration as for the cross-linking/MS studies. The SAXS parameters for all samples are listed in **Table 3**. In all samples, the scattered intensity $I(q)$ increased proportionally with sample concentration and the Guinier range (used to calculate the radius of gyration, R_g) showed a linear response at low scattering angles (not shown) indicating high quality data with no evidence of protein aggregation.

Dimer - We compared dimeric apoA-I^{Δ185-243} derived natively from *E. coli* (never denatured) or after denaturation/refolding. The SAXS profiles showed no difference in R_g (Guinier and real space), D_{max} and the two samples were of comparable molecular volumes across all concentrations tested. Furthermore, scattering profiles and the pairwise distribution plots of both dimers were nearly identical (**Fig. 5a** and **5b**, respectively) indicating similar structures. Dimeric apoA-I^{Δ185-243} had an R_g value of 36.17 ± 0.50 Å which, taken together with the pairwise distribution plot (**Fig. 5b**, violet) suggested an elongated rod-like structure. Transformation of the SAXS data exhibited a plateau in the $q^3 \cdot I(q)$ plot and lack of plateau in the $q^4 \cdot I(q)$ plot (**Fig. 5c** and **5d**, respectively) indicating structures were folded yet contained a flexible domain (31,32). Indeed, the low-resolution envelope generated via *ab initio* reconstructions (**Fig. 5e**) confirmed the elongated, rod-like appearance. Interestingly, although a slight curvature was apparent, there was no evidence of the crescent shape observed in the crystal structure. The normalized spatial discrepancy (NSD), shown in **Table 3**, was reasonable suggesting high quantitative similarity between independent reconstructions.

Monomer - Monomeric apoA-I^{Δ185-243} had R_g values of 23.20 ± 0.26 Å and a parabolic pairwise distribution plot (**Fig. 5b**, tan), consistent with a more globular structure than the dimer. Transformation of the monomer SAXS scattering profile showed a similar plateau in the $q^3 \cdot I(q)$ plot and lack of plateau in the $q^4 \cdot I(q)$ as compared to the dimer indicating a rigid structure with a flexible domain. In agreement with the pairwise distance

distribution profile, the *ab initio* reconstruction of the monomeric envelope (**Fig. 5f**) confirmed the globular appearance having similar width, but half the length of the dimer envelope. Indeed, the monomer envelope yielded a diameter of ~ 78 Å with an average volume of 53100 Å³, roughly half of the dimer (95700 Å; **Table 3**).

Comparison to the Crystal Structure—We used the cross-linking and SAXS experimental data derived in solution to test the predictions from the Mei crystal dimer. **Fig. 6a** shows a contact plot rendered from the crystal structure. Colored areas represent amino acids that fall within 26 Å of each other; i.e. within possible cross-linking distance. The cross-links in **Tables 1** and **2** were placed on the contact plot showing that 13 out of 20 were consistent. The majority of the violations were in the N-terminal region highlighted in **Fig. 6b**.

Next, the crystal structure was compared to the SAXS data using FoXS (22,23). **Fig. 7a** and **7b** shows significant deviation between the theoretical scattering profile and the experimental data with a poor χ of 3.25. We solvated the crystal structure and generated a theoretical pairwise distribution curve using Scatter (**Fig. 7c**, red). The biphasic pattern for the crystal structure clearly deviates from the experimental pairwise distribution curve. Lastly, we superimposed the crystal structure on the molecular envelopes generated from the SAXS experimental data (**Fig. 7d**) revealing large inconsistencies. Taking the cross-link violations with the SAXS discrepancies, we concluded that apoA-I^{Δ185-243} adopts an alternate configuration in solution. Thus, we set out to generate a solution-state model of apoA-I^{Δ185-243}.

Derivation of a Dimeric ApoA-I^{Δ185-243} Solution Model—Since many of the experimental cross-links were consistent with the crystal structure, we used it as a starting model. We used as few manipulations as possible to bring it in line with the cross-linking and SAXS data. Previous studies on discoidal HDL suggest the antiparallel 5/5 helices can form a hairpin (10,33). Implementation of hairpins and juxtaposition to residues 106-116 resolved some of the problematic cross-links and resulted in the overall shortening of the structure imposed by the SAXS data. Focusing on the N-termini, we observed cross-links consistent with the crystal structure (**Fig. 8**, top inset) but also links that strongly suggested an interaction between the N-terminus and middle of the molecule (**Fig. 8** bottom inset). We concluded that: (i) satisfying all cross-

links in a single dimer model was only feasible by collapsing the structure into a compact, spherical globular protein or (ii) the N-terminus is dynamic and moving back and forth from the end to the middle of the molecule and the cross-linker was capturing it in both positions. The elongated structure implied by the SAXS data ruled out possibility #1. Therefore, we generated two initial models differing in the placement of the N-terminus (residues 1-37). The “open” conformation has the N-terminus folded across the ends of the dimer to complete four helical bundles on each end (like the crystal structure). A second “closed” conformation uses the N-terminal minor helix (residues 37-42) as a hinge so the N-terminus can flip over and interact with the middle of the molecule.

Modeller v9.14 was used to generate 100 iterations of each initial structure constrained by experimental cross-links respective to the “open” and “closed” conformations. The best initial model for each conformation was determined using FoXS which gave improved fits ($\chi=1.32$ and $\chi=0.95$ for the “open” and “closed”, respectively) compared to the crystal structure ($\chi=3.25$). We performed an energy minimization on each best-fit conformation and then used AllosMod-FoXS (23,24) to generate 3000 alternate structures of each model. These were constrained to their respective cross-link sets and simulated at 300K to determine whether an alternate model at a physiological temperature better represents the SAXS profile. **Fig. 9a** and **9b** show the final “open” and “closed” models superimposed onto the SAXS molecular envelope. Both show good visual fit to the molecular envelopes and excellent agreement with the experimental SAXS profile shown in **Fig. 9c** ($\chi=0.73$ and 0.83 for the “open” and “closed”, respectively). Additionally, MultiFoXS was used on a pool of all 6000 independent models (3000 each of the open and closed) to determine if a two-state model better fit the SAXS profile. The crystal structure was included in the pool as a control. Indeed, MultiFoXS picked a single model from each conformation for an improved fit of $\chi=0.67$ weighted at 64% open and 36% closed (**Fig. 9c**, orange). Lastly, contact plots were generated and superimposed for both conformations in **Fig. 9d** showing that 100% of the observed cross-links are accounted for between the two models.

Derivation of a Monomeric ApoA-I^{Δ185-243} Solution Model—Mei et al. postulated that monomeric apoA-I^{Δ185-243} occurs when the long

helix (residues 118-184), which participates in the dimer domain swap, doubles back onto the four helix bundle in the exact same position that the same sequence from the other monomer would sit in the dimer (12). We generated a starting model of this concept based on the crystal structure (**Fig. 10a**). Interestingly, **Fig. 10b** shows that the cross-linking pattern was 100% compatible, even on the un-optimized model. Using Modeller 9.14, 100 iterations of the model were generated with the imposed cross-link constraints. The best-fit model was then run through AllosMod-FoXS to generate 3000 alternate structures at 300K. The final best-fit model is shown superimposed on the molecular SAXS envelope for monomeric apoA-I^{Δ185-243} (**Fig. 10c**) with an excellent fit of $\chi=0.78$ (**Fig. 10d**).

DISCUSSION

Deletion mutants have become a valuable work-around to the notorious problem of getting full-length exchangeable apolipoproteins to crystallize for high-resolution structural studies. Typically, apolipoproteins have low thermodynamic stability with highly dynamic domains that are poised to interact with lipid (34,35). Removing these domains may make crystallization more tractable, but there are drawbacks to this approach. First, it is never clear how the structure of the deletion mutant relates to that of the full-length protein. Even though the missing domain is flexible, it may participate in interactions that are key to the global structure of the WT protein. Second, apolipoproteins are likely more sensitive than most proteins to high concentrations and non-physiological precipitants required for crystallization. Therefore, we believe that it is critical to evaluate predictions from crystal structures on proteins that are in solution.

We previously reported the crystal structure of a deletion mutant of apoA-IV which showed a remarkably similar helical “swap” as that reported by Mei et al for apoA-I^{Δ185-243}. SAXS and cross-linking analyses of dimeric apoA-IV⁶⁴⁻³³³ in solution agreed quite well with the crystal structure (13) and we went on to propose a structure for full-length apoA-IV using the crystal structure as a template. In the case of apoA-I^{Δ185-243}, however, it was immediately clear that both the SAXS analysis and the cross-linking patterns were not fully compatible with the crystal structure, at least in the dimer. Despite these differences, important predictions of the crystal structure were clearly

borne out in solution. For example, the long swapped helix between the dimers was confirmed. Intermolecular cross-links observed between Lys-96 and Ser-167, Lys-118 and Lys-140, Lys-118 and Ser-142, and Lys-118 to Lys-133 are consistent with the extended helix swap. There was also confirmatory evidence of the folded helical hairpins at each end of the dimer with the intramolecular cross-link between Lys-23 and Lys-59, for example.

However, we reproducibly noted interactions that could not be reconciled in the rigid crystal structure. Implementation of the hairpins in helix 5 and their subsequent juxtaposition to residues 106-116 satisfied critical cross-links constraints while also improving the fit to the SAXS data. Several studies have postulated the existence of a hairpin in helix 5 despite its absence in either crystal structure (9,12). In the double belt disc model (10), Li and colleagues hypothesized that helix 5 hairpins could reduce the diameter of discoidal HDL. Applying to the Mei crystal structure (12), helix 5 flexibility might allow the formation of a “presentation tunnel” for the docking of LCAT and the subsequent influx of cholesterol ester; an idea supported by molecular dynamics studies (10,36). We also observed interactions between the N-terminus and the central domain of the dimer. For example, the intermolecular cross-link between the N-terminus and Lys-118 can only occur if the N-terminal major helix (residues 7-34) swings back across the dimer as illustrated in our closed model. However, we also saw the N-terminus interacting intramolecularly with Lys-77 which is consistent with the crystal structure. Since the N-terminus cannot be in two places at once, we were forced to postulate at least two models in equilibrium.

This result is intriguing when considering the transition of apoA-I to a lipid-bound species. Mei et al postulated that the crystal structure could form a discoidal HDL particle through a “sequential unhinging of the N-terminal bundle” (12). They proposed that the hairpin at each end of the dimer can swing away from the AB-repeating backbone and then unfurl to form a ring that approximates the double belt model of lipid-bound apoA-I (10). This movement is quite analogous to our “closed” (i.e. closed ring) conformation of the lipid-free dimer. Both transitions are predicted to occur using the N-terminal minor helix as a hinge.

Mei et al proposed that the N-terminal bundles are stabilized by two hydrophobic clusters at each

end of the bundle (**Fig. 11a**). The N-terminal aromatic cluster holds the first helix, the second helix B of H1, and the helix of the first A unit of H2 together. The C aromatic cluster holds the N-terminal helix and H4(AB2) together through π - π interactions. Influx of lipid was suggested to open the N-terminal helix bundle by disrupting one or both staple domains. Our data suggests that, in solution, these helical bundles may be more dynamic (**Fig. 11b** and **c**). Interestingly, we found similar aromatic clusters in the closed model that might contribute to its stability (**Fig. 11c**). Trp-8 is in close proximity to Trp-108 and Phe-104 in the long swapped helix. Additionally, Phe-33 from both N-terminal major helices are in close proximity in the middle of the molecule. Thus hydrophobic clusters could be pseudo-stabilizing features in both conformers in solution.

We caution that this proposed structural equilibrium is only relevant in the context of this particular deletion mutant and its role with respect to lipid binding or other functions in the full length protein is less clear. The missing C-terminus has been shown repeatedly to play a major role in lipid binding (34,37-40). It is possible that, when present, the C-terminus stabilizes the N-terminal bundles in much the same way they appear to be under crystallization conditions (30). Indeed, in apoA-IV, there are extensive stabilizing interactions between the N- and C-termini at both ends of its dimer (41). Engagement of the apoA-I C-terminus, by lipid or perhaps ABCA1, may free up the N-terminal helix to swing away from the helical bundle as part of the particle assembly process. Nevertheless, the absence of the C-terminus may have allowed the fortuitous visualization of a transition step (the opening of the N-terminal bundles) during lipid binding that would otherwise not be apparent in a static full-length structure.

Based on our SAXS data, the half circle curvature likely manifests as lipid accumulates. The curvature in the crystal structure may have arisen from a pseudo-lipid-like environment contributed by PEG or other additives during crystallization. Indeed, additives like isopropanol can induce a lipid-bound-like structure to otherwise lipid-free apoA-I (42). Crystal packing and other factors could also be responsible (43).

With regard to monomeric apoA-I ^{Δ 185-243}, our results are highly compatible with the monomer scheme proposed by Mei et al. The idea is quite

similar to the “pocket knife closing” model that we proposed for apoA-IV (13). The overall α -helicity of our model is 61% which matches nicely with circular dichroism data estimating 59% helicity for this mutant in solution (12). Although it is dangerous to make direct comparisons from a deletion mutant to the monomeric full-length version of apoA-I, we did find it interesting that our final model exhibited α -helical character in many of the same regions assigned by the hydrogen-deuterium exchange experiments of Chetty et al (44) in WT apoA-I; there was 65% overlap of helical residues (**Fig. 10e**). There was also some cross-link overlap between apoA-I $^{\Delta 185-243}$ vs those reported for full-length apoA-I. These data imply that some of the structural features of apoA-I $^{\Delta 185-243}$ could apply to WT apoA-I. Confirmation awaits more detailed studies on WT apoA-I.

Finally, we acknowledge the strong experimental evidence showing that apoA-I exhibits molten globule characteristics in solution (29,30,35,45). ApoA-I has a free energy of

denaturation that is well below that of most soluble proteins (45,46) with helical segments that are constantly folding and unfolding on a timescale of seconds (44). Although the molecules must have a distinct shape as evidenced by the SAXS data, the two dimeric structures reported here are probably best thought of as two general conformational classes, each representing a set dynamically related structures which co-exist at any given time.

In summary, we report two related models for soluble dimeric apoA-I $^{\Delta 185-243}$ that differ by the location of the N-terminus. We also provided strong evidence supporting the postulated monomeric structure of apoA-I $^{\Delta 185-243}$. This work emphasizes that high-resolution structural studies should be coupled with innovative in-solution experiments to better understand the dynamics of the exchangeable apolipoproteins. Such understanding will allow us to better understand how they transition in response to lipid. Current work is focused on deriving a structure for full-length apoA-I using these models as a foundation.

ACKNOWLEDGEMENTS: The SAXS experiments were conducted at the Advanced Light Source (ALS), a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the Department of Energy, Office of Basic Energy Sciences, through the Integrated Diffraction Analysis Technologies (IDAT) program, supported by DOE Office of Biological and Environmental Research. Additional support comes from the National Institute of Health project MINOS (R01GM105404). This work was supported by grants HL67903 (WSD) and GM098458 (WSD and TBT).

CONFLICT OF INTEREST: The authors declare that they have no conflicts of interest related to the contents of this article.

AUTHOR CONTRIBUTIONS: JTM conducted experiments, derived the models, and wrote the paper. JM and MC conducted experiments. RGW, TBT, MKJ, JPS and WSD analyzed data, contributed to model building and assisted in manuscript writing. DBL, PCC, FCG analyzed data.

REFERENCES

1. Soutar, A. K., Garner, C. W., Baker, H. N., Sparrow, J. T., Jackson, R. L., Gotto, A. M., and Smith, L. C. (1975) Effect of the human plasma apolipoproteins and phosphatidylcholine acyl donor on the activity of lecithin: cholesterol acyltransferase. *Biochemistry* **14**, 3057-3064
2. Bodzioch, M., Orso, E., Klucken, J., Langmann, T., Bottcher, A., Diederich, W., Drobnik, W., Barlage, S., Buchler, C., Porsch-Ozcurumez, M., Kaminski, W. E., Hahmann, H. W., Oette, K., Rothe, G., Aslanidis, C., Lackner, K. J., and Schmitz, G. (1999) The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. *Nat Genet* **22**, 347-351
3. Brooks-Wilson, A., Marcil, M., Clee, S. M., Zhang, L. H., Roomp, K., van Dam, M., Yu, L., Brewer, C., Collins, J. A., Molhuizen, H. O., Loubser, O., Ouelette, B. F., Fichter, K., Ashbourne-Excoffon, K. J., Sensen, C. W., Scherer, S., Mott, S., Denis, M., Martindale, D., Frohlich, J., Morgan, K., Koop, B., Pimstone, S., Kastelein, J. J., and Hayden, M. R. (1999) Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency. *Nat. Genet.* **22**, 336-345
4. Rust, S., Rosier, M., Funke, H., Real, J., Amoura, Z., Piette, J. C., Deleuze, J. F., Brewer, H. B., Duverger, N., Deneffe, P., and Assmann, G. (1999) Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1. *Nat. Genet.* **22**, 352-355
5. Oram, J. F., Lawn, R. M., Garvin, M. R., and Wade, D. P. (2000) ABCA1 is the cAMP-inducible apolipoprotein receptor that mediates cholesterol secretion from macrophages. *Journal of Biological Chemistry* **275**, 34508-34511
6. Shah, A. S., Tan, L., Long, J. L., and Davidson, W. S. (2013) Proteomic diversity of high density lipoproteins: our emerging understanding of its importance in lipid transport and beyond. *J Lipid Res* **54**, 2575-2585
7. Segrest, J. P. (1977) Amphipathic helices and plasma lipoproteins: thermodynamic and geometric considerations. *Chem.Phys.Lipids* **18**, 7-22
8. Segrest, J. P., and Feldmann, R. J. (1977) Amphipathic helices and plasma lipoproteins: a computer study. *Biopolymers* **16**, 2053-2065
9. Borhani, D. W., Rogers, D. P., Engler, J. A., and Brouillette, C. G. (1997) Crystal structure of truncated human apolipoprotein A-I suggests a lipid-bound conformation. *Proc.Natl.Acad.Sci.U.S.A* **94**, 12291-12296
10. Li, L., Chen, J., Mishra, V. K., Kurtz, J. A., Cao, D., Klon, A. E., Harvey, S. C., Anantharamaiah, G. M., and Segrest, J. P. (2004) Double belt structure of discoidal high density lipoproteins: molecular basis for size heterogeneity. *J Mol Biol* **343**, 1293-1311
11. Borrell, B. (2009) Fraud Rocks Protein Community. *Nature* **462**, 970
12. Mei, X., and Atkinson, D. (2011) Crystal structure of C-terminal truncated apolipoprotein A-I reveals the assembly of high density lipoprotein (HDL) by dimerization. *J Biol Chem* **286**, 38570-38582
13. Deng, X., Morris, J., Dressmen, J., Tubb, M. R., Tso, P., Jerome, W. G., Davidson, W. S., and Thompson, T. B. (2012) The Structure of Dimeric Apolipoprotein A-IV and Its Mechanism of Self-Association. *Structure.* **20**, 767-779
14. Deng, X., Walker, R. G., Morris, J., Davidson, W. S., and Thompson, T. B. (2015) Role of Conserved Proline Residues in Human Apolipoprotein A-IV Structure and Function. *J Biol Chem* **290**, 10689-10702
15. Tubb, M. R., Smith, L. E., and Davidson, W. S. (2009) Purification of recombinant apolipoproteins A-I and A-IV and efficient affinity tag cleavage by tobacco etch virus protease. *Journal of Lipid Research* **50**, 1497-1504
16. Walker, R. G., Deng, X., Melchior, J. T., Morris, J., Tso, P., Jones, M. K., Segrest, J. P., Thompson, T. B., and Davidson, W. S. (2014) The structure of human apolipoprotein A-

- IV as revealed by stable isotope-assisted cross-linking, molecular dynamics and small angle X-ray scattering. *Journal of Biological Chemistry*
17. Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951) Protein measurement with the Folin phenol reagent. *J Biol Chem* **193**, 265-275
 18. Lima, D. B., de Lima, T. B., Balbuena, T. S., Neves-Ferreira, A. G., Barbosa, V. C., Gozzo, F. C., and Carvalho, P. C. (2015) SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis. *J Proteomics*
 19. Dyer, K. N., Hammel, M., Rambo, R. P., Tsutakawa, S. E., Rodic, I., Classen, S., Tainer, J. A., and Hura, G. L. (2014) High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. *Methods Mol Biol* **1091**, 245-258
 20. Franke D, S. D. (2009) DAMMIF, a program for rapid ab initio shape determination in small-angle scattering. *J Appl Crystallogr* **42**, 342-346
 21. Sali, A., and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815
 22. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A., and Sali, A. (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys J* **105**, 962-974
 23. Schneidman-Duhovny, D., Hammel, M., and Sali, A. (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res* **38**, W540-544
 24. Weinkam, P., Pons, J., and Sali, A. (2012) Structure-based model of allostery predicts coupling between distant sites. *Proc.Natl.Acad.Sci.U.S.A* **109**, 4875-4880
 25. Laccotripe, M., Makrides, S. C., Jonas, A., and Zannis, V. I. (1997) The carboxyl-terminal hydrophobic residues of apolipoprotein A-I affect its rate of phospholipid binding and its association with high density lipoprotein. *Journal.of.Biological.Chemistry.* **272**, 17511-17522
 26. Swaim, C. L., Smith, J. B., and Smith, D. L. (2004) Unexpected products from the reaction of the synthetic cross-linker 3,3'-dithiobis(sulfosuccinimidyl propionate), DTSSP with peptides. *J Am.Soc.Mass Spectrom.* **15**, 736-749
 27. Leavell, M. D., Novak, P., Behrens, C. R., Schoeniger, J. S., and Kruppa, G. H. (2004) Strategy for selective chemical cross-linking of tyrosine and lysine residues. *J Am Soc Mass Spectrom* **15**, 1604-1611
 28. Silva, R. A., Hilliard, G. M., Fang, J., Macha, S., and Davidson, W. S. (2005) A three-dimensional molecular model of lipid-free apolipoprotein A-I determined by cross-linking/mass spectrometry and sequence threading. *Biochemistry* **44**, 2759-2769
 29. Pollard, R. D., Fulp, B., Samuel, M. P., Sorci-Thomas, M. G., and Thomas, M. J. (2013) The conformation of lipid-free human apolipoprotein A-I in solution. *Biochemistry* **52**, 9470-9481
 30. Segrest, J. P., Jones, M. K., Shao, B., and Heinecke, J. W. (2014) An experimentally robust model of monomeric apolipoprotein A-I created from a chimera of two X-ray structures and molecular dynamics simulations. *Biochemistry* **53**, 7625-7640
 31. Rambo, R. P., and Tainer, J. A. (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* **95**, 559-571
 32. Hammel, M. (2012) Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). *Eur.Biophys.J.* **41**, 789-799
 33. Jones, M. K., Catta, A., Li, L., and Segrest, J. P. (2009) Dynamics of activation of lecithin:cholesterol acyltransferase by apolipoprotein A-I. *Biochemistry* **48**, 11196-11210
 34. Davidson, W. S., Hazlett, T., Mantulin, W. W., and Jonas, A. (1996) The role of apolipoprotein AI domains in lipid binding. *Proc Natl Acad Sci U S A* **93**, 13605-13610
 35. Saito, H., Dhanasekaran, P., Nguyen, D., Holvoet, P., Lund-Katz, S., and Phillips, M. C. (2003) Domain structure and lipid interaction in human apolipoproteins A-I and E, a general model. *Journal of Biological Chemistry* **278**, 23227-23232

36. Segrest, J. P., Jones, M. K., Catta, A., and Thirumuruganandham, S. P. (2012) Validation of previous computer models and MD simulations of discoidal HDL by a recent crystal structure of apoA-I. *J Lipid Res* **53**, 1851-1863
37. Palgunachari, M. N., Mishra, V. K., LundKatz, S., Phillips, M. C., Adeyeye, S. O., Alluri, S., Anantharamaiah, G. M., and Segrest, J. P. (1996) Only the two end helices of eight tandem amphipathic helical domains of human apo A-I have significant lipid affinity - Implications for HDL assembly. *Arterioscl Throm Vas* **16**, 328-338
38. Holvoet, P., Zhao, Z., Vanloo, B., Vos, R., Deridder, E., Dhoest, A., Taveirne, J., Brouwers, E., Demarsin, E., Engelborghs, Y., and et al. (1995) Phospholipid binding and lecithin-cholesterol acyltransferase activation properties of apolipoprotein A-I mutants. *Biochemistry* **34**, 13334-13342
39. Minnich, A., Collet, X., Roghani, A., Cladaras, C., Hamilton, R. L., Fielding, C. J., and Zannis, V. I. (1992) Site-directed mutagenesis and structure-function analysis of the human apolipoprotein A-I. Relation between lecithin-cholesterol acyltransferase activation and lipid binding. *Journal of Biological Chemistry* **267**, 16553-16560
40. Panagotopoulos, S. E., Witting, S. R., Horace, E. M., Hui, D. Y., Maiorano, J. N., and Davidson, W. S. (2002) The role of apolipoprotein A-I helix 10 in apolipoprotein-mediated cholesterol efflux via the ATP-binding cassette transporter ABCA1. *J Biol Chem* **277**, 39477-39484
41. Deng, X., Morris, J., Chaton, C., Schroder, G. F., Davidson, W. S., and Thompson, T. B. (2013) Small-angle X-ray scattering of apolipoprotein A-IV reveals the importance of its termini for structural stability. *J Biol Chem* **288**, 4854-4866
42. Leroy, A., and Jonas, A. (1994) Native-like structure and self-association behavior of apolipoprotein A-I in a water/n-propanol solution. *Biochim.Biophys.Acta* **1212**, 285-294
43. Segrest, J. P., Jones, M. K., Catta, A., Manchekar, M., Datta, G., Zhang, L., Zhang, R., Li, L., Patterson, J. C., Palgunachari, M. N., Oram, J. F., and Ren, G. (2015) Surface Density-Induced Pleating of a Lipid Monolayer Drives Nascent High-Density Lipoprotein Assembly. *Structure* **23**, 1214-1226
44. Chetty, P. S., Mayne, L., Lund-Katz, S., Stranz, D., Englander, S. W., and Phillips, M. C. (2009) Helical structure and stability in human apolipoprotein A-I by hydrogen exchange and mass spectrometry. *Proc.Natl.Acad.Sci.U.S.A* **106**, 19005-19010
45. Gursky, O., and Atkinson, D. (1996) Thermal unfolding of human high-density apolipoprotein A-1: implications for a lipid-free molten globular state. *Proc Natl Acad Sci U S A* **93**, 2991-2995
46. Reijngoud, D. J., and Phillips, M. C. (1982) Mechanism of Dissociation of Human Apolipoprotein a-I from Complexes with Dimyristoylphosphatidylcholine as Studied by Guanidine-Hydrochloride Denaturation. *Biochemistry* **21**, 2969-2976

FOOTNOTES:

This work was supported, in whole or in part, by National Institutes of Health Grants R01 GM098458 (to W.S.D. and T.B.T.) and HL67093 (to W.S.D.).

The abbreviations used are: apo, apolipoprotein; HDL, High density lipoprotein; LCAT, lecithin:cholesterol acyl transferase; SAXS, small angle X-ray scattering; TEV, tobacco etch virus; BS³, bis-(sulfosuccinimidyl) suberate.

FIGURE LEGENDS

Figure 1: Principle behind isotope-assisted cross-linking: Recombinant proteins produced with either ^{14}N amino acids (light gray) or ^{15}N amino acids (dark gray) are mixed at a 1:1 ratio under denaturing conditions and allowed to reassemble resulting in the four combinations on the left. Proteins are locked into position by cross-linking and digested with trypsin. Intermolecular cross-links result in four mass peaks as shown in the top panel while intramolecular cross-links result in two mass peaks shown in the bottom panel.

Figure 2: Expression and purification of lipid-free apoA-I $^{\Delta 185-243}$. ApoA-I $^{\Delta 185-243}$ was expressed and purified from bacteria as described in *Methods*. **(a):** SDS-PAGE analysis of wild-type apoA-I (lane 1) and apoA-I $^{\Delta 185-243}$ (Lane 2). **(b):** Resolution and molecular weight determination of ^{14}N and ^{15}N apoA-I $^{\Delta 185-243}$ using mass spectrometry.

Figure 3: Redistribution of lipid-free apoA-I $^{\Delta 185-243}$ after denaturation and reassembly. Distribution of dimeric and monomeric apoA-I $^{\Delta 185-243}$ before (dotted line) and after denaturation and reassembly (solid line) using size exclusion chromatography.

Figure 4: Separation and purification of monomeric and dimeric lipid-free ApoA-I $^{\Delta 185-243}$ by gel filtration and resulting MS spectra of intramolecular and intermolecular cross-links identified from each species. ^{14}N and ^{15}N ApoA-I $^{\Delta 185-243}$ were denatured, mixed 1:1, and reassembled as described in *Methods*. Proteins were cross-linked and subjected to gel filtration chromatography. **(a):** Mixture of cross-linked monomeric and dimeric apoA-I $^{\Delta 185-243}$. **(b):** Purified monomeric (lane 1) and dimeric (lane 2) samples after separation by gel filtration chromatography. Gels were stained with Coomassie Blue. **(c):** Mass spectra for the $[\text{M}+4\text{H}]^{4+}$ ion of the cross-linked peptides spanning residues 107-116 and 117-123. Monomeric (blue) and dimeric (green) molecules exhibit a clear intramolecular span in both. **(d):** Mass spectra for the $[\text{M}+4\text{H}]^{4+}$ ion of the cross-linked peptides spanning residues 117-123 and 132-140 exhibiting a clear intermolecular span in the dimer and intramolecular span in the monomer.

Figure 5: Small Angle X-ray Scattering (SAXS) analysis of lipid-free apoA-I $^{\Delta 185-243}$ dimer and monomer. **(a):** Intensity distribution of the SAXS scattering function of both dimer isolations (purple and green) and the monomer (tan). **(b):** Pairwise distribution function where AU equals arbitrary units. **(c):** $q^3 \cdot I(q)$ versus $q^3(\text{\AA}^{-1})^3$. **(d):** $q^4 \cdot I(q)$ versus $q^4(\text{\AA}^{-1})^4$. **(e):** The SAXS *ab initio* reconstructions of the dimeric and monomeric **(f)** apoA-I $^{\Delta 185-243}$. The envelopes were generated using DAMMIF and DAMAVER. The averaged envelope was rendered and displayed using UCSF Chimera. Data was from 3 samples each at a different concentration. Each concentration was run at 4 different exposure times.

Figure 6: Comparison of dimeric apoA-I $^{\Delta 185-243}$ crystal structure with experimental cross-links. **(a):** The cross-linking data from **Tables 1 and 2** are superimposed on a molecular contact plot generated for the reported crystal structure of dimeric apoA-I $^{\Delta 185-243}$ (PDB entry 3R2P). The X and Y axes delineate the residue number of apoA-I from 1-243. Dark areas represent the truncated C-terminal segments of the protein. The diagonal line bisects the figure with the bottom/right showing intramolecular interactions and the upper left showing intermolecular interactions. **(b):** Structural representation showing the cross-links (red dots circled in the contact plot) that deviate from predictions from the crystal structure.

Figure 7: Comparison of dimeric apoA-I $^{\Delta 185-243}$ with experimental SAXS data. **(a):** The theoretical scattering profile of the crystal dimer generated by FoXS superimposed onto the experimental scattering profile from the SAXS analysis with the Chi distribution values in parenthesis. **(b):** The residuals normalized to the experimental scattering profile (black dotted line). **(c):** The pairwise distribution function of the experimental dimer (black) and the theoretical distribution generated after solvation using ScÅtter where AU equals arbitrary units. **(d):** the crystal structure superimposed onto the molecular envelopes generated from the SAXS analysis.

Figure 8: Derivation of the starting model(s) of dimeric apoA-I $^{\Delta 185-243}$. The open model of apoA-I $^{\Delta 185-243}$ was generated by applying a hairpin in helix 5 of each apoA-I molecule. The “open” model consists of

cross-links consistent with the placement of the N-terminus (red) in the crystal structure (top inset). The “closed” initial model was generated by folding the N-terminal segments (red) to the middle of the molecule guided by remaining intermolecular cross-links observed in that region (right inset).

Figure 9: AllosMod-FoXS modeling of the open and closed conformations of dimeric apoA-I^{Δ185-243}. The best-fit “open” and “closed” conformations are superimposed onto the *ab initio* dimeric apoA-I^{Δ185-243} molecular envelope in panels (a) and (b), respectively. (c): Comparison of the experimental dimeric apoA-I^{Δ185-243} x-ray scattering profile (black line) to the best-fit “open” (green line), “closed” (purple line) and two-state (orange line) theoretical profiles with the Chi distribution in parenthesis. (d): The residual scatter plot normalized to the experimental scatter profile (black dotted line). (e): Cross-linking data from Tables 1 and 2 superimposed on a molecular contact blot generated from the best fit of the combined “open” (dark) and “closed” (light) models.

Figure 10: AllosMod-FoXS modeling of the monomeric apoA-I^{Δ185-243}. (a): Derivation of the initial model as postulated by Mei. et. al. (b): Cross-linking data from Tables 1 and 2 superimposed on a molecular contact blot generated for the theoretical model. (c): Single best fit conformation from AllosMod-FoXS output superimposed on the molecular envelope. (d): Comparison of the experimental monomeric apoA-I^{Δ185-243} x-ray scattering profile (black line) to the single best-fit initial (green line) and final (purple line) profiles generated by AllosMod-FoXS. (e): Regions of α -helicity in our final model of monomeric apoA-I^{Δ185-243} (magenta) vs that determined in monomeric full-length apoA-I by Chetty et. al. (black) and the crystal structure (green). The boxes represent α -helical segments.

Figure 11: Disruption of hydrophobic “staples” that allow mobility of the N-terminus in solution. (a): One monomer of the Mei crystal structure of apoA-I showing the C aromatic cluster (comprised of residues Phe33, Phe104, Trp108) and an N aromatic cluster (Trp8, Phe71, Trp72) that is proposed to stabilize the four helical bundles on each end of the dimer structure. (b): The open form of the solution model reported here showing the plausibility of both interactions. (c): The closed form of the solution model showing disruptions of the N aromatic cluster allowing the N-terminal major helix, using the N-terminal minor helix (blue) as a hinge, to swing toward the middle of the dimer. Stabilization of the N-terminal helix is postulated to occur with two alternative aromatic clusters, *i*) between Phe33 on the N-terminal major helices of each molecule and *ii*) Trp8 at the end of the N-terminal major helix with Trp108 and Phe104 on the domain swap.

Table 1

Identified INTRA-peptide BS³ cross-links in isolated lipid-free apoA-I^{A185-243} monomeric and dimeric samples derived from mixed ¹⁴N and ¹⁵N labeled proteins

Cross-link	Peptides involved ^a	Mod. ^b	Sample	<i>Da</i>	
				¹⁴ N mass ^c	¹⁵ N mass
K88-K94	84-QEMSKDLEEV K AK-96	XL	M	1671.86	1688.81
			D	1671.86	1688.81
K12-K23	11-VKDLATVYVDVL K D S GR-27	XL	M	2015.13	2037.07
			D	2015.12	2037.06
K23-S25	11-VKDLATVYVDVL K D S GR-27	XL,H	M	2171.22	2193.16
			D	2171.21	2193.15
K133-K140	132-QKLHELQE K LSPLGEEMR-149	XL	M	2302.24	2329.16
			D	2302.23	2329.15
K133-S142	132-QKLHELQE K LSPLGEEMR-149	XL	M	2302.24	2329.16
			D	-- ^d	--
K140-S142	132-QKLHELQE K LSPLGEEMR-149	XL,H	M	2458.31	2485.24
			D	--	--
K106-K107	97-VQPYLDDFQ K KWQEEMEL Y R -116	XL	M	2782.39	2811.32
			D	--	--

^a Lysines or Serines involved in cross-links are in bold.

^b Chemical modifications: XL = 1 complete cross-link (+138.068 Da), H = 1 hydrolyzed cross-linker (+156.079 Da).

^c Experimentally derived monoisotopic mass for each peptide with each isotope and the combinations.

^d Not detected. These ions were detectable in one sample (i.e. monomer or dimer) but not in the other.

Table 2

Identified INTER-peptide BS³ cross-links in isolated lipid-free apoA-I^{A185-243} monomeric and dimeric samples derived from mixed ¹⁴N and ¹⁵N labeled proteins

Cross-link	Peptides involved ^a	Mod. ^b	Sample ^c	Peptide mass ^c				Span
				¹⁴ N	¹⁴ N/ ¹⁵ N	¹⁵ N/ ¹⁴ N	¹⁵ N	
Da								
K118-S142	117-QKVEPLR-123	XL	M	2037.12	--	--	2061.06	Intra
	141-LSPLGEEMR-149		D	-- ^d	--	--	--	--
K118-K133	117-QKVEPLR-123	XL	M	2158.25	--	--	2185.17	Intra
	132-QKLHELQEK-140		D	2158.24	2170.21	2173.20	2185.17	INTER
NT-K182	-1-GDEPPQSPWDR-10	XL	M	2236.10	--	--	2261.03	Intra
	178-LEALKEN-184		D	-- ^d	--	--	--	--
NT-K118	-1-GDEPPQSPWDR-10	XL	M	-- ^d	--	--	--	--
	117-QKVEPLR-123		D	2289.17	2301.13	2305.12	2317.07	INTER
K12-K182	11-VKDLATVYVDVVK-23	XL	M	2415.39	--	--	2439.32	Intra
	178-LEALKEN-184		D	-- ^d	--	--	--	--
K107-K118	107-KWQEEMELYR-116	XL	M	2417.28	--	--	2445.20	Intra
	117-QKVEPLR-123		D	2417.27	--	--	2445.19	Intra
K12-K118	11-VKDLATVYVDVVK-23	XL	M	-- ^d	--	--	--	--
	117-QKVEPLR-123		D	2468.46	2480.42	2483.42	2495.38	INTER
NT-K107	-1-GDEPPQSPWDR-10	XL	M	-- ^d	--	--	--	--
	107-KWQEEMELYR-116		D	2831.32	2847.28	2847.28	2863.25	INTER
NT-K12	-1-GDEPPQSPWDR-10	XL	M	2882.52	--	--	2913.43	Intra
	11-VKDLATVYVDVVK-23		D	2882.50	--	--	2913.42	Intra
K96-S167	95-AKVQPYLDDFQK-106	XL	M	2889.50	--	--	2921.41	Intra
	161-THLAPYSDEL-171		D	2889.49	2905.45	2905.45	2921.41	INTER
K118-K140	117-QKVEPLR-123	XL	M	-- ^d	--	--	--	--
	134-LHELQEKLSPLGEEMR-149		D	2914.60	2926.55	2937.53	2949.50	INTER
K118-K142	117-QKVEPLR-123	XL	M	-- ^d	--	--	--	--
	134-LHELQEKLSPLGEEMR-149		D	2914.60	2926.55	2937.53	2949.50	INTER
K88-K96	84-QEMSKDLEEVK-94	XL	M	2923.50	--	--	2953.41	Intra
	95-AKVQPYLDDFQK-106		D	2923.49	--	--	2953.41	Intra
K40-K118	28-DYVSQFEGSALGKQLNLK-45	XL	M	-- ^d	--	--	--	--
	117-QKVEPLR-123		D	3002.65	3014.62	3025.55	3037.54	INTER
K118-K140	117-QKVEPLR-123	XL	M	3326.85	--	--	3365.74	Intra
	132-QKLHELQEKLSPLGEEMR-151		D	-- ^d	--	--	--	--
NT-K23	-1-GDEPPQSPWDR-10	XL,H	M	3453.79	--	--	3491.68	Intra
	11-VKDLATVYVDVVKDSDGR-27		D	3453.77	--	--	3491.67	Intra
K77-K182	62-EQLGPVTQEFWDNLEKETEGLR-83	XL	M	3570.84	--	--	3609.73	Intra
	178-LEALKEN-184		D	-- ^d	--	--	--	--
NT-140	-1-GDEPPQSPWDR-10	XL,H	M	-- ^d	--	--	--	--
	132-QKLHELQEKLSPLGEEMR-149		D	3740.87	3756.79	3767.81	3783.76	INTER
NT-K77	-1-GDEPPQSPWDR-10	XL	M	4037.97	--	--	4083.85	Intra
	62-EQLGPVTQEFWDNLEKETEGLR-83		D	4037.95	--	--	4083.82	Intra
K23-K59	11-VKDLATVYVDVVKDSDGR-27	XL,H	M	4052.22	--	--	4096.09	Intra
	46-LLDNWDSVTSTFSKLR-51		D	4052.19	4074.11	4074.11	4096.07	Intra
K40-K140	28-DYVSQFEGSALGKQLNLK-45	XL,H	M	4454.38	--	--	4504.24	Intra
	132-QKLHELQEKLSPLGEEMR-149		D	4454.35	--	--	4504.22	Intra
K88-K107	95-AKVQPYLDDFQKQWQEEMELYR-116	XL,H,H	M	4628.35	--	--	4674.21	Intra
	84-QEMSKDLEEVK-94		D	-- ^d	--	--	--	--

^a Lysines or Serines involved in cross-links are in bold.

^b Chemical modifications: XL = 1 complete cross-link (+138.068 Da), H = 1 hydrolyzed cross-linker (+156.079 Da).

^c Experimentally derived monoisotopic mass for each peptide with each isotope and the combinations.

^d Not detected. These ions were detectable in one sample (i.e. monomer or dimer) but not in the other.

Table 3

Experimental parameters from SAXS sampling of apoA-I^{A185-243}

<i>apoA-I</i> ^{A185-243} Species	I(O) (Guinier) cm ⁻¹	R _g (Guinier) Å	Real Space R _g Å	D _{max} Å	Volume Å ³	DAMMIF NSD Å
Monomer						
4.0 mg/ml	1460	22.9	23.34	80	45642	
2.0 mg/ml	859	23.4	23.75	78	49575	0.678 ± 0.059
1.0 mg/ml	396	23.3	23.55	76	64078	
Dimer (Denatured/Refolded)						
4.0 mg/ml	2270	34.5	37.76	123	89421	
2.0 mg/ml	1530	37.4	38.08	130	94124	1.204 ± 0.529
1.0 mg/ml	890	40.2	38.41	125	103611	
Dimer (Not Denatured)						
2.6 mg/ml	1400	36.1	37.4	119	83759	
1.7 mg/ml	925	36.7	37.7	122	83373	0.906 ± 0.324
1.0 mg/ml	522	35.7	37.7	120	103468	

Figure 1

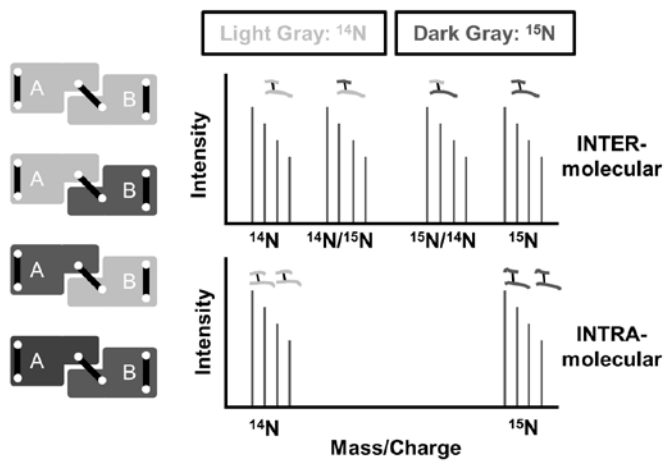


Figure 2

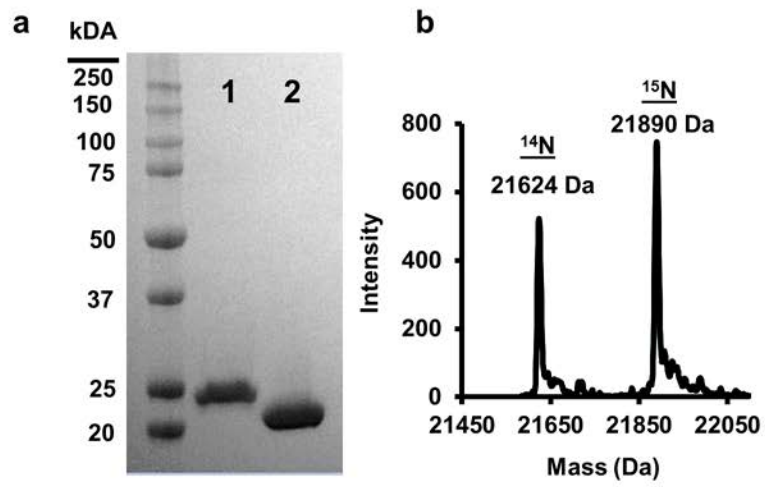


Figure 3

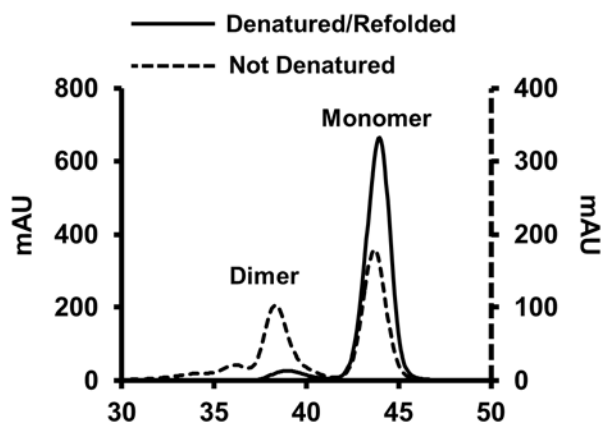


Figure 4

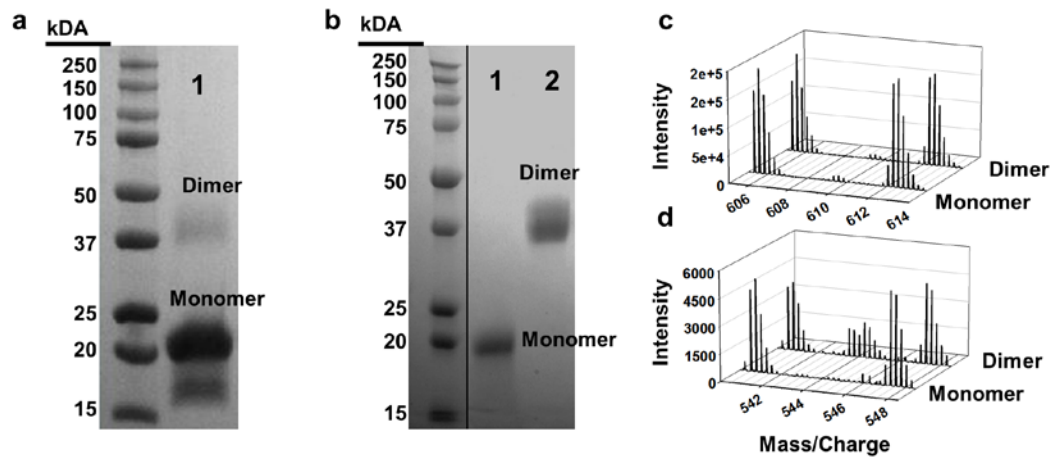


Figure 5

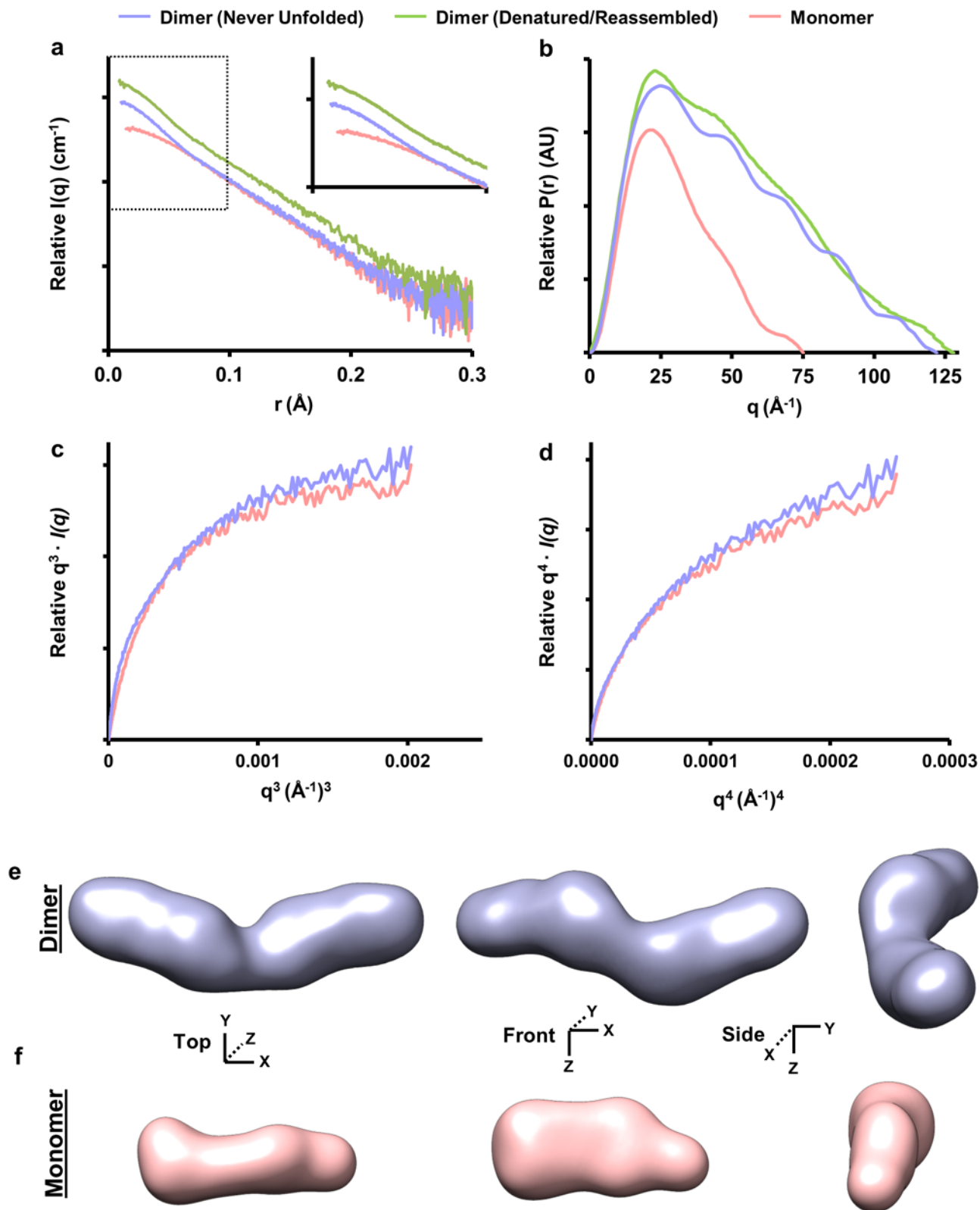


Figure 6

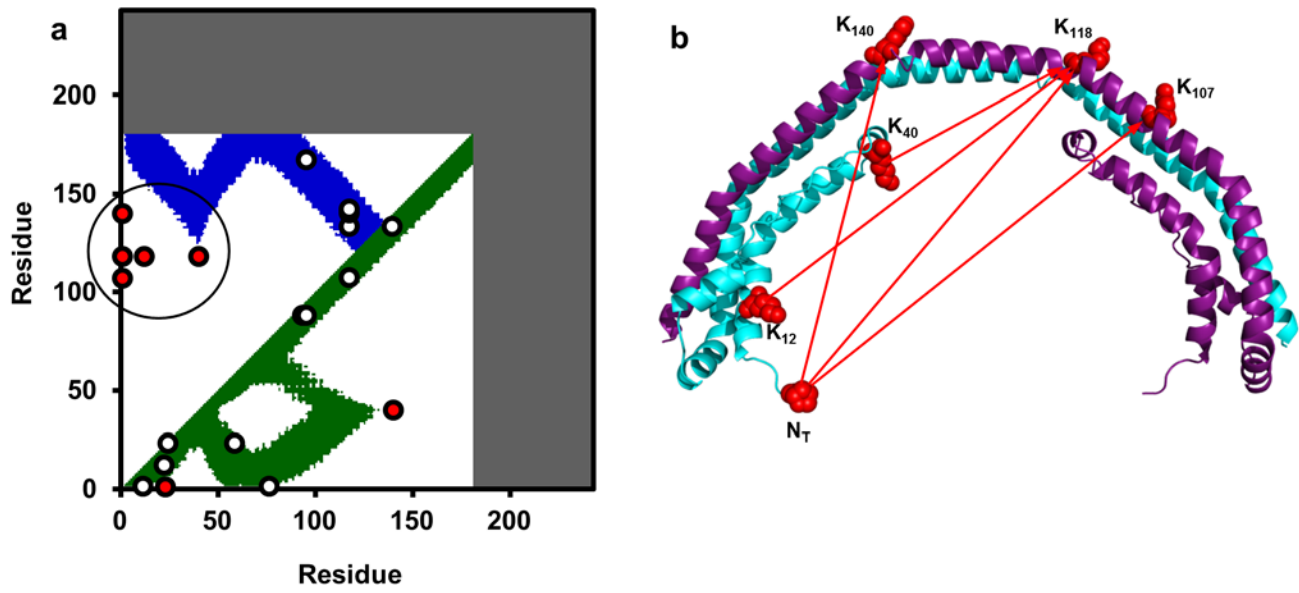


Figure 7

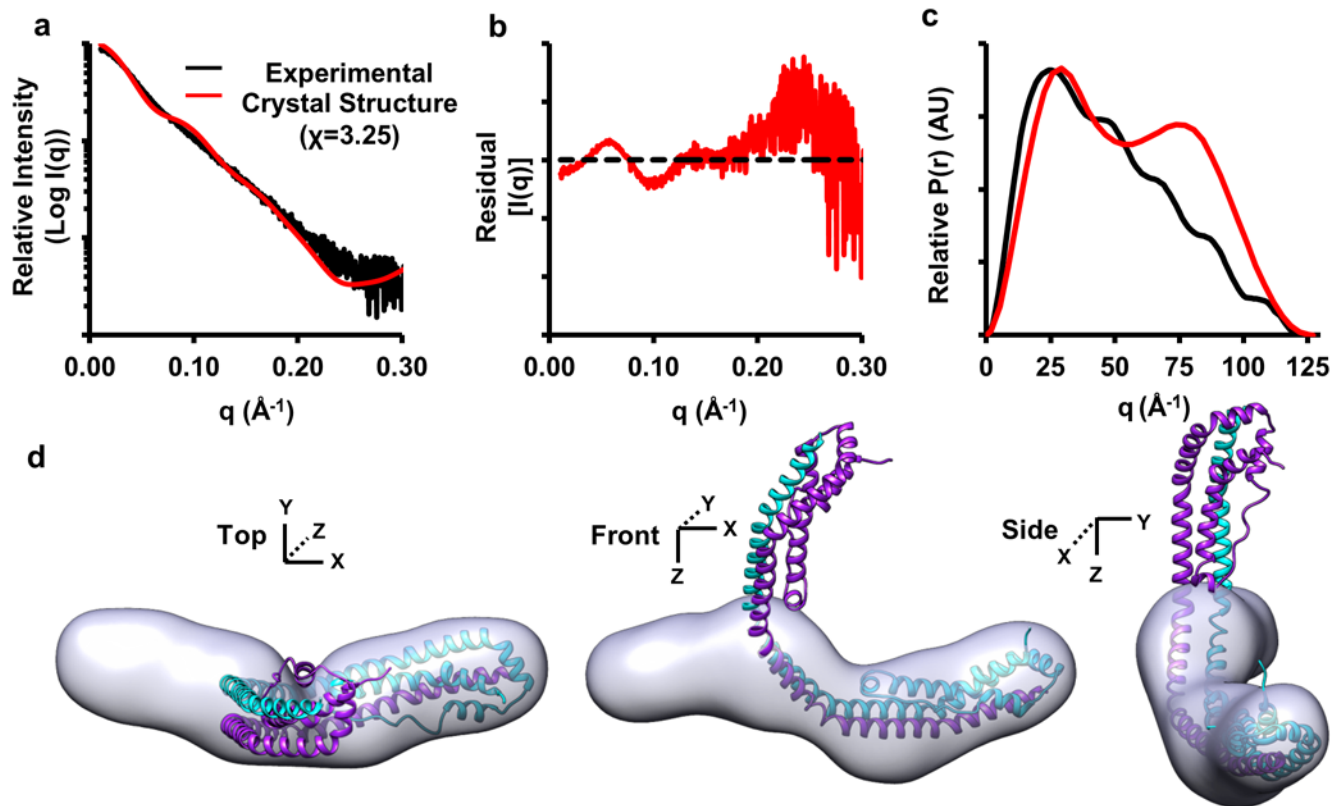


Figure 8

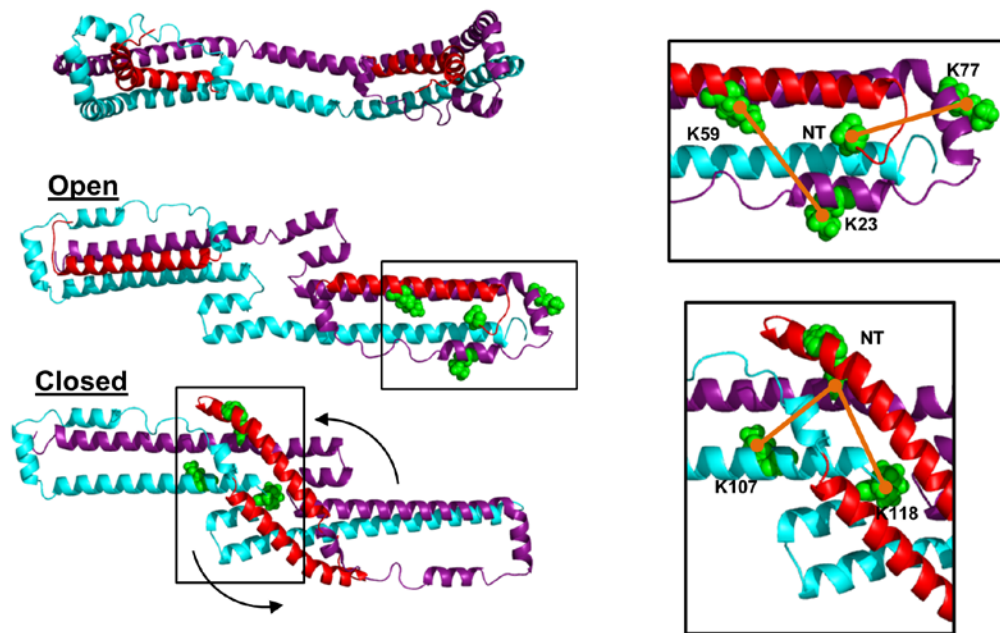


Figure 9

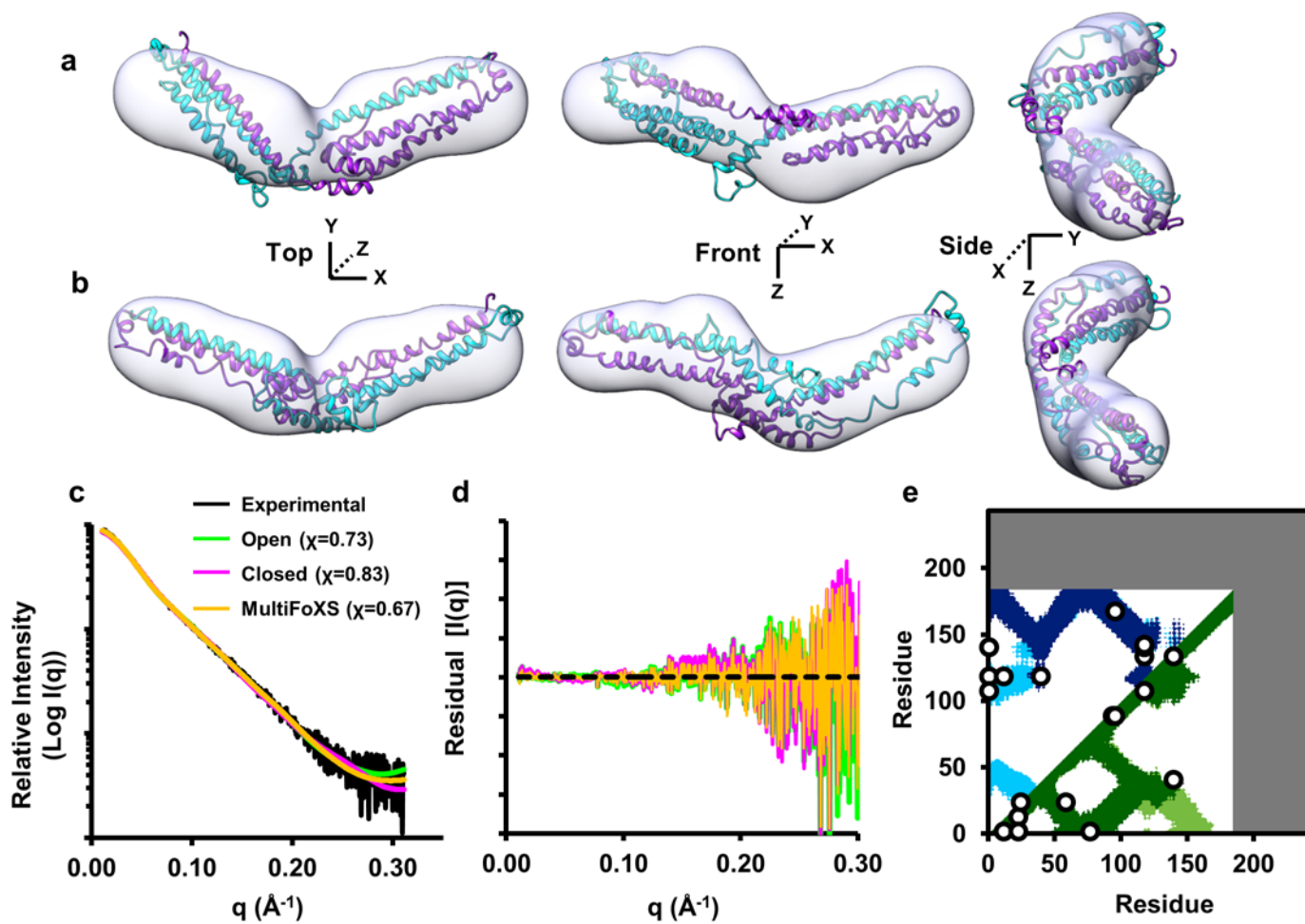


Figure 10

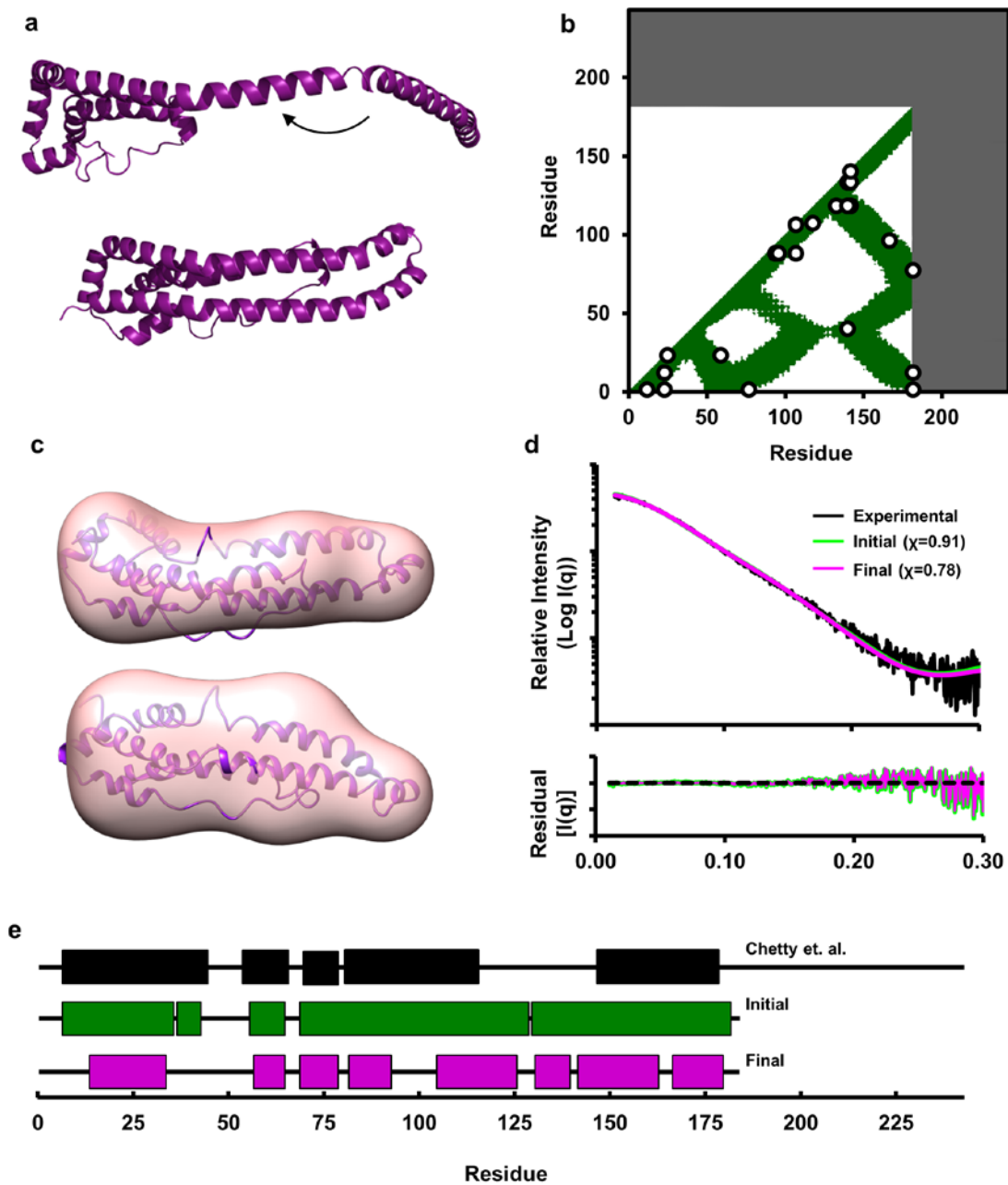
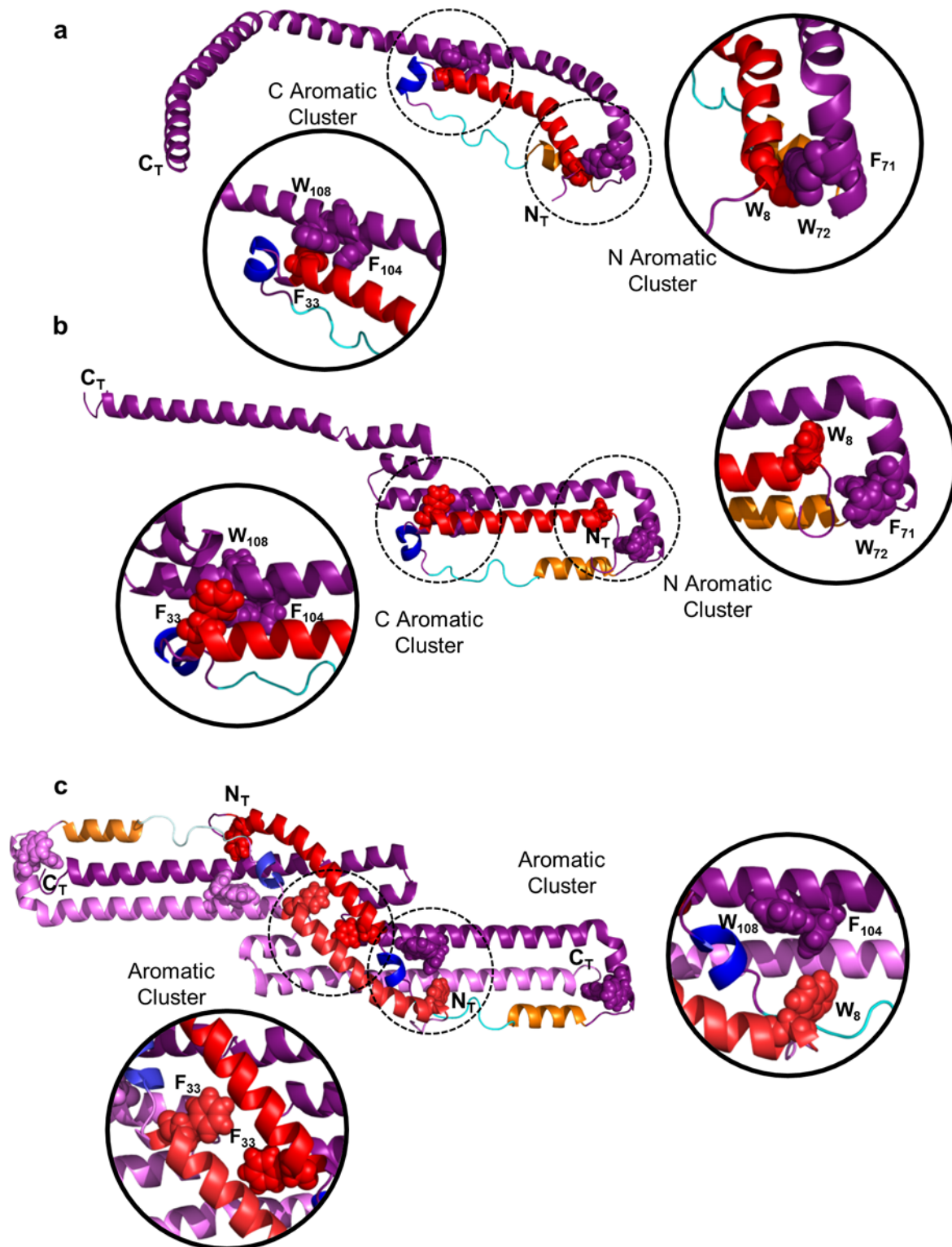


Figure 11



Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0

Paulo C Carvalho^{1,2,7}, Diogo B Lima^{1,7}, Felipe V Leprevost^{1,3}, Marlon D M Santos¹, Juliana S G Fischer¹, Priscila F Aquino⁴, James J Moresco⁵, John R Yates III⁵ & Valmir C Barbosa⁶

¹Computational Mass Spectrometry Group, Carlos Chagas Institute, Fiocruz Paraná, Curitiba, Brazil. ²Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil. ³Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA. ⁴Leonidas e Maria Deane Institute, Fiocruz Amazonas, Manaus, Brazil. ⁵Laboratory for Biological Mass Spectrometry, The Scripps Research Institute, La Jolla, California, USA. ⁶Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. ⁷These authors contributed equally to this work. Correspondence should be addressed to P.C.C. (paulo@pccarvalho.com).

Published online 10 December 2015; doi:10.1038/nprot.2015.133

PatternLab for proteomics is an integrated computational environment that unifies several previously published modules for the analysis of shotgun proteomic data. The contained modules allow for formatting of sequence databases, peptide spectrum matching, statistical filtering and data organization, extracting quantitative information from label-free and chemically labeled data, and analyzing statistics for differential proteomics. PatternLab also has modules to perform similarity-driven studies with *de novo* sequencing data, to evaluate time-course experiments and to highlight the biological significance of data with regard to the Gene Ontology database. The PatternLab for proteomics 4.0 package brings together all of these modules in a self-contained software environment, which allows for complete proteomic data analysis and the display of results in a variety of graphical formats. All updates to PatternLab, including new features, have been previously tested on millions of mass spectra. PatternLab is easy to install, and it is freely available from <http://patternlabforproteomics.org>.

INTRODUCTION

Shotgun proteomics has revolutionized biochemical and biomedical research by enabling the identification and quantification of thousands of proteins in complex biological samples such as organelles, cell lysates, biological fluids and tissues¹. The field's denomination of shotgun proteomics describes a strategy developed in the Yates laboratory to characterize proteins that are analyzed indirectly through peptides obtained by proteolysis, in analogy to shotgun genomic sequencing². The core of the discipline relies on state-of-the-art nanochromatography coupled with mass spectrometry, which is one of the most sensitive methods of analytical chemistry, in order to dissociate peptide ions in the mass spectrometer and to ultimately obtain peptide sequences; from these sequences, one can infer and quantify the proteins found in complex mixtures. Hundreds of thousands of tandem mass spectra are commonly generated in an experiment, and therefore advanced bioinformatics algorithms are required to make sense of all the data. In a typical experiment, peptides are fractionated by liquid chromatography on-line with tandem mass spectrometry, and protein identification is achieved by comparing experimental spectra against those theoretically generated from a sequence database. Proteins are then inferred by matching the identified peptide sequences to the sequences in the database; as peptides can match more than one protein, proteins can be further grouped according to a maximum parsimony criterion³. Peptide spectrum-matching (PSM) algorithms commonly leverage data from existing genomic projects. As a postgenomic discipline, the goals of shotgun proteomics are far more ambitious than those of genome sequencing, as shotgun proteomics aims to report protein expression, interaction, localization, post-translational modifications, turnover time and so on, when comparing different biological states. During the past decade, shotgun proteomics has been applied in many different ways to advance biological discovery. Notable examples can be found in studies describing differential protein expression between subcellular compartments⁴, pinpointing changes in proteomic

profiles of cancer biopsies⁵ and describing the contents of venoms to ultimately aid in biotechnological applications^{6,7}.

The field was jump-started by the creation of SEQUEST, an algorithm that correlates tandem mass spectra with theoretical spectra generated from a sequence database⁸. In what followed, the coupling of strong cation-exchange chromatography with reversed-phase chromatography on-line with tandem mass spectrometry set new heights in terms of the number of peptide identifications. This technology, later renamed as multidimensional protein identification technology (MudPIT)⁹, as well as competing strategies that use ultra-long chromatography gradients¹⁰, was adopted and thus raised the bar in terms of challenges, both in the handling of the new computational burden and in how to statistically deal with what was considered 'big data' at the time. In response, a new class of algorithms appeared, geared toward postprocessing the search engine results in order to statistically pinpoint identifications with confidence; examples of pioneering efforts are Peptide Prophet¹¹ and DTASelect¹². At the same time, breakthroughs on how to quantify complex peptide mixtures analyzed by mass spectrometry were being attained; the two main pillars of these breakthroughs were labeled and label-free approaches. Examples of the former are the isobaric tags¹³ and stable isotope amino acid labeling¹⁴, and examples of the latter are spectral counting^{15,16} and extracted-ion chromatograms (XICs)¹⁷. Naturally, intensive software development tailored toward enabling these quantification approaches became necessary. As the possibilities for how to mine the 'proteosphere'¹⁸ continued to expand, a plethora of new software programs began to be sparsely distributed among members of the community, each addressing very specific niches. These have included, for example, algorithms for scoring phosphosites¹⁹, deconvoluting mass spectra²⁰ and even for dealing with unsequenced organisms¹⁸.

PatternLab and other widely adopted proteomic pipelines

With so many options to choose from for analyzing shotgun proteomic data, efforts were shifted toward the creation of unified

pipelines: indeed, deciding which software to pick and making them interact with one another were challenging problems. Thus, the first pipelines emerged, including the trans-proteomic pipeline (TPP)^{21,22}, OpenMS²³, MaxQuant^{24,25} and PatternLab for proteomics²⁶, each having its own set of advantages and limitations. In what followed, SkyLine²⁷ and Galaxy²⁸ emerged to overcome some of the limitations of the aforementioned tools at the time. Although there is great overlap among these software pipelines, each has a special set of features that provides advantages when analyzing data originating from a certain setup.

TPP and Galaxy are tailored to (but not limited to) working on computing clusters, thus relieving users from the burden of processing large amounts of data (and therefore vastly mobilizing resources, such as storage) locally on their own computers. As these tools are generally remotely accessed through a web-based interface or command-line tools, no requirements are imposed on the local operating system or hardware configuration. Through the years, several leading groups have worked together on developing modules for TPP, and thus ultimately questions regarding details of how each module works can be addressed directly by the corresponding specialists. In contrast, the team behind Galaxy focuses on making available a customizable workflow management system, and thus efforts have been channeled toward providing a sophisticated environment for users to integrate several data analysis tools and protocols (as opposed to developing the data analysis modules themselves). In fact, this strategy culminated in making Galaxy an environment capable of integrating genomic, transcriptomic, proteomic and metabolomic data²⁹.

In contrast, MaxQuant, OpenMS, Skyline and PatternLab are all designed exclusively to be used locally on one's computer, with some clear benefits over their web-based counterparts. For example, when an update is done on a web-based pipeline, there is the possibility of immediate (and sometimes undesired) impact on ongoing analyses. Desktop users, on the other hand, have control over when to update their software. Moreover, it must be noted that today's high-end desktops, and even notebooks, have become so powerful that they are fully capable of analyzing the data from large-scale proteomic experiments efficiently.

MaxQuant, Skyline and PatternLab all require Microsoft's Windows 7 (or later) operating system, as they are based on .NET, which is a software framework that runs primarily on Microsoft Windows. In contrast, OpenMS can be executed on any operating system, as it is based on the C++ programming language. Another advantage of OpenMS is that its modules are all available as stand-alone tools, which facilitates integration into third-party workflows or the design of custom, local bioinformatics pipelines. As for the other tools, MaxQuant has been known to excel in stable isotope labeling by amino acids in cell culture (SILAC) experiments and Skyline in its unmatched capabilities in experiments addressing selected reaction monitoring (SRM) and parallel reaction monitoring (PRM). More recently, Skyline became capable of analyzing data-independent acquisition (DIA) data, as described in a previous protocol³⁰. PatternLab, in turn, provides one of the most complete and user-friendly experiences, owing to its very refined and interactive graphical user interface. As for its hallmarks, we believe that they lie in analyzing label-free data through the T-Fold³¹ module and in the isobaric (e.g., isobaric tags for relative and absolute quantification (iTRAQ) or tandem mass tags (TMT)) analyzer module.

Some of its unique features include providing an integrated cloud service³², modules for statistically filtering and performing assembly of *de novo* sequencing data³³, statistically scoring phosphopeptides³⁴, dealing with time-course experiments³⁵ and offering a module for integrated Gene Ontology analysis³⁶. Modules yet to be integrated in future versions are capable of deisotoping and decharging mass spectra¹⁸, and of identifying cross-linked peptides to address protein-protein interaction and to aid in providing structural data³⁷ (the latter is described in a recent protocol³⁸). Therefore, even though all mentioned tools, web- and desktop-based alike, overlap substantially with one another, each has its own hallmarks and unique features and may, as such, be more suitable for one's working style and needs.

PatternLab is freely available software, and it is flexible enough to be used in the analysis of most shotgun proteomic experiments. We advise using PatternLab on any experiment requiring label-free quantification, or on experiments in which the data have been chemically labeled with isobaric markers.

Development of the protocol

Since its launch in 2008, PatternLab has undergone continual improvement and expansion. The very first version was limited to working with spectral counting, and it offered strategies for pinpointing differentially expressed proteins, but all modules from that time have since been replaced by more sophisticated versions. Such major updates led us to release the system's first major protocol in 2010 (ref. 39). Thanks to the continual influx of suggestions from their various users, the modules continued to evolve and new modules appeared, such as the Search Engine Processor⁴⁰ (SEPro) for filtering and organizing shotgun proteomics data, and a module for XICs. A revised version of that first protocol was then published in 2012 (ref. 41). The PatternLab version at the time, PatternLab for proteomics 2.0, consisted of a series of modular software. A major request from its community of users was for the installation process of so many modules (one at a time) to be simplified. In addition, there was a desire for greater integration among the (then-independent) modules so that they would not have to be dealt with separately. Moreover, installing the modules could sometimes require installing third-party software such as the Java Runtime Environment, as well as having to deal with configuration files. Simply put, PatternLab needed to be reengineered to be completely installable at a single click of the mouse, as well as to work as a unit. PatternLab for proteomics 3.0 achieved this in 2013, by uniting all modules under a single graphical user interface and thus fulfilling all user requests of that time.

Since 2013, PatternLab has acquired new modules and functionalities. Some examples are as follows: Búzios, which allows the clustering of similar proteomic profiles⁵; the XD Scoring system, for evaluating the confidence in phosphosites³⁴; PepExplorer, a tool for analyzing shotgun proteomic data of unsequenced organisms³³; tools for performing analysis of variance (ANOVA); the incorporation of the Comet search engine, wrapped in a graphical user interface⁴², for analyzing isobaric experiments (e.g., iTRAQ and TMT); and a cloud service that enables large-scale quantitative predictions and comparisons of protein domains³². Some existing modules were significantly upgraded, such as the one for XICs. PatternLab for proteomics 4.0 is the culmination of these various changes; some of these changes are major, to the point of

spanning the complete workflow, but they always aim to simplify the process of analyzing shotgun proteomic data in an increasingly integrated environment. This protocol introduces the freely available PatternLab for proteomics 4.0, and it shows how to operate the latest modules and how to deal with the new, simplified workflow. For those modules that underwent no changes, readers are referred to the corresponding sections of the previously published protocols.

Experimental design

PatternLab is adaptable to many experimental designs, and as such it is applicable to analyzing data from most proteomic experiments. The topic of sample preparation and data acquisition in the mass spectrometer is an extensive one, and it encompasses tasks that must be performed before analyzing the data; in this regard, we recommend following the steps in the protocol by Richards *et al.*⁴³.

Sequence database preparation. Databases of protein sequences are required so that theoretical mass spectra generated from them can be compared with experimental spectra. For the widely adopted PSM approach, we recommend downloading sequences from UniProt⁴⁴, as some downstream analysis tools (e.g., the Gene Ontology explorer) can take advantage of this knowledgebase. Regardless, any type of sequence database in the FASTA format is supported, so users can download sequences from the US National Center for Biotechnology Information (NCBI) or even use an in-house-generated database. The UniProt knowledgebase comprises the Swiss-Prot and the TrEMBL databases; the former contains manually annotated and reviewed sequences, whereas the latter's sequences are automatically annotated but not reviewed. We recommend downloading, whenever possible, only the species-specific database, which contains entries from both Swiss-Prot and TrEMBL. This is achieved by navigating to the UniProt website at <http://www.uniprot.org>, clicking on the large 'Proteomes' square, and then naming the species in the search box. The sequences can be obtained by clicking on the number in the 'Protein count' column beside the desired species, clicking on the download button, and then selecting the FASTA format. If wishes to use the Gene Ontology as a downstream tool, an additional download of the sequences, in the 'Text' format, must be done.

Subsequently, a target-decoy database must be generated before searching with PatternLab's integrated version of Comet. PatternLab contains a module that allows the automatic generation of decoys by reversing each sequence of the target database. A PatternLab option that we strongly recommend is to automatically include the 127 common contaminants found in proteomic experiments (keratin, BSA and so on). Even though there are many possible ways to generate decoy sequences, sequence reversal has been the most widely adopted one, as it conserves the complexity of the database (e.g., approximately the same number of decoy peptides and target peptides after an *in silico* digestion⁴⁵).

Peptide identification from tandem mass spectra. PatternLab adopts Comet for the comparison of experimental and theoretically generated mass spectra. Comet is a fast and sensitive open-source search engine that stemmed from the widely adopted SEQUEST⁸. Comet is constantly being updated, and PatternLab's automatic updates may include an updated built-in Comet search

engine. A complete description of Comet's parameters is available at the Comet project's website http://comet-ms.sourceforge.net/parameters/parameters_201502/; PatternLab allows the setting of these parameters through its graphical user interface.

When searching for peptide candidates within a database, a precursor mass tolerance must be specified. When using high-resolution instruments such as an Orbitrap Velos (Thermo, San Jose), we recommend using no less than 40, even if the mass spectrometer used provides, say, 5 p.p.m. The suggestion for the adoption of wide search windows is empirical and comes from experimenting with the search engine. Nevertheless, our experience is aligned with that of John S. Cottrell and David M. Creasy, from whom we quote, "The common observation is that FDR (false discovery rate) increases rather than decreases for very narrow precursor tolerances because the reliability of the scoring is reduced by the small numbers of candidates"⁴⁶. Finally, we note that Comet's results will later be statistically filtered and postprocessed by SEPro. At that final stage, any matching containing more than a tighter tolerance (e.g., 5 p.p.m.), will be discarded.

Peptides absent from the database cannot be identified by classical PSM. The PSM strategy is therefore blind to mutations and polymorphisms, and it may not work satisfactorily on organisms that lack a reference peptide sequence database. Moreover, post-translational modifications must be specified a priori. Often these are unknown for the experiment at hand, so usually only carbamidomethylation of cysteine and oxidation of methionine are specified as fixed and variable modifications, respectively. By having a quick look at UniMod (<http://www.unimod.org>), the protein modification for mass spectrometry database, one can take note of the variety of modifications that can occur in a sample. To cope with these limitations, approaches stemming from *de novo* sequencing have emerged. Among them we highlight Spectral Networks⁴⁷, Mod-A⁴⁸, MS-Blast⁴⁹ and PepExplorer³³. The first two are capable of pinpointing unanticipated modifications, whereas the last two start with *de novo* sequencing results and align them against sequence databases of homolog organisms so that similar proteins can be determined. In particular, PepExplorer is integrated into PatternLab's workflow, but notwithstanding this we recommend that the user consider other applications when working with unsequenced organisms. Being based on different paradigms, such applications may provide complementary results.

Statistically filtering peptide spectrum matches. The sensitivity of a PSM search engine is intimately related to how the search results are postprocessed. PatternLab relies on SEPro⁴⁰ to statistically filter its results in order to achieve a predetermined FDR. The filtered results can be saved as a 'sepr' file and shared with collaborators. In this regard, anyone can open these files and have access to a dynamic report that enables sorting proteins according to various criteria (coverage, normalized spectral abundance factors, spectral counts, and so on), as well as access to annotated mass spectra and search engine scores, and also accomplish much more within a few clicks of the mouse. Even though PatternLab houses Comet, SEPro (and consequently PatternLab) is compatible with ProLuCID⁵⁰, SEQUEST⁸ and the Spectrum Identification Machine for PITC⁵¹. Our 2012 protocol provides the main steps for using SEPro⁴¹. At the time of this writing, PatternLab still required several separate downloads for installation and relied mostly on ProLuCID, but SEPro has now been ported to the

main interface. Only the features that were implemented since 2012 are highlighted herein.

Quantitative proteomics. PatternLab can work with label-free quantification and with chemically labeled relative quantification. Among the label-free strategies, spectral counting has often been used in experiments with multidimensional separation (e.g., MudPIT). A spectral count refers simply to the number of tandem mass spectra associated with a protein, and it is used as a surrogate for the protein's relative abundance. The community has proposed various ways for normalizing data of this type, and PatternLab optionally allows normalization by the normalized spectral abundance factor (NSAF) approach, which takes into account a protein's length during the normalization process⁵². PatternLab also allows quantification by XICs, which are frequently used in single-shot experiments and are obtained by plotting the intensity of a given *m/z* value, plus or minus a given tolerance, over a given span of time. The area underneath this curve, or integral, can then be used as a surrogate for a peptide's relative abundance in the mixture and as such provide a basis for comparison against the XIC of the same peptide in different mixtures.

A popular strategy for chemically labeling peptides to increase confidence in relative quantification has been the use of isobaric tags; PatternLab also makes available modules for analyzing such data. Examples of widely adopted, commercially available tags are iTRAQ¹³ and TMT⁵³, which enable experiments to be multiplexed. Currently, the most commonly adopted configurations are the 4-plex iTRAQ, 6-plex TMT and 8-plex iTRAQ; we point out that higher degrees of multiplexing are also available. These reagents rely on stable isotope-labeled molecules that covalently bind to the side-chain amines and the N terminus of polypeptide chains. PatternLab used to rely on the now deprecated SEProQ module (then available as a separate download) for dealing with XICs and isobaric tag data, but this module has been substantially re-designed and integrated into PatternLab for proteomics 4.0. A limitation of relative quantification by isobaric tags has been the interference of the nearly isobaric peptides that are co-fragmented in the mass spectrometer along with the desired precursor ion, which generates a false relative quantification as the reporter ions' signals get mixed with those from the nearly isobaric molecules. To overcome this limitation, elaborate methods such as MultiNotch, which is only applicable to state-of-the-art or customized mass spectrometers, have been developed⁵⁴. As far as we know, PatternLab's isobaric module, described herein, is the only one to support MultiNotch acquisition while still providing a solution to standard data acquisition by automatically identifying and discarding multiplexed spectra.

The project must be organized in terms of what run belongs to which condition. This is performed using PatternLab's Project Organization module, which ultimately generates a file that contains all identifications and the quantification data of all

runs from the entire experiment for use in downstream analyses by several modules. Examples of such analyses are clustering proteins or peptides with similar expression profiles for time-course experiments, clustering data, pinpointing differentially expressed proteins or proteins found in only one condition, performing ANOVA and even Gene Ontology analyses. In this protocol, we provide the main steps, highlighted in the graphical summary in **Figure 1**, involved in these analyses. An accompanying video, which demonstrates PatternLab for proteomics 4.0 in action, is available that provides an overview of the software (**Supplementary Video 1**).

Limitations of PatternLab for proteomics 4.0

The following are the major limitations of the current PatternLab version:

- No handling of data from N15 labeling quantitative proteomic experiments.
- No handling of SILAC data.
- No handling of SRM or PRM data⁵⁵.
- Not yet fully integrated with a public repository such as PRIDE⁵⁶.

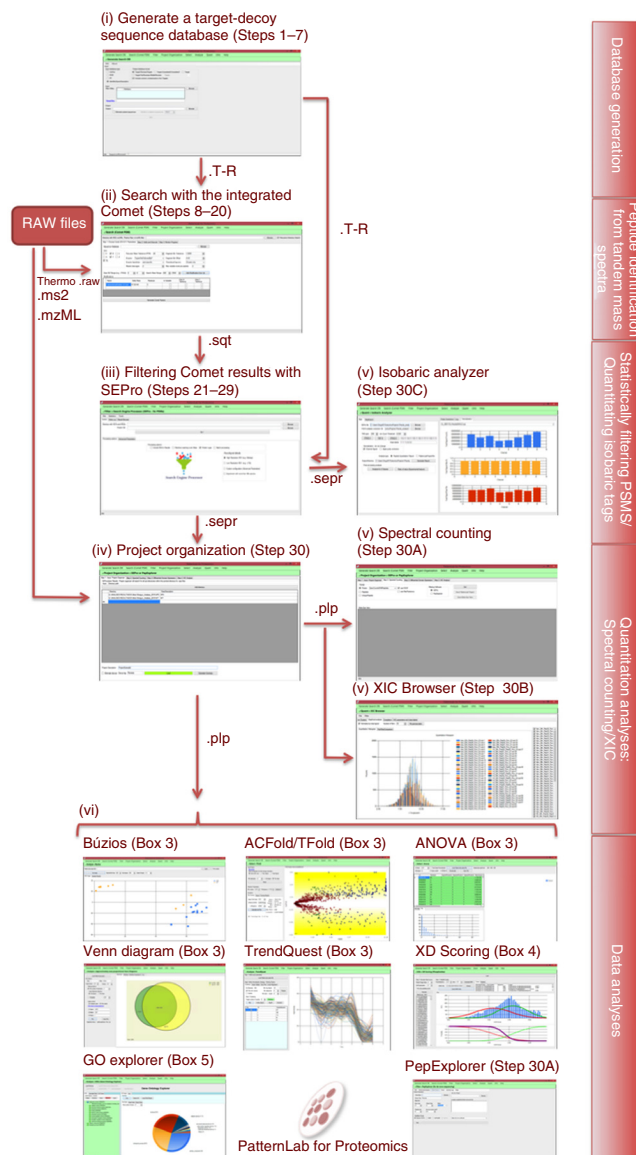


Figure 1 | Overview of PatternLab's workflow. In a general workflow, a target-decoy database is prepared (i), the mass spectra are searched (ii) and statistically filtered to meet a user-defined FDR (iii), the project is organized in terms of which mass spectral files belong to what biological conditions (iv), quantitative information is extracted (v) and then the various downstream modules for data analysis can be used (vi). The main modules for database generation, peptide identification, statistical filtering and quantification of PSMs, and data analysis are presented. The protocol steps pertinent to each module are also given.



PROTOCOL

- Cannot handle top-down data (that is, mass spectrometry of intact proteins).
- The seamless integration with raw data from mass spectrometers other than those from Thermo requires exporting data to text-based formats such as MS2, mzXML, mzML or MGF.
- Requires a computer with Microsoft Windows 7 or later.

We are working to overcome most of these limitations, although we are not currently looking into addressing the third limitation, as Skyline already does a good job on that. Tackling the seventh limitation requires updates in the .NET environment from Microsoft's end. The sixth limitation can be overcome by referring to the ProteoWizard project⁵⁷.

MATERIALS

EQUIPMENT

Hardware requirements

- A personal computer with at least 6 GB of RAM and an ×86–64 processor
▲ **CRITICAL** We strongly recommend having a multicore processor, as it can effectively deal with the parallel computation performed by some of the modules, and having at least 16 GB of RAM.
- Local storage is required for processing mass spectrometer RAW files. The space occupied by these files can vary substantially, depending on the mass spectrometer used

Data files

- Mass spectra data files in any of these formats: mzML⁵⁸, mzXML, MS2 (ref. 59) or Thermo's RAW

Software requirements

- Microsoft Windows 7 or later (64-bit version) ▲ **CRITICAL** 'Regional and Language Options' have to be set to English, as several modules are tied to its decimal system.
- .NET Framework 4.5 or later needs to be installed. The .NET Framework is made freely available by Microsoft; a new computer should already have this requirement fulfilled. Nonetheless, if the .NET Framework is

not detected during PatternLab's installation, an attempt will be made to automatically install it through Microsoft's website. The latest version, as of the time of this writing, is available from <http://www.microsoft.com/en-us/download/details.aspx?id=42642>

- Thermo Scientific MSFileReader should be installed in case the user wishes to work directly from the RAW instrument files. Instructions on obtaining this file are available from <https://thermo.flexnetoperations.com/control/thmo/download?element=6306677>

EQUIPMENT SETUP

PatternLab setup Go to the PatternLab home page at <http://patternlabforproteomics.org> and click on the 'Download' link. If the .NET Framework 4.5 or later is already installed on the computer, clicking on the 'launch' link will automatically install PatternLab; otherwise, click on the 'Install' button. After PatternLab is installed for the first time, its main screen will pop up (Fig. 2). ▲ **CRITICAL** Administrative access privileges are required for installation. ▲ **CRITICAL** If PatternLab fails to install, you may need to update to .NET 4.5 or later. You can manually download and install the latest version of the .NET framework from Microsoft's website.

PROCEDURE

Generating a target-decoy sequence database ● **TIMING** 5 s to several hours, depending on settings

▲ **CRITICAL** A target-decoy sequence database must be generated before PSM.

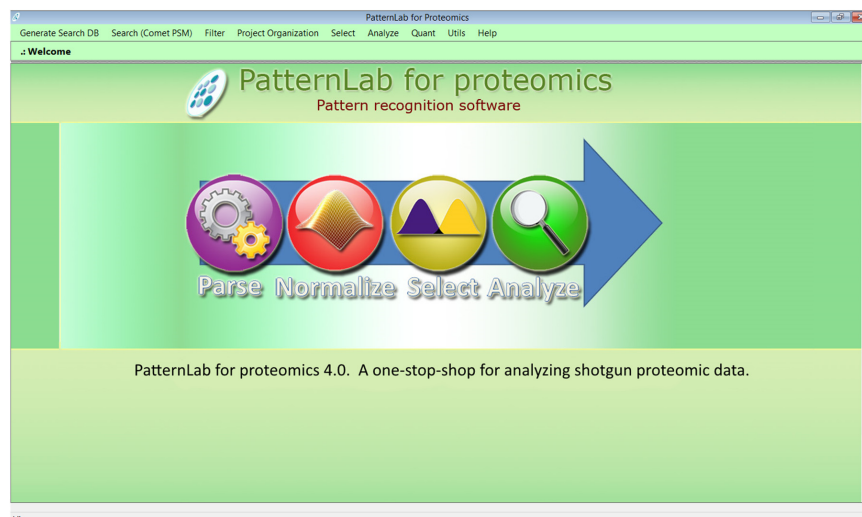
1| Click on 'Generate Search DB' in the upper-left corner of the interface. The sequence database module will load (Supplementary Fig. 1).

2| Select an input database file format (UniProt, NCBI, and so on). A generic format called 'Identifier Space Description' can be used for any FASTA file.

3| Choose the output database format. We strongly recommend using the target-decoy approach that automatically includes a reverse version of each sequence in the database (with a 'Reverse_' attached to the beginning of the identifier). The other formats are made available for very specific purposes of software benchmarking.

4| Check the 'Include common contaminants in the Targets' checkbox to include the sequences of 127 common contaminants to mass spectrometry (e.g., keratins) at the beginning of the output sequence database.

Figure 2 | PatternLab's main screen. The general PatternLab workflow is indicated by the order in which the pull-down menus appear. Generally, a target-decoy sequence database is prepared, searched with Comet and filtered to achieve a given FDR using SEPro or PepExplorer (in the case of *de novo* sequencing). The project is then organized in order to indicate which files belong to which biological condition. Downstream analysis is achieved by using the modules in the 'Select and Analyze' menus.



- 5] Click on the 'Browse' button in the Input group box and select sequence databases that were downloaded from the Internet. More than one database can be selected by pressing the Ctrl key while clicking on the file names in the file selection window.
- 6] Click on the 'Save as' button in the Output group box, and specify the name of the new database. A checkbox reading 'Eliminate subset sequences' is available for the elimination of sequences that meet a user-specified identity within other sequences in the database. When this happens, a note is appended to the remaining protein's sequence description with a reference to the eliminated sequence. Specifying an identity below 100% will significantly increase the time for generating the database.
- 7] Press the 'Go' button to generate the new database. For proteogenomic studies, consider taking the extra measures described by Nesvizhskii⁶⁰ so that the FDR is not underestimated. This is recommended.

Performing PSM with the integrated Comet search engine ● TIMING 1–2 min to >1 d, depending on sample complexity and equipment used

- 8] Click on the 'Search (Comet PSM)' option from the main menu. The Comet graphical user interface will appear (Fig. 3).
- 9] Indicate a directory containing Thermo RAW, MS2, mzXML or mzML mass spectra files in the topmost textbox. The 'Recursive Directory Search' box must be checked for multiple directories to be searched.
- 10] Specify a target-decoy sequence database.
- 11] Specify a precursor mass tolerance. We suggest using the default 40 p.p.m., even for high-resolution mass spectrometers, as discussed in the INTRODUCTION.
- 12] For species-specific databases, set the parameter 'Enzyme specificity' to 'semi-specific'. This is recommended, and it will increase the search space and reduce the search engine speed. However, having an estimate of how many semi-tryptic peptides were obtained after a tryptic digest can shed light on how well the sample was digested. If the sample was markedly degraded, we expect >20% of the peptides to be semi-specific. Contrasting with this, samples with no more than 5% semi-specific peptides should be taken as having undergone almost no degradation (optional; see Box 1). We note that some degradation is always expected.
- 13] Specify the number of missed cleavages allowed. We recommend allowing up to two misses for standard shotgun proteomic searches.
- 14] Specify the 'Fragment Bin Tolerance', 'Fragment Bin Offset' and 'Theoretical Fragment Ions' parameters. For low-resolution tandem MS, as generally provided in a Thermo LTQ, we recommend setting these values to 1.0005, 0.4 and 'M peak only', respectively. For high-resolution tandem MS, provided by a Thermo Q-Exactive, we recommend experimenting also with 0.02, 0 and 'default peak shape', respectively. The latter setting may slow the software substantially and, in our hands, it has usually led to little improvement in the search results.
- 15] Post-translational modifications (PTMs) should be specified by clicking on the 'Add Modification from Lib' button, which makes the modification library window pop up (Supplementary Fig. 2). To select one or more PTMs, click on the corresponding row header, which highlights the entire row, and then on the 'Add selected row to my Search.xml' button.

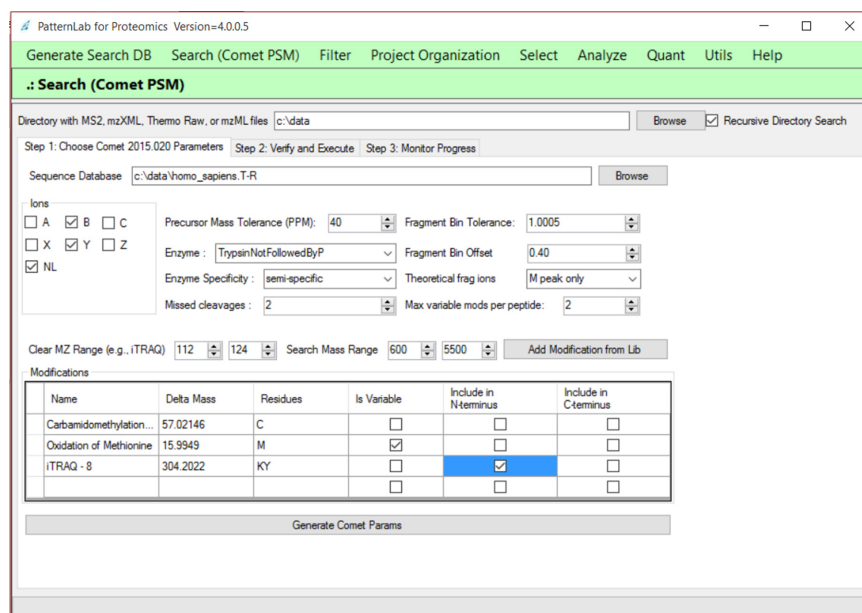


Figure 3 | PatternLab's Comet search engine graphical user interface.



Box 1 | On enzymatic specificity

The Comet search can be performed in the fully specific or semi-specific search spaces. Fully specific refers to considering only peptides originating from a complete digestion (i.e., with enzyme cleavage sites at both the C terminus and the N terminus). Semi-specific makes Comet lift the constraint that both cleavage sites be present, allowing instead the presence of only one. For example, in the sequence R.APBCK.A, where '.' denotes the occurrence of cleavage, selecting semi-specific will make Comet consider A, AP, APB, APBC, K, CK, BCK and PBCK, in addition to APBCK. Otherwise (i.e., if fully specific is selected), the search space will be limited to APBCK.

16| Optionally, new PTMs can be saved to the library. To do this, simply fill out the empty row (always the bottommost one) with the corresponding information and click on the 'Update my Lib' button.

17| Indicate whether the modification is variable, and which of the two termini it applies to, by checking the corresponding boxes. For example, if not all methionines in the sample are expected to be oxidized, then the modification should be checked as variable. However, for modifications that are expected in all occurrences of the amino acid, such as, say, carbamidomethylation of cysteine, leave the variable option unchecked. **Figure 3** exemplifies a situation in which the iTRAQ 8-plex is to be considered as a fixed modification on the N terminus and for the K and Y amino acids, whereas variable oxidation is expected for the M amino acid.

18| For experiments making use of isobaric tags (e.g., iTRAQ or TMT), enter the m/z range that spans the reporter ions as a 'Clear MZ Range' option. This will have the software ignore the signal of these reporter ions when matching the theoretical spectra with the experimental one.

19| Click on the 'Generate Comet Params' button. The user will be transferred to the next tab, 'Step 2: Verify and Execute'. The user should then simply click on the 'Save Comet Params' button, thus saving all search engine specifications in a text file in the search directory. We note that the contents of this file are made available in the upper section of the window, which provides the experienced user with the possibility of manually altering the search engine specifications.

20| Click on the 'Go!' button. The user will be automatically transferred to the 'Step 3: Monitor progress' tab, which in turn is automatically updated as the search makes progress. Comet's terminal screen will also pop up for each new search. The results files in the SQT format will be generated.

! CAUTION Closing the Comet pop-up terminal screen will terminate the search.

? TROUBLESHOOTING

Statistically filtering Comet results with SEPro ● TIMING ~30 s

21| Load SEPro by clicking on the 'Filter' menu and then on 'Search Engine Processor (SEPro – for PSM)'. SEPro's entry screen will appear as in **Supplementary Figure 3**.

22| Copy and paste the directory containing the SQT files into the topmost textbox. This can also be achieved by clicking on the corresponding 'Browse' button and navigating to the directory. If the corresponding directory contains a comet.params file, then SEPro will automatically detect the path to the sequence database and fill out the next textbox (Protein DB).

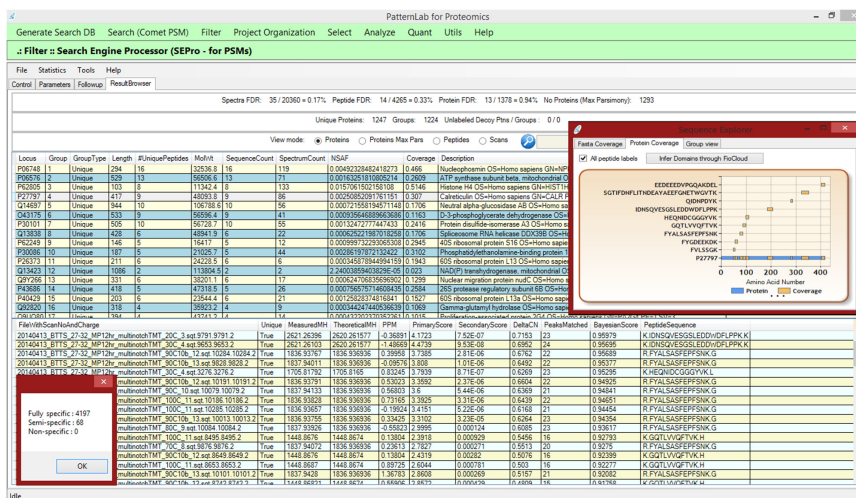
23| Choose from one of SEPro's default filtering parameter configurations. For this, click on one of the appropriate radio buttons in the lower panel, 'High Resolution MS1' or 'Low Resolution MS1'. Regardless, all SEPro parameters, as described in the 2012 protocol⁴¹, can be set, and they are readily available by clicking on the 'Advanced parameters' tab.

▲ CRITICAL STEP The 'High Resolution MS1' mode is advised for data from instruments that provide less than 20 p.p.m. for MS1 and more than 20k resolution. For example, if an Orbitrap was used to obtain MS1 and an LTQ to obtain the MS2, then the 'High Resolution MS1' option should be chosen; this configuration is also suitable for instruments that provide high-resolution MS2, such as a Q-Exactive HF instrument. The 'Low Resolution MS1' mode is recommended when all data are obtained, for example, on an LTQ-Velos instrument (Thermo, San Jose).

24| Check the 'Include MS2 in results' box in case inclusion of the mass spectra of the identified peptides in the report is desired. This will allow double-clicking on an identification, and thus enabling the spectrum browser to be opened.

▲ CRITICAL STEP If the experiment uses isobaric tags for downstream relative quantification, checking this option is required.

Figure 4 | SEPro's Result Browser. SEPro provides a dynamic report that can be sorted according to any column. The top panel lists protein identifications, and clicking on any one of them causes the lower panel to display all matches associated with the corresponding protein, together with their respective scores. Double-clicking on a protein result brings up a window (in the upper-right corner) displaying a graphical coverage representation, a FASTA coverage representation and a group view (i.e., other proteins that share peptides). By double-clicking on a row in the lower panel, the annotated mass spectrum pops up. The lower-left corner displays one of the many new features in PatternLab for proteomics 4.0; clicking on the 'Tools' menu and then on 'Evaluation of Enzyme Specificity' will display a window informing how many fully specific, semi-specific and nonspecific peptides were identified in the mixture.



25| Select the 'Experiment with more than 50k spectra' option in case it is estimated that there are ~50,000 or more mass spectra in the data; such volume is easily obtained when performing MudPIT experiments or using last-generation instruments (e.g., Orbitrap Elite) with long (3 h or more) gradients. This will make SEPro group identifications according to precursor charge state and enzymatic status (i.e., fully specific and semi-specific) in order to generate discriminatory functions that are independent of both charge state and enzymatic status.

26| Click on the 'Go' button. The user will be redirected to the 'Follow up' tab where the tool's progress is reported.

? TROUBLESHOOTING

27| When the tool finishes processing, click on the 'Result Browser' tab to access the results (**Fig. 4** and **Supplementary Fig. 3**).

28| Save the results by accessing the 'File' menu and then by choosing 'Save SEPro results'. Note that many formats are made available other than SEPro's own; for example, one can save in the DTASelect format¹² or in a tab-delimited file for use with spreadsheet software.

▲ CRITICAL STEP If the user performed a 'Batch Processing' by checking the corresponding box in the entry page, the SEPro results files will be automatically saved to their corresponding directories. Batch processing is useful when there are several directories lying directly one level below a main directory; in this case, the user needs only to specify the path to filter the main directory, select the batch processing option and press the 'Go' button. **Figure 4** shows SEPro's graphical user interface while browsing through filtered results.

Box 2 | Project organization

One of the goals of proteomics is the study of differences in protein expression throughout different biological states. Others include analyzing time series data or samples originating from different tissues. In this regard, PatternLab must be informed which samples come from which biological condition or point in time. The Project Organization module deals with this matter. For example, suppose that one performed a five-point time-course experiment with three biological replicates at each point. Data were acquired using 12-step MudPIT, and now the user wishes to perform relative quantification by spectral counting. This hypothetical experiment would encompass a total of 180 LC-MS/MA files. These files would need to be arranged in directories as follows. First, a directory for each time point would need to be created: for this example, say, T0, T1, T2, T3 and T4. Within each directory, directories for each biological replicate would also need to be created, so, for example, within the T0 directory we would create the directories TOB1, TOB2 and TOB3. (We urge the user not to provide simplified names as, say, B1, because this same name might ambiguously refer to B1 in directory T1 and some modules of PatternLab require each directory to have a unique name.) Finally, within TOB1, for example, the RAW files, SQT files and the sepr file would be placed. We note that this organization can also be arranged before using Comet; in this way, only the main directory would need to be provided and PatternLab would have Comet search within each directory (consequently making the SQT files already appear in the corresponding directories). Similarly, SEPro can perform batch filtering if the main directory is provided. Structuring the files as described enables PatternLab to ultimately compile a PatternLab project file, which contains cross-experiment identification and quantification data; in turn, these are required for downstream analysis. During the next steps of the protocol, the user should decide whether quantification should be performed by spectral counting, by XICs or through reporter ion signals provided by isobaric markers. Although the latter originates from sample preparation, the former two remain an open choice; we recommend using spectral counting for MudPIT experiments and XICs for single shots.



Box 3 | Differential proteomics using the ACFold/T-Fold/Venn diagram modules/principal component analysis ● TIMING <3 s

Once a PatternLab project file is generated, the ACFold or T-Fold³¹ and area-proportional Venn diagram modules can be used for pinpointing differentially expressed proteins and proteins exclusive to a biological condition, respectively. Other modules for performing ANOVA, principal component analysis (PCA) (Búzios) and for analyzing time-course experiment data (TrendQuest) are also available.

These modules are all demonstrated in **Supplementary Video 1**, and they have been described in our previous protocols, so we refer the reader to them⁴¹. Notwithstanding this, we note that these modules' previous versions required the use of the 'index.txt' and 'sparseMatrix.txt' files to store all the identification and quantification data of the experiment. In the current version, they were replaced by a single PatternLab project file, generated in the Project Organization module as explained in **Box 2**. PatternLab for proteomics 4.0 provides a tool for migrating the legacy format to the updated PatternLab project file in the 'Utils' menu.

? TROUBLESHOOTING

29| (Optional) A frequent community request has been for the user to be able to concatenate the results of several SEPro files. To do this, place the desired files in the same directory, select the option 'SEPro Fusion' from the 'Tools' menu and then click on the 'Save new SEPro file' button in the pop-up window. A new SEPro file will be generated that joins the data from all the SEPro files pertaining to that directory.

Quantification analysis using spectral counting, XIC or analysis of multiplex experiments with isobaric tags

30| At this point, it is possible to choose option A for quantification analysis by spectral counting, option B for XIC or option C for analysis of experiments using isobaric labels. For project organization, see **Box 2**. Once this step is finalized, downstream data analysis involving differential proteomics (**Box 3**), scoring phosphopeptide sites (**Box 4**) or analyzing results under the light of the Gene Ontology (**Box 5**) is then possible.

(A) Quantification analysis with spectral counting ● TIMING ~20 s

- (i) *Project organization.* Click on the 'Project Organization' menu, and then on the 'SEPro or PepExplorer' button. The interface will look like that shown in **Figure 5**.
- (ii) Include each directory, prepared as specified in **Box 2**, in the Input Control. For the example in **Figure 5**, two biological conditions were inserted (i.e., BiologicalCondition1 and BiologicalCondition2).
- (iii) Include a brief (~10 words) description of the experiment in the Project Description text box.
- (iv) Click on the 'Load' button.
- (v) To obtain Spectral counting data for downstream analysis, click on the 'Step 2: Spectral Counting' tab. There you can optionally select for NSAF⁵² normalization, and choose whether the quantification will be mapped at the peptide or protein level. Next, click on the 'Go' button, followed by the 'Save PatternLab project' button.

Box 4 | Scoring phosphopeptide localizations with the XD Scoring module ● TIMING ~35 s

Confidently determining phosphorylation sites is crucial to understanding the regulatory mechanisms in biological systems. PatternLab for proteomics 4.0 includes a false-localization rate probabilistic module, termed XD Scoring, that enables unbiased phosphoproteomics studies²⁵. Briefly, the XD Scoring algorithm infers a probabilistic function from the distribution of the identified phosphopeptides' XCorr delta scores (XD scores) and provides *P* values by relying on Gaussian mixture models and a logistic function.

For a mass spectrum whose top-scoring candidate is a phosphopeptide, the XD score is calculated as the difference between the top two XCorr scores of alternative phosphorylation sites in the same peptide sequence. In this regard, for this module to work efficiently, we recommend having the search engine report at least the top 20 scoring candidates in its search results. When using the Comet search in PatternLab, this amounts to editing the line that starts with 'num_output_lines = ' to indicate 20, after clicking on the 'Generate Comet Params' button.

1. Access the XD Scoring module by clicking on the 'Utils' menu and then on 'XD Scoring (Phosphosite)'.
2. Click on the 'Load SQT files' button and select the Comet results files by pressing and holding the 'Ctrl' key while left-clicking on the desired search results files.
3. Click on the 'Calculate' button. A list containing the logarithms of the delta scores for all phosphopeptides will appear in the lower textbox.
4. Click on the 'Generate GMM' button. This will enable PatternLab to generate a Gaussian mixture model whose two Gaussians come from a histogram on the natural logarithms of the XD score. At the bottom of the interface, a green curve shows the cumulative distribution of the green Gaussian and a red curve shows the complementary cumulative distribution of the red Gaussian (**Supplementary Fig. 6**). A complementary logistic function is then generated based on the former two distributions (purple curve). The desired *P* values are given by this function.
5. Specify a SEPro file; this enables the program to output a table associating a *P* value to each site attribution.

Box 5 | The gene ontology explorer ● TIMING ~5 min

The Gene Ontology Explorer (GOEx) allows users to analyze their data under the light of the Gene Ontology; this module has been well documented^{21,40}. In order to analyze the data, a ‘precomp’ object must be generated; this is done by joining the Gene Ontology OBO-format file (available at <http://geneontology.org/page/download-ontology>) with an annotation file. Our original version worked only with annotation files provided at the Gene Ontology website, but the updated GOEx module can work with any organism available in the UniProt base. As this has been the only update to this module, what follows pertains exclusively to the steps for generating a precomp file using UniProt.

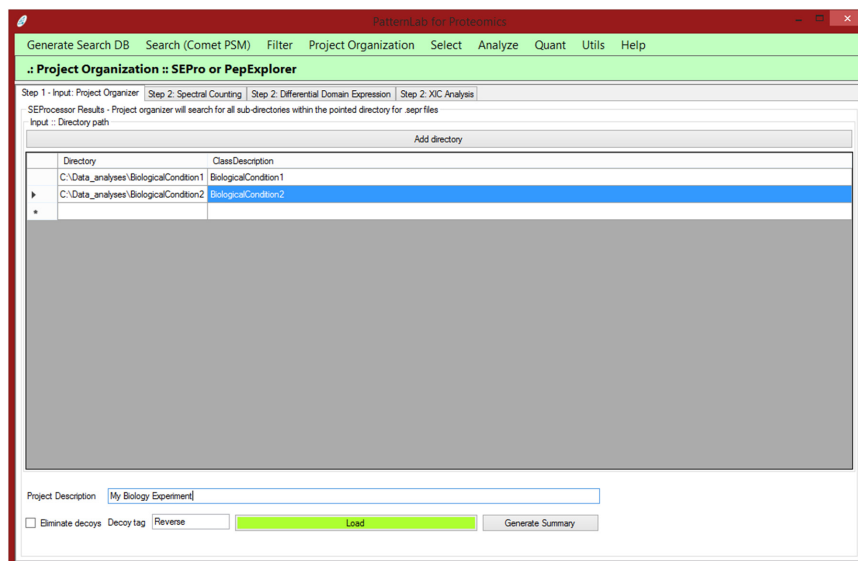
1. Download the data for the desired organism from UniProt as previously described, but instead of selecting the FASTA format choose the text format.
2. Download the latest Gene Ontology OBO file.
3. Access the Gene Ontology by clicking on the ‘Analyze’ menu and then on ‘GOEx (Gene Ontology Explorer)’. The GOEx interface will appear.
4. Click on the ‘Load GO DAG’ button and select the GO.OBO file. This will cause GOEx to perform some optimizations that should take ~2 min.
5. Click on the ‘Load Associations’ button; a window will pop up. The new option for using UniProt text files will be available and selected by default.
6. Click on the ‘Browse for conversion file’ button and load the file downloaded from UniProt.
7. Click on the ‘Save Precomp’ button. The next time a GO analysis is performed, instead of having to repeat all these steps the user can proceed directly to loading the precomp file by clicking on the ‘Load precomp’ button.
8. Refer to the previous publications on GOEx^{36,39} for a complete set of instructions for operating this module.

(vi) Optionally, map spectral counts to protein domains by selecting the ‘Step 2: Differential Domain Expression’ tab. This tab offers controls that enable the generation of a PatternLab project file, as previously described³².

(B) Quantification analysis with XIC ● TIMING ~30–40 s for each mass spectrum raw file

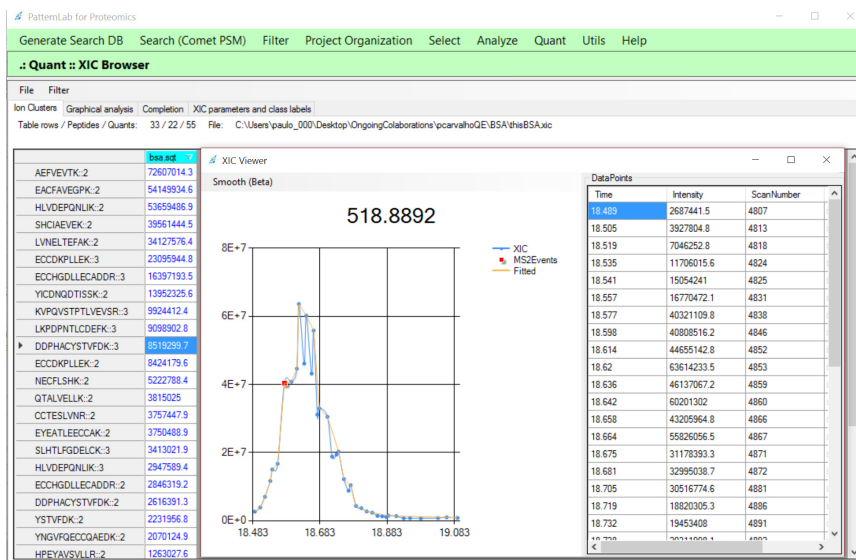
- (i) Follow Step 30A(i–iv).
- (ii) Click on the ‘Step 2: XIC Analysis’ tab if XICs are to be obtained. This tab offers controls that will ultimately produce an XIC file, viewable within PatternLab’s XIC Browser module, which is available through the ‘Quant’ menu by selecting ‘XIC Browser’. The XIC Browser module is then used to generate a PatternLab project file, as described in the Using PatternLab’s XIC Browser section.
- (iii) Click on the ‘Quant’ menu and then select ‘XIC Browser’.
- (iv) Click on ‘File’, and then on ‘Load’ and ‘Bin’ to load an .xic file generated using PatternLab’s Project Organizer. This is a binary file by default, yet the XIC Browser allows files to be saved in the JavaScript Object Notation (JSON), which is a lightweight text-data-interchange format that simplifies the parsing by other software.
- (v) Review the list of cross-experiment identified peptides that will appear as soon as the file finishes loading. Note that each column will be named after a search file (e.g., SQT) and list the XIC values for each peptide. Double-click on an XIC value to open an XIC plot together with a table discriminating the plotted values, as exemplified in **Figure 6**. The table discriminating the quantification values can also be copied and the values pasted onto some spreadsheet software.
- (vi) Click on the ‘Graphical Analysis’ tab to view a histogram of the label-free quantification values for all peptides (**Supplementary Fig. 4**). Note that many experiments can be simultaneously assessed.
- (vii) Optionally, use the XIC Browser to reduce the effects from undersampling. Undersampling is a common problem in proteomics, as not all peptides are sampled by the mass spectrometer. The XIC Browser can help with this limitation by relying on the retention times and precursor masses of peptides identified in a run to estimate the XIC of a peptide

Figure 5 | PatternLab’s Project Organizer. This module is responsible for joining the information of the various biological or technical replicates from all biological conditions. Directories containing results filtered by SEPro should be indicated for each biological condition.



PROTOCOL

Figure 6 | PatternLab's XIC Browser. By clicking on the XIC values (blue numbers), a window displaying the corresponding XIC plot will pop up.



in another run, one in which that peptide was not sampled. To accomplish this, first click on the 'Completion' tab; a list of all liquid chromatography–tandem mass spectrometry (LC-MS/MS) runs in the experiment will be provided in one column, along with another column to which the user can input a number for each run. Label the runs that should be grouped for inferring XICs by placing the same number beside each one (Fig. 7). Finally, click on 'Filter' and then on

'Fill in the gaps'. The new XICs, completed by using the retention times and the precursor masses of peptides identified in compatible runs, will be listed in the XIC Browser in green. Identifications with no XICs, or XICs not passing a minimum quality criterion, will have values of -1 and be listed in red.

- (viii) The same peptide is usually identified through different charge states and consequently with different precursor m/z values. The XIC Browser makes available an option, through the 'Filter' menu and then by selecting 'Retain Optimum Signal', for only the best (higher-value) XICs for a given charge state to be retained. So, for example, if in general the charge-(+2) peptide precursors for a given peptide have XIC values greater than their charge-(+3) counterparts, then all XICs from the latter version of that peptide will be discarded. Arguably, by considering only the more intense XIC versions of the peptide, less noise gets into the model and a more accurate relative quantification can be obtained (data not shown).
- (ix) Click on the 'File' menu followed by 'Save' and then by 'PatternLab project file' to generate a PatternLab project file for downstream analysis.

(C) Analyzing multiplex experiments labeled with isobaric tags ● TIMING 20–50 s for each mass spectrum raw file

▲ **CRITICAL** SEPro files to be analyzed with the 'Isobaric module' must have been processed using the 'include MS2 in results' option.

- (i) Click on the 'Quant' menu, and then on 'Isobaric Analyzer'.
- (ii) If data were acquired according to the MultiNotch approach, extract the MS3 data from the RAW file. For this, click on PatternLab's 'Utils' menu, select the RawReader module, then check the 'MS3' checkbox and the directory containing the mass spectra raw files and click on the 'Go' button. We note that this step can also be accomplished by any software that is capable of extracting MS3 files, such as RawExtractor, for example, made available at <http://fields.scripps.edu/researchtools.php> (ref. 59). Once this is done, click on the 'MultiNotch' tab, specify the path to the SEPro file and to the MS3 directory, and click on the 'Go' button. This procedure will patch the SEPro file to include the MS3 data from the reporter ions so that downstream analysis can be performed.
- (iii) (Optional) Remove multiplexed tandem mass spectra from the data set. This step is recommended for data not acquired using MultiNotch. For this, execute YADA²⁰ with its default configuration on the extracted MS1 and MS2 files. This will generate a corrected batch of MS2 files in which the multiplexed MS2 data have their multiple precursors

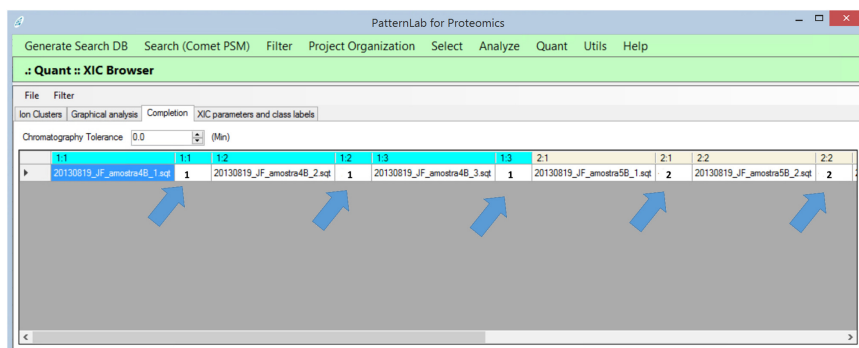


Figure 7 | The XIC Browser's completion tab allows for establishing rules for grouping files that can be used to search for m/z and chromatographic retention times of possibly undersampled peptides. Search results originating from each biological condition have their column header in a different color to facilitate the process. In this example, the user labeled runs from biological conditions 1 and 2 are shown with '1' and '2', respectively, as indicated by the blue arrows. This will make the software use as references only files with the same labels to try and complete the XICs of undersampled peptides.

indicated in the spectrum heading. Then, back in PatternLab's Isobaric Analyzer module, specify the YADA output directory; multiplexed spectra will no longer be considered.

- (iv) Specify the reporter ion masses in the third textbox from the top; predefined masses can be automatically filled in by pressing the 'iTRAQ 4', 'TMT6' or 'iTRAQ8' buttons.
- (v) Specify a data normalization strategy; we strongly recommend using the 'Channel Signal' normalization (default). This normalization adds up the signals of all spectra for each channel (i.e., isobaric marker), and the normalized values for each spectrum are obtained by dividing each reporter ion signal by the corresponding channel's sum.
- (vi) (Optional) Check the 'Apply purity correction' box to correct for the distortions inherent to isobaric tags. These are not 100% pure, and therefore they come with a datasheet per batch, which indicates for each reporter ion reagent the percentages by which its mass differs from the quoted mass by -2, -1, +1 and +2 Da. This enables PatternLab to use Cramer's rule to account for and correct such distortions. If the purity correction numbers provided by the manufacturer differ from those provided in the Isobaric Analyzer's 'Purity Correction' tab, manually alter the values in the software to reflect those provided by the manufacturer. This correction tends to yield very subtle improvements, particularly when compared with the normalization of Step 49.
- (vii) Click on the 'Generate Report' button. This will generate a text file discriminating each peptide contained in the SEPro results, together with its spectral count and redundancy (i.e., how many proteins in the database it matches), followed by the scan numbers and the corresponding normalized TMT or iTRAQ signals in each channel. PatternLab's screen will look like the one in **Supplementary Figure 5**.
- (viii) Generate a PatternLab project file by clicking on the 'PatternLab project file' radio button, and then on the 'Generate Report' button. This file is useful when analyzing experiments with more than two biological conditions.
- (ix) Comparing isobaric tag results from different channels: Click on the 'Two conditions experiment' button; a new window will pop up.
- (x) Specify the 'Class labels' parameter for each channel. As this is a pairwise comparison, only 1 and 2 should be used as labels. In case a channel is not to be included in the statistics, it should be labeled as -1. So, for example, if an iTRAQ 8-plex experiment was carried out, channels 1, 2 and 3 are related to biological condition 1 (i.e., class 1), and channels 5, 6 and 7 are related to class 2. Channels 4 and 8 are not related to the experiment, so the class labels should be 1, 1, 1, -1, 2, 2, 2 and -1, respectively.
- (xi) Click on the 'Browse' button and select the peptide quantification report generated in Step 30C(vii).
- (xii) Press the 'Go' button. The software will load the report and then automatically switch to the next tab, 'Result Browser', and display results as in **Figure 8**.

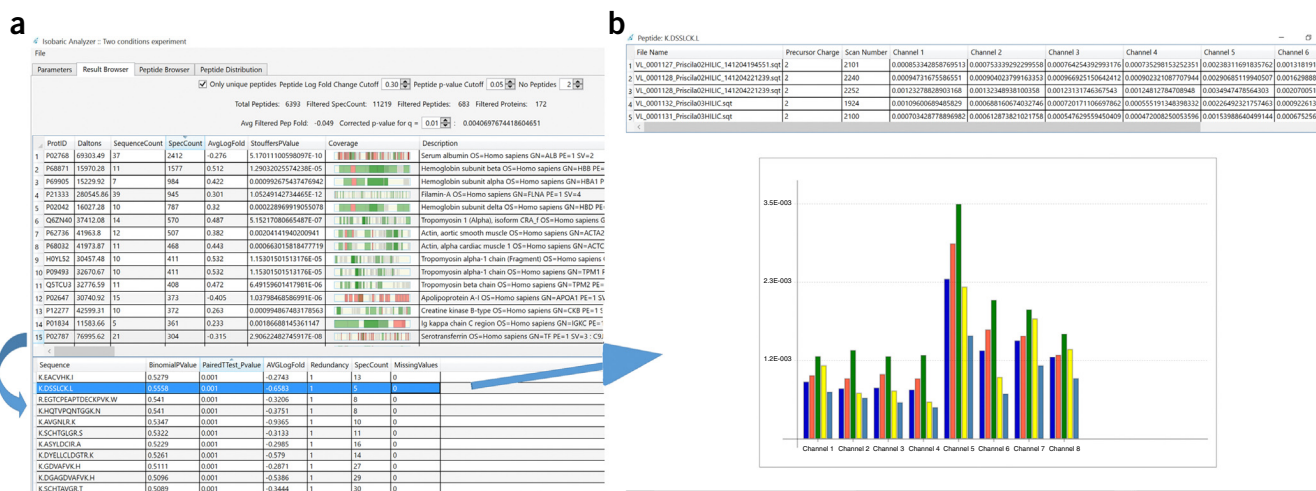


Figure 8 | Result Browser for PatternLab's Isobaric Analyzer, two conditions experiment. (a) The main view when browsing results. The top section displays controls that allow the user to dynamically filter acceptable results according to only unique peptides, only peptides that present an absolute fold change greater than a specified log fold change value, peptides with a binomial or paired *t*-test *P* value lower than a given cutoff and, finally, only proteins containing at least a user-specified number of peptides satisfying these constraints. In what follows, the software reports the total number of peptides identified in the experiment and how many mass spectra, peptides and proteins abide by the cutoff values. The software also suggests a *P* value cutoff at the protein level (corrected *P* value) based on the Benjamini-Hochberg procedure. The upper portion of **a** displays the protein identifications and various details. For example, we note the 'StouffersPValue' column, which represents a meta-analysis of the *P* values of the various peptides belonging to that protein as to whether the protein can be considered as presenting a differential abundance or not. Another key column is 'Coverage', where green sections represent identified peptides with a higher abundance in condition 1, red for condition 2 and gray sections for peptides not satisfying the user-established criteria. When clicking on a protein row, the lower portion of **a** refreshes to provide details, at the peptide level, for that protein. (a,b) Double-clicking on a peptide row (a) causes a window to pop up (b), which displays the reporter ion signals for each pertinent mass spectrum, as exemplified in the lower portion of **b**.



PROTOCOL

(xiii) Specify values for the parameters given in the following table.

Parameter	Description
Only unique peptides	Makes the software consider only peptides that map to one protein in the sequence database
No. of peptides	For example, setting this to 2 means that only proteins that have 2 or more peptides will be considered in the analysis
Peptide log fold change cutoff	Establishes a lower boundary on the absolute value of the natural logarithm of peptides' fold changes. Peptides falling below the bound will be eliminated
Peptide <i>P</i> value cutoff	Peptides whose paired <i>t</i> test or binomial <i>P</i> value does not fall below this cutoff will be eliminated
Corrected <i>P</i> value for <i>q</i>	Allows the user to control the theoretical false-discovery rate by specifying a <i>q</i> value. A corrected <i>P</i> value is calculated according to the Benjamini-Hochberg procedure

(xiv) Click on the 'File' menu, and then on 'Export Protein Results', to export the filtered proteins, together with information on the corresponding peptides, to a text file.

(xv) Click on the 'Peptide Browser' tab to review the list of identified peptides. Recall that peptides appearing only in one biological condition achieve low binomial *P* values. The paired *t*-test *P* value, in contrast, indicates whether the peptide achieved a statistical change in the mean of its reporter ions when comparing the two biological conditions.

(xvi) Click on the 'Peptide Distribution' tab to view a volcano plot at the peptide level. Green circles indicated peptides having a higher abundance in condition 1, and red circles indicate those with a higher abundance in condition 2. The gray translucent circles indicated peptides that did not pass the user-specified criteria. Hover the mouse over a circle to review the pop-ups that discriminate the corresponding peptide sequence, fold change and *P* value. An iTRAQ 8-plex example data set is available for practice. It can be downloaded and the results obtained with it can be compared against those provided on PatternLab's website.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**. If you require help for anything not covered in this protocol, describe the problem in our PatternLab Google group, which is made available through the project's website at <http://patternlabforproteomics.org>, or through the 'Help' menu in the graphical user interface by clicking on 'Troubleshooting and user forum'.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Possible solution
20	Comet tries to read Thermo RAW files and displays the message: 'Retrieving the COM class factory for component with CLSID failed due to the following error: 80040154 Class not registered'	The MSFileReader lib is not installed	Install the MSFileReader, available from Thermo's website
26	The message 'Not enough spectra in decoy or target class to make robust statistic. ANALYSIS WILL BE DISCONTINUED'	There are not sufficient decoy peptides or spectra	Disable the options 'Group by charge state' and/or 'Group by enzymatic no termini' in SEPro's advanced parameter tab
Box 3	There are results from previous versions of PatternLab (i.e., index and sparse matrix) that cannot be opened in the current version	Results must be upgraded to the new PatternLab project file	Use the module 'IndexSparseMatrixLegacy' available in the 'Utils' menu

● TIMING

Steps 1–7, generating a target-decoy sequence database: this step usually takes 5 s of computing time. However, when the 'Eliminate subset sequences' option is selected, time quickly scales up to minutes or even hours, growing quadratically with the number of sequences in the database. For the RefSeq *Homo sapiens* database (20,247 sequences), selecting this option led to ~2 min for the step to complete

Steps 8–19, performing PSM with the integrated Comet search engine: 1–2 min

Step 20, by far, the most time-demanding step is the search itself: search time can range from a few minutes up to more than a day, varying mostly with sample complexity, the number of variable PTMs considered, the mass spectrometer used, LC gradient length and so on, as well as the computer's processor. We exemplify the computational burden of an iTRAQ 8-plex experiment obtained from human biopsies of gastric cancer; two fractions of HILIC were obtained and each analyzed using a 2-h RP chromatography coupled online to an Orbitrap Velos instrument. This example data set and sequence database are made available on PatternLab's website as an exercise to certify that one can reproduce our results as indicated. The search, considering only the fixed modifications of carbamidomethylation of cysteine, and the iTRAQ 8 modification at the N terminus and at the K and Y residues, took 1,035 s on our 24-core (2 × X5675 Xeon) server. All other steps happen almost instantaneously (30 s at most), but users will want to spend time on the modules to assess results (browse through the list of identified proteins and the annotated spectra, experiment with the Gene Ontology and so on)

Steps 21–29, statistically filtering Comet results with SEPro: filtering time can vary greatly according to the experimental design and the number of spectra. It is expected to fall somewhere near 30 s for a typical 2-h LC-MS/MS analysis acquired on an Orbitrap Velos instrument

Step 30A, quantification analysis with spectral counting: computing time should be ~20 s per SEPro file, assuming each file originated from a typical 2-h LC-MS/MS analysis acquired on an Orbitrap Velos

Step 30B, quantification analysis with XIC: computing time should be ~30–40 s for each mass spectrum raw file, assuming that each file originated from a typical 2-h LC-MS/MS analysis acquired on an Orbitrap Velos

Step 30C, analyzing multiplexed experiments labeled with isobaric tags: computing time should be ~20–50 s for each mass spectrum raw file, assuming that each file originated from a typical 2-h LC-MS/MS analysis acquired on an Orbitrap Velos

Box 3, differential proteomics: typically takes <3 s of computing time for any of the modules

Box 4, scoring phosphopeptides: the overall computing time is ~35 s

Box 5, setting up the Gene Ontology Explorer module: generating or loading a .precomp file can take ~5 min. Computing time for exploring one's data is practically negligible

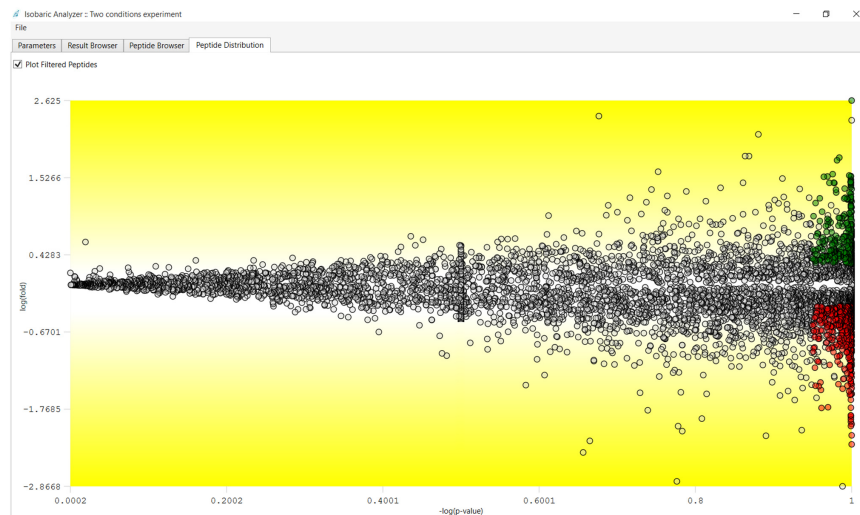
ANTICIPATED RESULTS

PatternLab for proteomics 4.0 is the culmination of the interaction between our group and the proteomics community since 2008. It has been tested on millions of spectra by various groups and aided in the research of a wide range of biological questions. Indeed, PatternLab's goal has been to help scientists identify, quantify and attempt to make sense of the thousands of proteins identified by shotgun proteomics in order to ultimately make a difference in the understanding of biological processes^{61,62}. The present protocol emphasizes only the new features and major changes, including some modules that were replaced with completely re-designed substitutes. For example, PatternLab's new Project Organizer replaces the former 'Regrouper', doing away with the 'index.txt' and 'SparseMatrix.txt' files and introducing the PatternLab project file instead, which is used by many modules for performing quantitative proteomic analyses. The current version also includes a tool, which is accessible through the 'Utils' menu, that allows one to upgrade the legacy format to the new one. In addition, the SEProQ functionalities (XIC and Isobaric browser) were substantially upgraded, and they are now integrated into the same graphical user interface. New modules, such as PepExplorer, whose functionality is similar to that of SEPro but for *de novo* sequencing³³, and the XD Scoring system (**Supplementary Fig. 6**) for phosphopeptide localization, are also part of the new version.

Some representative works illustrating the types of results that can be expected from this protocol are the following. Webb *et al.* used PatternLab to analyze data originating from an online two-dimensional liquid chromatography separation consisting of 39 strong cation-exchange steps followed by a short 18.5-min reversed-phase gradient⁶³. This large-scale data generation approach enabled the identification of 4,269 proteins from 4,189 distinguishable protein families from yeast during log phase growth. In this study, PatternLab's T-Fold module was used to pinpoint differentially abundant proteins, according to spectral counting, during the yeast cellular quiescence, thus providing an overview of most of the yeast proteome. The works from Christie-Oleza *et al.*^{64,65} constitute another example in which PatternLab and spectral counting were used to pinpoint differentially abundant proteins, this time comparing marine bacteria under several natural conditions. Aquino *et al.*⁵ used PatternLab's XIC module to explore the proteomic landscape of a gastric tumor biopsy. In the latter, the biopsy was sectioned into ten parts, and each part was subjected to MudPIT analysis; the authors identified several proteins whose abundance gradually increases/decreases as a function of the distance to the center of the tumor. Chaves *et al.*⁶⁶ used PatternLab's Isobaric analyzer module to analyze TMT data from aging soleus and extensor digitorum longus rat muscles, disclosing quantitative data for more than 4,000 proteins. Finally, Shah *et al.*⁶⁷ used PatternLab's TrendQuest module to group protein expression profiles of *Jatropha curcas* seeds during five developmental stages.

One should always be able, when following a protocol, to reproduce previous results. To help make sure that this is the case, PatternLab's project website (<http://www.patternlabforproteomics.org>) makes available, through its download tab, previously analyzed data sets whose download and re-analysis we recommend strongly to those using PatternLab for the first time. All intermediate files, acquired step by step along the protocol, are also available. The new user can then practice with the protocol to reproduce our results. **Figure 9** exemplifies good results provided by PatternLab's Isobaric module on data

Figure 9 | PatternLab's Isobaric Analyzer. The screenshot shows the result of an analysis. Each dot represents a peptide that is mapped according to its log fold change (y -axis) and its differential abundance P value (x -axis). Peptides colored in green or red are those that satisfied user-specified cutoff criteria for fold-change and P value.



acquired using the MultiNotch approach on TMT-labeled peptides analyzed using an Orbitrap Fusion (Thermo, San Jose). This is so because peptides (dots) are evenly distributed along the y -axis and assume a disposition similar to the eruption of a volcano, thus constituting a so-called volcano plot.

As with any software pipeline or even individual scientist, it is the feedback from collaborators and other peers that drives improvement. In the case of PatternLab, all the feedback, suggestions and even bug fixes have been the most important assets we could count on, helping our suite of tools become more and more sophisticated and hopefully ever closer to supporting answers to questions that were previously intangible. In this regard, we look forward to receiving user feedback through the newly created forum so we can continue to improve on this community-driven and freely available tool.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação do Câncer, Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for its BBP grant and Programa de Apoio à Pesquisa Estratégica em Saúde da Fiocruz (PAPES VII). J.R.Y. acknowledges funding from the US National Institutes of Health (P41 GM103533, R01 MH067880, and R01 MH100175) and the National Heart, Lung and Blood Institute (NHBLI) Proteomics Center at the University of California at Los Angeles (UCLA) (HHSN268201000035C). J.J.M. acknowledges NIH research resources (5P41RR011823) and funding from the National Institute of General Medical Sciences (8 P41 GM103533).

AUTHOR CONTRIBUTIONS P.C.C., J.R.Y. and V.C.B. have participated in the PatternLab project since its beginning in 2008. D.B.L. participated in updating features from several modules and the graphical user interface, as well as in helping migrate to the new PatternLab project file format. F.V.L. developed the PepExplorer module together with P.C.C. M.D.M.S. developed several functions in PepExplorer and had a major participation in the development of the isobaric quantification module. J.S.G.F., P.F.A. and J.J.M. have been participating in PatternLab since early versions by continuously performing beta testing, pointing out required features and providing suggestions on how to make the software more user-friendly. P.C.C. and D.B.L. created the supplementary video. P.C.C. and V.C.B. wrote the manuscript. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

1. Hebert, A.S. *et al.* The one-hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–347 (2014).
2. Yates, J.R. Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* **33**, 1–19 (1998).
3. Zhang, B., Chambers, M.C. & Tabb, D.L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **6**, 3549–3557 (2007).
4. Hwang, S.-I. *et al.* Systematic characterization of nuclear proteome during apoptosis: a quantitative proteomic study by differential extraction and stable isotope labeling. *Mol. Cell. Proteomics* **5**, 1131–1145 (2006).

5. Aquino, P.F. *et al.* Exploring the proteomic landscape of a gastric cancer biopsy with the shotgun imaging analyzer. *J. Proteome Res.* **13**, 314–320 (2014).
6. Calvete, J.J., Sanz, L., Angulo, Y., Lomonte, B. & Gutiérrez, J.M. Venoms, venomomics, antivenomics. *FEBS Lett.* **583**, 1736–1743 (2009).
7. Valente, R.H., Dragulev, B., Perales, J., Fox, J.W. & Domont, G.B. BJ46a, a snake venom metalloproteinase inhibitor. Isolation, characterization, cloning and insights into its mechanism of action. *Eur. J. Biochem* **268**, 3042–3052 (2001).
8. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
9. Washburn, M.P., Wolters, D. & Yates, J.R. III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
10. Köcher, T., Pichler, P., Swart, R. & Mechtler, K. Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients. *Nat. Protoc.* **7**, 882–890 (2012).
11. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
12. Cociorva, D., L Tabb, D. & Yates, J.R. Validation of tandem mass spectrometry database search results using DTASelect. *Curr. Protoc. Bioinformatics* **16** **74**, 13.4.1–13.4.14 (2007).
13. Ross, P.L. *et al.* Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).
14. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591–6596 (1999).
15. Carvalho, P.C., Hewel, J., Barbosa, V.C. & Yates, J.R. III. Identifying differences in protein expression levels by spectral counting and feature selection. *Genet. Mol. Res.* **7**, 342–356 (2008).
16. Liu, H., Sadygov, R.G. & Yates, J.R. III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
17. Neilson, K.A. *et al.* Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* **11**, 535–553 (2011).
18. Shevchenko, A., Valcu, C.-M. & Junqueira, M. Tools for exploring the proteomesphere. *J. Proteomics* **72**, 137–144 (2009).
19. Beausoleil, S.A., Villén, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).



20. Carvalho, P.C. *et al.* YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics* **25**, 2734–2736 (2009).
21. Keller, A., Eng, J., Zhang, N., Li, X. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 2005.0017 (2005).
22. Deutsch, E.W. *et al.* Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin. Appl.* **9**, 745–754 (2015).
23. Kohlbacher, O. *et al.* TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **23**, e191–e197 (2007).
24. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
25. Cox, J. *et al.* A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat. Protoc.* **4**, 698–705 (2009).
26. Carvalho, P.C., Fischer, J.S.G., Chen, E.I., Yates, J.R. & Barbosa, V.C. PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics* **9**, 316 (2008).
27. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
28. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451–1455 (2005).
29. Boekel, J. *et al.* Multi-omic data analysis using Galaxy. *Nat. Biotechnol.* **33**, 137–139 (2015).
30. Egertson, J.D., MacLean, B., Johnson, R., Xuan, Y. & MacCoss, M.J. Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nat. Protoc.* **10**, 887–903 (2015).
31. Carvalho, P.C., Yates, J.R. III. & Barbosa, V.C. Improving the TFold test for differential shotgun proteomics. *Bioinformatics* **28**, 1652–1654 (2012).
32. Leprevost, F.V. *et al.* Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for Proteomics. *J. Proteomics* **89**, 179–182 (2013).
33. Leprevost, F.V. *et al.* PepExplorer: A similarity-driven tool for analyzing *de novo* sequencing results. *Mol. Cell. Proteomics* **13**, 2480–2489 (2014).
34. Fischer, J. *et al.* A scoring model for phosphopeptide site localization and its impact on the question of whether to use MSA. *J. Proteomics* **129**, 42–50 (2015).
35. Fischer, J. *et al.* Dynamic proteomic overview of glioblastoma cells (A172) exposed to perillyl alcohol. *J. Proteomics* **73**, 1018–1027 (2010).
36. Carvalho, P.C. *et al.* GO Explorer: a gene-ontology tool to aid in the interpretation of shotgun proteomics data. *Proteome Sci.* **7**, 6 (2009).
37. Lima, D.B. *et al.* SIM-XL: a powerful and user-friendly tool for peptide cross-linking analysis. *J. Proteomics* **129**, 51–55 (2015).
38. Borges, D. *et al.* Using SIM-XL to identify and annotate cross-linked peptides analyzed by mass spectrometry. *Protoc. Exch.* doi:10.1038/protex.2015.015 (2015).
39. Carvalho, P.C., Yates Iii, J.R. & Barbosa, V.C. Analyzing shotgun proteomic data with PatternLab for proteomics. *Curr. Protoc. Bioinformatics* **30**, 13.13.1–13.13.15 (2010).
40. Carvalho, P.C. *et al.* Search engine processor: filtering and organizing peptide spectrum matches. *Proteomics* **12**, 944–949 (2012).
41. Carvalho, P.C., Fischer, J.S.G., Xu, T., Yates, J.R. III. & Barbosa, V.C. PatternLab: from mass spectra to label-free differential shotgun proteomics. *Curr. Protoc. Bioinformatics* **40**, 13.19.1–13.19.18 (2012).
42. Eng, J.K., Jahan, T.A. & Hoopmann, M.R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
43. Richards, A.L. *et al.* One-hour proteome analysis in yeast. *Nat. Protoc.* **10**, 701–714 (2015).
44. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **41**, D43–D47 (2013).
45. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
46. Cottrell, J.S. & Creasy, D.M. Response to: the problem with peptide presumption and low mascot scoring. *J. Proteome Res.* **10**, 5272–5273 (2011).
47. Bandeira, N. Spectral networks: a new approach to *de novo* discovery of protein sequences and posttranslational modifications. *BioTechniques* **42**, 687 (2007).
48. Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell. Proteomics* **11**, M111.010199 (2012).
49. Shevchenko, A. *et al.* Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73**, 1917–1926 (2001).
50. Xu, T. *et al.* ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol. Cell Proteomics* **5**, S174 (2006).
51. Borges, D. *et al.* Effectively addressing complex proteomic search spaces with peptide spectrum matching. *Bioinformatics* **29**, 1343–1344 (2013).
52. Zybailov, B. *et al.* Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* **5**, 2339–2347 (2006).
53. Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).
54. McAlister, G.C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150–7158 (2014).
55. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).
56. Vizcaino, J.A. *et al.* The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2013).
57. Chambers, M.C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
58. Martens, L. *et al.* mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, R110.000133 (2011).
59. McDonald, W.H. *et al.* MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **18**, 2162–2168 (2004).
60. Nesvizhskii, A.I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).
61. de Miguel, N. *et al.* Proteome analysis of the surface of *Trichomonas vaginalis* reveals novel proteins and strain-dependent differential expression. *Mol. Cell. Proteomics* **9**, 1554–1566 (2010).
62. Clair, G., Armengaud, J. & Dupont, C. Restricting fermentative potential by proteome remodeling: an adaptive strategy evidenced in *Bacillus cereus*. *Mol. Cell. Proteomics* **11**, M111.013102 (2012).
63. Webb, K.J., Xu, T., Park, S.K. & Yates, J.R. Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. *J. Proteome Res.* **12**, 2177–2184 (2013).
64. Christie-Oleza, J.A., Piña-Villalonga, J.M., Bosch, R., Nogales, B. & Armengaud, J. Comparative proteogenomics of twelve *Roseobacter* exoproteomes reveals different adaptive strategies among these marine bacteria. *Mol. Cell. Proteomics* **11**, M111.013110 (2012).
65. Christie-Oleza, J.A., Fernandez, B., Nogales, B., Bosch, R. & Armengaud, J. Proteomic insights into the lifestyle of an environmentally relevant marine bacterium. *ISME J.* **6**, 124–135 (2012).
66. Chaves, D.F.S. *et al.* Comparative proteomic analysis of the aging soleus and extensor digitorum longus rat muscles using TMT labeling and mass spectrometry. *J. Proteome Res.* **12**, 4532–4546 (2013).
67. Shah, M. *et al.* Proteomic analysis of the endosperm ontogeny of *Jatropha curcas* L. seeds. *J. Proteome Res.* **14**, 2557–2568 (2015).

Exploring the Proteomic Landscape of a Gastric Cancer Biopsy with the Shotgun Imaging Analyzer

Priscila Ferreira Aquino,^{†,‡,¶} Diogo Borges Lima,^{‡,¶} Juliana de Saldanha da Gama Fischer,[‡] Rafael Donadelli Melani,[†] Fabio C. S. Nogueira,[†] Sidney R. S. Chalub,[§] Elzalina R. Soares,[⊥] Valmir C. Barbosa,^{||} Gilberto B. Domont,^{†,*} and Paulo C. Carvalho^{‡,*}

[†]Proteomics Unit, Rio de Janeiro Proteomics Network, Department of Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro 21941-909, Brazil

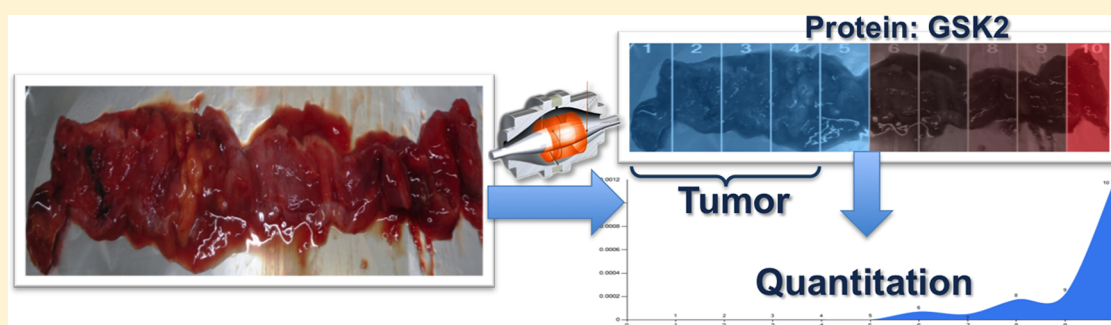
[‡]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz Paraná 81350-010, Brazil

[§]Departamento de Cirurgia Digestiva, Universidade Federal do Amazonas, Amazonas, Brazil

[⊥]Departamento de Química, Universidade Federal do Amazonas, Amazonas, Brazil

^{||}Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro 21941-909, Brazil

S Supporting Information



ABSTRACT: Accessing localized proteomic profiles has emerged as a fundamental strategy to understand the biology of diseases, as recently demonstrated, for example, in the context of determining cancer resection margins with improved precision. Here, we analyze a gastric cancer biopsy sectioned into 10 parts, each one subjected to MudPIT analysis. We introduce a software tool, named Shotgun Imaging Analyzer and inspired in MALDI imaging, to enable the overlaying of a protein's expression heat map on a tissue picture. The software is tightly integrated with the NeXtProt database, so it enables the browsing of identified proteins according to chromosomes, quickly listing human proteins never identified by mass spectrometry (i.e., the so-called missing proteins), and the automatic search for proteins that are more expressed over a specific region of interest on the biopsy, all of which constitute goals that are clearly well-aligned with those of the C-HPP. Our software has been able to highlight an intense expression of proteins previously known to be correlated with cancers (e.g., glutathione S-transferase Mu 3), and in particular, we draw attention to Gastrone-2, a “missing protein” identified in this work of which we were able to clearly delineate the tumoral region from the “healthy” with our approach. Data are available via ProteomeXchange with identifier PXD000584.

KEYWORDS: chromosome 15, shotgun imaging, gastric cancer, Gastrone

■ INTRODUCTION

The study of proteomic landscapes has surfaced as a key technology, with applications in diagnosing and in better understanding pathologies. One of the pioneers of this field, Richard Caprioli, has continually used MALDI (matrix-assisted laser desorption ionization) imaging to pinpoint molecular changes previously undetected by immunohistochemistry or morphological assessments, and thus diagnosed as “normal” but having nevertheless many characteristics in common with tumors.^{1,2} The widely adopted MALDI imaging approach is capable of profiling tissues by describing their landscape as a

group of mass spectral peaks with corresponding intensities. Software applications are then used to perform image analyses to allow the visualization of tissue areas where some mass spectral peak appears that is more commonly found in tumors. One of the limitations of this strategy is that protein identification is fairly compromised, given the limited availability of samples (thin tissue slices) and the challenges

Special Issue: Chromosome-centric Human Proteome Project

Received: September 10, 2013

Published: November 22, 2013

associated with locally trypsinizing and identifying in-depth a variety of components, although efforts have been made to overcome these issues compared to that of shotgun proteomic strategies. Shotgun proteomics is a “bottom-up” approach that enables large-scale protein analysis of almost complete proteomes;³ here we employ a shotgun approach for imaging, yielding as a disadvantage the lack of spatial resolution that MALDI imaging does. Thus, advantages and disadvantages of each methodology make them complementary. Nevertheless, we show shotgun imaging to be an important research tool perfectly suited to collect and supply qualitative and quantitative protein data that are unavailable by other techniques; MALDI and shotgun complement each other.

The multidimensional protein identification technology (MudPIT) embodies a powerful shotgun proteomic strategy that employs two-dimensional liquid chromatography (LC/LC) online with tandem mass spectrometry (MS/MS).⁴ We previously employed this approach to compare gastric cancer biopsies with both resection margins and healthy tissues and concluded that resection margins “looked more proteomically alike” to cancer biopsies than to healthy tissues.⁵ Among the conclusions, we postulated that resection margins play a key role in Paget’s “soil to seed” hypothesis, which states that cancer cells require a special microenvironment to nourish and that understanding it could ultimately lead to more effective treatments. With this as motivation, here we take an additional step by further exploring the spatial proteomic landscape of gastric cancer biopsies. To this end, we sliced a biopsy into 10 sections and then meticulously analyzed each piece by MudPIT. While the widely adopted approach would be to use MALDI imaging to characterize a profile for the tissue at hand, we believe that this would have the drawback of not leading to as large-scale protein identification as that of MudPIT.⁶ Although our proposed data analysis strategy does not offer the spatial resolution from MALDI imaging approaches, it is nonetheless backed up by an in-depth quantitative MudPIT analysis that, in all, we have termed shotgun imaging.

Another motivation for this work stems from the chromosome-centric human proteome project (C-HPP) initiated by the Human Proteome Organization (HUPO), whose goal is to map the entire human protein set, which lacks any mass spectrometric evidence, the so-called “missing proteins”. One of the goals of NeXtProt has therefore been to keep track of which proteins have mass spectrometric identifications. NeXtProt is then helping to pave the way along the C-HPP roadmap by filling in as a human protein knowledgebase that includes data on protein expression from tissues and cells and by tracking the progress in identifying the remaining 3844 missing proteins. With this as motivation, we have taken the foundations of HUPO (C-HPP and B/D-HPP) as core guidelines for our shotgun imaging pipeline and exemplify it with the proteomic profiles of gastric cancer to search for missing proteins and provide information linked to this disease. As a result, we have tailored our imaging analysis software to be able to join MudPIT results with images and to extract information from NeXtProt, aiming to address C-HPP goals such as browsing proteins by chromosomes, easily listing the identifications of missing proteins, and helping to track the ones whose expressions correlate with diseased regions. The entire workflow described herein is integrated into the PatternLab for Proteomics pipeline, which offers an arsenal of tools for analyzing shotgun proteomic data, ranging from

protein quantitation, to differential proteomics, to gene ontology analysis, to name a few.⁷

In summary, here we present a software application that allows for overlaying quantitative shotgun proteomic data onto images. As described above, there are well-established approaches for achieving this goal, but they are not backed up by in-depth protein identification nor easily integrated with a software. Regardless, existing strategies are complementary, and therefore, the purpose determines what methods are best.

■ MATERIALS AND METHODS

Subjects

This study was approved by the Ethics Committee of the Clementino Fraga Filho University Hospital of the Federal University of Rio de Janeiro (HUCFF/UFRJ: MEMO, No. 10252913.5.0000.5257, CAAE). The tumor biopsy was collected at the Oncology Control Foundation Center of the Amazonas State (FCECON) after the signing of informed consent. The biopsy was acquired from an area along the stomach that included tumor and resection margin, during the operation procedures on a male patient. Briefly, the tumor was located in the gastric antrum, and the resection margin was macroscopically defined during the surgery as a 10 cm rim of healthy-looking tissue surrounding the tumor. The biopsy was then subtyped, and the clinical stage of the disease was determined according to the tumor, node, and metastasis (TNM) classification of the American Joint Committee on Cancer (AJCC). The histological type was determined to be adenocarcinoma and classified as T4. Subsequently, the biopsy was vertically divided into 10 sections of equal length to be further analyzed by MudPIT.

Protein Solubilization with RapiGest and Digestion with Trypsin

All biopsy sections were pulverized with liquid nitrogen. Each protein pellet was resuspended independently with RapiGest SF according to the manufacturer’s instructions to a final concentration of 0.1%. The samples were quantified using the Qubit fluorimetric quantitation (Invitrogen) as described in the manufacturer’s instructions. One hundred micrograms of each sample was reduced with 20 mM dithiothreitol (DTT) at 60 °C for 30 min. The samples were cooled to room temperature and incubated in the dark with 66 mM of iodoacetamide (IAA) for 30 min. Afterward, all samples were digested overnight with trypsin (Promega) at the ratio of 1/50 w/w (E/S) at 37 °C. Following digestion, all reactions were acidified with 10% formic acid (1% final concentration) to stop the proteolysis. The samples were centrifuged for 30 min at 20 800 rcf to remove insoluble material.

LC/LC/MS/MS Data Acquisition

Sixty micrograms of the digested peptide mixture was desalted using a reversed-phase column manually packed in a tip using the Poros R2 resin (Applied Biosystems). The desalted peptides were resuspended in solution A (5 mM KH₂PO₄ + 25% ACN pH 3) and loaded into a strong cation exchange microspin column from Harvard Apparatus. Peptides were eluted from the column in a stepwise manner by applying solution A with increasing KCl concentrations of 85, 150, 250, and 400 mM. Each fraction was desalted once again and analyzed on a reversed-phase column coupled to an Orbitrap Velos mass spectrometer (Thermo, San Jose, CA). The peptide mixtures were separated in a 20 cm analytical column (75 μm

inner diameter) that was packed in-house with 3 μm C18 beads (Reprosil-AQ Pur, Dr. Maisch). The flow rate at the tip of the reverse column was 200 nL/min when the mobile phase composition was 5% acetonitrile in 0.1% formic acid. We then applied a 120 min gradient: using first 5–50% acetonitrile in 0.1% formic acid for 100 min, then 50–95% acetonitrile in 0.1% formic acid for 20 min. The effluent from the nLC column was directly electrosprayed into the mass spectrometer.

The LTQ Orbitrap Velos instrument was operated in data-dependent acquisition mode to automatically switch between full scan MS and MS/MS acquisition with a dynamic exclusion of 90 s. For the HCD/CID top 6 method, survey full scan MS spectra (from m/z 350–2000) were acquired in the Orbitrap system with resolution of 60 000 at m/z of 400. The six most intense peptide ions with charge states of 2+ or 3+ were sequentially isolated and fragmented in the CID and HCD collision cells using normalized collision energies of 35 and 50, respectively. The resulting fragments were detected in the Orbitrap system with a resolution of 7500. Standard mass spectrometric conditions for all experiments were as follows: 2.5 kV spray voltage, no sheath and auxiliary gas flow, heated capillary temperature of 235 $^{\circ}\text{C}$, predictive automatic gain control (AGC) enabled, and an S-lens RF level of 70%. Mass spectrometer scan functions and nLC solvent gradients were controlled by the Xcalibur data system (Thermo, San Jose, CA).

Peptide Spectrum Matching

Mass spectra were extracted to the MS2 format using PatternLab's RawReader (available at: <http://proteomics.fiocruz.br/Softwares.aspx>). The NeXtProt database release 2013-07-15 was downloaded. A target decoy database was then generated using PatternLab to include a reversed version of each sequence found in the database plus those from 127 common mass spectrometry contaminants. The ProLuCID search engine (v 1.3) was used to compare experimental tandem mass spectra against those theoretically generated from our sequence database and to select the most likely peptide spectrum matches (PSMs).⁸ Briefly, the search was limited to fully and semitryptic peptide candidates and imposed carbamidomethylation and oxidation of methionine as fixed and variable modification, respectively. The search engine accepted peptide candidates within a 50 ppm tolerance from the measured precursor m/z and used the XCorr and Z-Score as the primary and secondary search engine scores, respectively.

Assessment of PSMs

The validity of the PSMs was assessed using the Search Engine Processor (SEPro)⁹ (v 2.1.0.23). Identifications were grouped by charge state (+2 and \geq 3) and then by tryptic status (fully tryptic, semitryptic), resulting in four distinct subgroups. For each result, the ProLuCID XCorr, DeltaCN, and Z-Score values were used to generate a Bayesian discriminator. The identifications were sorted in nondecreasing order according to the discriminator score. A cutoff score was established to accept a false-discovery rate (FDR) of 1% based on the number of labeled decoys. This procedure was independently performed on each data subset, resulting in an FDR that was independent of tryptic status or charge state. Additionally, a minimum sequence length of six amino acid residues was required. Results were postprocessed to only accept PSMs with less than 5 ppm and proteins supported by two or more independent evidence (e.g., identification of a peptide with different charge states, a modified and a nonmodified version of

the same peptide, or two different peptides). This last filter led to a 0% FDR in all search results. For our quantitative shotgun imaging analysis, we only considered proteins having at least one unique peptide identified.

Relative Protein Quantitation

The MS1 was extracted with RawReader and deconvoluted using YADA's default settings for bottom-up shotgun proteomics.¹⁰ Extracted ion chromatograms (XICs) were obtained utilizing SEPro's Quantitation module as previously described.⁷ In what followed, PatternLab's Regrouper module normalized the quantitative data according to the distributed normalized ion abundance factor (dNIAF)⁵ approach, which is an adaptation of dNSAF for XICs.¹¹ Finally, PatternLab's index and sparse matrix files were generated;⁷ briefly, these are text files that summarize all the results from all MudPIT runs (i.e., identification and quantification) and serve as input to our imaging software together with NeXtProt and the tumor's image.

Shotgun Imaging Software

We have developed a shotgun imaging software to allow the visual assessment of proteomic results. We henceforth refer to it as the shotgun imaging analyzer (SIA). SIA was implemented in C# NetFramework 4.5 for the Windows operating system (7 or later) and is available for download at <http://proteomics.fiocruz.br/software/shotgunImaging>. Its graphical user interface (GUI) displaying an image of the biopsy used in this work is shown in Figure 1. The GUI contains two main tabs: one for

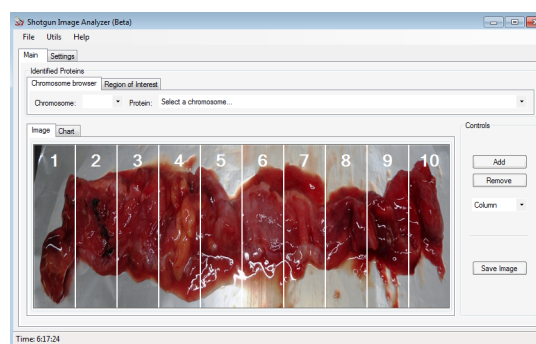


Figure 1. Shotgun imaging graphical user interface displaying an image of the gastric biopsy analyzed in this work.

viewing the results and the other for entering the index and sparse matrix files generated with PatternLab for Proteomics,⁷ NeXtProt, and a path to an image file of the biopsy to overlay an expression heat map. From there on, the user can then generate the heat map's wireframe, indicating the locations where the biopsy was sectioned. Technically, this translates into each section corresponding to a class label of the sparse matrix that should be generated according to a previously published bioinformatics protocol.⁷

SIA has been tailored to take advantage of NeXtProt and therefore fits well in the context of C-HPP. A pull-down menu enables the browsing of identified proteins according to a chromosome of interest. By proceeding in this way, only proteins identified in the experiment at hand for that respective chromosome will be listed. Once a protein is selected, the GUI will overlay a heat map using a gray/light-bluish scale to indicate regions (lanes) where the protein was not identified and different tones of red to indicate its relative abundance in the lanes where it was identified. Quantitative values are made

available in a table located in the Utils menu. Another option is that this menu allows a report to be displayed containing the identified proteins that were marked as missing by NeXtProt.

Once all the data have been loaded through the GUI and the biopsy sections have been determined using the controls on the right, the user can save all information into a unified file for sharing with colleagues by means of the Save option available in the File menu; this file has an “.sip” extension for shotgun imaging project. Finally, SIA offers an option to automatically search for proteins that present a higher expression in a region of interest (ROI). For the experiment at hand, this translates into informing the software the numbers of the lanes that correspond to the tumor in order to enable the sorting by SIA of the identified proteins in a nonincreasing order according to the Golub index (i.e., the difference of the means of the quantitation values divided by sum of their standard deviations).

RESULTS AND DISCUSSION

We have described SIA, a software for shotgun imaging that enables the correlation of high-throughput identification and quantitation of proteins with a visual assessment of their locations. SIA follows a chromosome-centric approach that enables the listing of identified proteins according to the chromosomes to which they correspond. Moreover, it can automatically pinpoint proteins that show a higher expression in a given ROI (e.g., the diseased area) and is aligned with some of the goals of C-HPP.^{12,13} Our analysis has identified up to 1936 proteins (with redundancy), of which 657 have at least one unique peptide. Our identifications are made available through the *.sepr files and can be viewed with the Search Engine Processor viewer; more on their availability in the end of this article; Supporting Information Table 1 lists the proteins identified in each of the 10 biopsy sections. A desired feature in the quest for novel biomarker candidates is to pinpoint tissue-specific proteins.

Gastrokinins have been described as abundant proteins that are specifically expressed in the superficial gastric epithelium and have high evolutionary conservation. In particular, we draw attention to the identification of gastrokine isoform 2 (GKN2), a chromosome 2 protein that lacks mass spectrometric evidence of expression according to NeXtProt and therefore belongs to the list of the so-called missing proteins. Most importantly, GKN2 poses as a key protein for gastric cancer prognosis according to complementary approaches (e.g., western, PCR, microarray).^{14–16} That said, as far as we know, this is the first report to demonstrate the protein expression of GKN2 and its use for delineating the tumoral region using mass spectrometric evidence. Interestingly, GKN2, an abundant protein for the tissue at hand, is down-regulated and many times described as absent in gastric cancer.¹⁷ In a previous report, Moss and collaborators used a transcriptional profiling approach to describe a decreased expression of GKN2 and its relation to tumor prognosis; the authors reported an expression loss of GKN2 in 85% of diffuse intestinal-type cancers and correlated this with a significantly worse outcome ($p < 0.03$).¹⁶ More recently, Mao and collaborators corroborated these results using RT-PCR, Western blot, and immunohistochemistry for GKN1.^{14,15} They further discuss that this isoform function is to protect and maintain the integrity of the gastric epithelium, and that a restoration of GKN1 expression can suppress the gastric cancer cell viability. Our shotgun imaging result for GKN2 is found in Figure 2; we note that the shotgun image

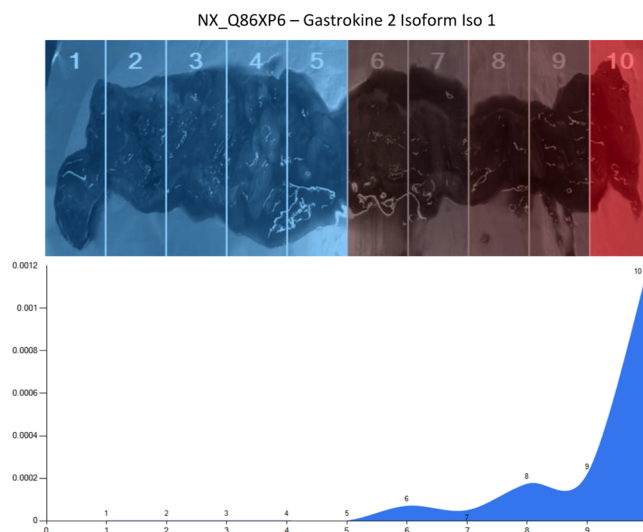


Figure 2. Shotgun imaging analyzer result for protein NX_Q86XP6, Gastrokine 2 isoform. The top panel refers to the overlaid quantitation heat map. The bluish colors indicate absence of NX_Q86XP6, darker red (e.g., as found in lane 10) is correlated with a higher expression of this protein by XIC. In the panel below, the *y*-axis represents the distributed normalized spectral abundance factor values (i.e., quantitation values), and the *x*-axis was aligned to the shotgun image to correspond to the slice numbers. A spline regressor was used to plot the area chart in blue. Lanes 1–4 are the tumoral regions, and regions 6–10 were classified as nontumoral.

identification of GKN1 isoform is very similar to that of Figure 2 (data not shown). Our identification of GKN2 contains four unique peptides, of which we exemplify the mass spectrum of one of them in Figure 3. Results clearly demonstrate increasing expressions in slices further apart from the tumoral region (1–4).

It is worthy to note that we also identified trefoil factor isoform 1 (TFF1) from chromosome 21 (NX_P04155), a protein that has been found in high levels in the upper gastric mucosal cells and known to interact with GKN2. Their modes of action remain unsolved; however, there was a recent demonstration of the existence of a GKN2–trefoil factor 1 heterodimer.¹⁸ TFF1 expression was found in all tissue sections except in lane 3, which lies within the tumoral region. Although we cannot make a statistical claim (i.e., $p > 0.05$), the average values of DNIAF are 1.7 higher in the nontumoral region. This result is supported by the literature that states that TFF1 should be absent (or under-expressed) in the tumor area.¹⁶

As previously described, SIA can sort all protein identifications according to the Golub index when comparing differences in ROIs, which for the case at hand are sections 1–4 (the tumor ROIs). Among the top-ranking proteins in the results of the automatic search, we point out annexin, glutathione S-transferases Mu 3, and lumican. We discuss these proteins next and then take further note of the role played by tropomyosins in tumor development and of their presence in the biopsy in question.

Annexin is one of the calcium and membrane-binding proteins. The literature correlates this protein with a wide variety of cellular functions, such as membrane aggregation, phagocytosis, proliferation, apoptosis, and even tumorigenesis.¹⁹ A lower abundance of annexin has been associated with a poor prognosis in prostate cancer,²⁰ head and neck cancer,²¹ sinonasal adenocarcinomas,²² and hepatocellular cancer.²³

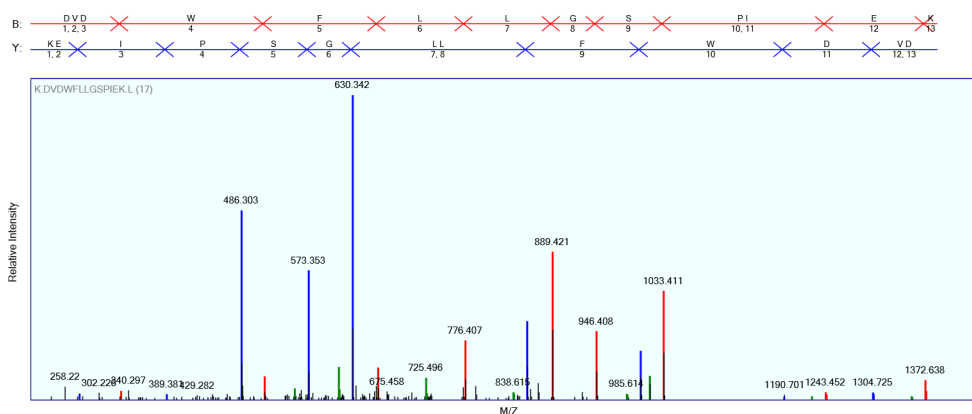


Figure 3. Tandem mass spectrum of the tryptic peptide K.DVDWFLGSPIEK.L uniquely found in the GKN-2 sequence.

Interestingly, at the molecular level, our results have shown annexin A1 (chromosome 9, NX_P04083) to be less abundant in cancer ROIs, which is in good agreement with previous work that, by complementary approaches such as DNA microarrays, immunoblotting, and immunohistochemistry,^{19,24,25} pinpoint the annexin gene as a potential marker for gastric cancer. While the role of this protein in gastric cancer is still unknown, there is increasing evidence that changes in annexin expression and its subcellular localization can contribute to the development and progression of the disease.

Glutathione S-transferases (chromosome 1) constitute a class of proteins that have been correlated with detoxification and a protective mechanism against the development of cancer. The gene family for this class has been shown to be highly polymorphic, being expressed in only 50% of individuals.^{26–28} A previous study correlated an elevated expression of this protein's mRNA with a poor prognosis.²⁹ Our results indicate a higher abundance of glutathione S-transferases Mu 3 in the tumoral ROIs.

Lumican is one of the small leucine-rich proteoglycans (SLRPs), with molecular weight of about 40 kDa comprised in four domains.³⁰ Previous works suggest that it plays a key role in cell–cell adhesion, in regulating stromal collagen matrix assembly, and it may also serve as a regulatory molecule of several cellular functions, such as cell migration and proliferation, apoptosis, differentiation, and inflammatory responses.^{30–32} In cancers, lumican expression has been correlated with tumoral growth and metastasis in colorectal and pancreatic tumors, benign prostatic hyperplasia, and especially in breast tumors, where it has been detected in stromal cells adjacent to tumor cells.^{33,30,34,35} Furthermore, high expression levels of this protein can be associated with high pathological tumor grades.³⁵ Its role in tumorigenesis remains elusive, but recent studies demonstrate that overexpression of lumican suppresses transformation induced by v-src and v-K-ras.^{30,34} These observations may be explained by the possibility that the lumican receptors mediate different signaling transduction pathways in a cell-type-specific manner, which in turn provides a scaffold for macrophage infiltration.^{30,35,36} In our imaging results, lumican expression is mostly found in the tumoral area.

Tropomyosins are components of actin filaments that play a critical role in regulating the interaction of actin and myosin. They are, therefore, important modulators of the adhesion dynamics that determines cell migration.³⁷ Previous studies demonstrate that different tropomyosin isoforms perform

distinct physiological roles, and that their expression profile is profoundly changed during the malignant progression of cancer cells since such cells change the organization of cytoskeletal and cellular morphology.^{38,39} In some types of cancers, such as breast cancer, a low expression of these proteins in the primary carcinoma has been detected, while on the other hand, high levels have been observed in metastatic tumors.⁴⁰ Similar results are reported for gastric cancer in the literature.³⁹ In our results, tropomyosins were found with elevated and approximately uniform expression levels throughout all biopsy sections.

We would also like to point out that in this work we identified E-cadherin only in the resection margin; this result is aligned with our previous report.⁵ In the literature, this protein has been associated with an important role in the early stage of tumorigenesis by modulating intracellular signaling which can promote tumor growth and be associated with metastasis.⁴¹

Finally, we wish to conclude with some remarks concerning the methodology we have used. First of all, we fully concur with the wider community in recognizing that MALDI imaging, with its remarkable spatial resolution, is arguably the current gold standard for exploring proteomic landscapes. On the other hand, we do also recognize that electrospray approaches are emerging as competitors. An example comes from the recently published work headed by L. Eberlin and R. Graham Cooks, in which they propose the use of ambient mass spectrometry (DESI) in a lipidomic approach to the intraoperative molecular diagnosis of brain tumors.⁴² Moreover, their work demonstrates the potential of DESI-MS to identify the histological type of brain tumors. These authors' approach includes a statistical classifier trained from the same mass spectrometry imaging used for histopathology diagnosis. Another example is a work led by Vladislav Petyuk and Richard Smith, where the authors utilized shotgun proteomics with electrospray ionization to profile a mammalian brain; they were able to produce shotgun images with in-depth identification but bound by similar spatial resolution as in our work.^{43,44} Reports such as this demonstrate the ever-growing importance of in-depth, high-throughput biomolecular mapping correlated with spatial disposition in potentially impacting medical treatment. While our approach exhibits poor spatial resolution, it is backed up by MudPIT, itself the gold standard in terms of number of identifications, which enabled us to identify a key missing protein and its partner related to gastric cancer. Taken together, these three strategies (MALDI, DESI, and shotgun imaging) can work in complementary ways to aid in developing more effective treatments for diseases.

Availability

SIA and the project file analyzed in this work are available through our site, <http://proteomics.fiocruz.br/software/shotgunImaging>. By downloading the project file into SIA, one can view shotgun imaging heat maps for any of our protein identifications. The RAW data, together with all ProLuCID (*.sqt) and SEPro files, are available at <http://dm64.ioc.fiocruz.br/sia/>. The mass spectrometry proteomics data have also been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository⁴⁵ with the data set identifier PXD000584.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary Table 1. A list of identified proteins for each of the 10 tumor sections. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: paulo@pcarvalho.com. Phone: +55(41)3316-3230. Fax: +55(41)3316-3267.

*E-mail: gilberto@iq.ufjr.br. Phone: +55(21) 2562-7353.

Author Contributions

[¶]P.F.A. and D.B.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

P.C.C. was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and RPT02H PDTIS/Carlos Chagas Institute-Fiocruz Parana. D.B.L. was funded by Instituto de Biologia Molecular do Paraná. V.C.B. acknowledges partial support from CNPq, CAPES, and a FAPERJ BBP grant. G.B.D. acknowledges support from CNPq (BP 308819/2011) and FAPERJ (E-26/110.138/2013). P.F.A. acknowledges scholarship support from CAPES.

■ REFERENCES

- (1) Chaurand, P.; Norris, J. L.; Cornett, D. S.; Mobley, J. A.; Caprioli, R. M. New developments in profiling and imaging of proteins from tissue sections by MALDI mass spectrometry. *J. Proteome Res.* **2006**, *5* (11), 2889–2900.
- (2) Oppenheimer, S. R.; Mi, D.; Sanders, M. E.; Caprioli, R. M. Molecular analysis of tumor margins by MALDI mass spectrometry in renal carcinoma. *J. Proteome Res.* **2010**, *9* (5), 2182–2190.
- (3) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., III. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394.
- (4) Washburn, M. P.; Ulaszek, R.; Deciu, C.; Schieltz, D. M.; Yates, J. R., III. Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem.* **2002**, *74* (7), 1650–1657.
- (5) Aquino, P. F.; Fischer, J. S.; Neves-Ferreira, A. G.; Perales, J.; Domont, G. B.; Araujo, G. D.; Barbosa, V. C.; Viana, J.; Chalub, S. R.; Lima de Souza, A. Q.; Carvalho, M. G.; Leao de Souza, A. D.; Carvalho, P. C. Are gastric cancer resection margin proteomic profiles more similar to those from controls or tumors? *J. Proteome Res.* **2012**, *11* (12), 5836–5842.
- (6) Kim, H. K.; Reyzer, M. L.; Choi, I. J.; Kim, C. G.; Kim, H. S.; Oshima, A.; Chertov, O.; Colantonio, S.; Fisher, R. J.; Allen, J. L.; Caprioli, R. M.; Green, J. E. Gastric cancer-specific protein profile

identified using endoscopic biopsy samples via MALDI mass spectrometry. *J. Proteome Res.* **2010**, *9* (8), 4123–4130.

- (7) Carvalho, P. C.; Fischer, J. S.; Xu, T.; Yates, J. R., III; Barbosa, V. C. PatternLab: From Mass Spectra to Label-Free Differential Shotgun Proteomics. In *Current Protocols in Bioinformatics*; Wiley: New York, 2012; Chapter 13, Unit 13-19.

- (8) Xu, T.; Venable, J. D.; Park, S. K.; Cociorva, D.; Lu, B.; Liao, L.; Wohlschlegel, J.; Hewel, J.; Yates, J. R., III. ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol. Cell. Proteomics* **2006**, *5*, S174.

- (9) Carvalho, P. C.; Fischer, J. S.; Xu, T.; Cociorva, D.; Balbuena, T. S.; Valente, R. H.; Perales, J.; Yates, J. R., III; Barbosa, V. C. Search engine processor: Filtering and organizing peptide spectrum matches. *Proteomics* **2012**, *12* (7), 944–949.

- (10) Carvalho, P. C.; Xu, T.; Han, X.; Cociorva, D.; Barbosa, V. C.; Yates, J. R., III. YADA: A tool for taking the most out of high-resolution spectra. *Bioinformatics* **2009**, *25* (20), 2734–2736.

- (11) Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L. Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem.* **2010**, *82* (6), 2272–2281.

- (12) Aebbersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12* (1), 23–27.

- (13) Huhmer, A. F.; Paulus, A.; Martin, L. B.; Millis, K.; Agreste, T.; Saba, J.; Lill, J. R.; Fischer, S. M.; Dracup, W.; Lavery, P. The chromosome-centric human proteome project: a call to action. *J. Proteome Res.* **2013**, *12* (1), 28–32.

- (14) Mao, W.; Chen, J.; Peng, T. L.; Yin, X. F.; Chen, L. Z.; Chen, M. H. Downregulation of gastrokine-1 in gastric cancer tissues and restoration of its expression induced gastric cancer cells to apoptosis. *J. Exp. Clin. Cancer Res.* **2012**, *31* (1), 49.

- (15) Mao, W.; Chen, J.; Peng, T. L.; Yin, X. F.; Chen, L. Z.; Chen, M. H. *Helicobacter pylori* infection and administration of non-steroidal anti-inflammatory drugs down-regulate the expression of gastrokine-1 in gastric mucosa. *Turk. J. Gastroenterol.* **2012**, *23* (3), 212–219.

- (16) Moss, S. F.; Lee, J. W.; Sabo, E.; Rubin, A. K.; Rommel, J.; Westley, B. R.; May, F. E.; Gao, J.; Meitner, P. A.; Tavares, R.; Resnick, M. B. Decreased expression of gastrokine 1 and the trefoil factor interacting protein TFIZ1/GKN2 in gastric cancer: influence of tumor histology and relationship to prognosis. *Clin. Cancer Res.* **2008**, *14* (13), 4161–4167.

- (17) Oien, K. A.; McGregor, F.; Butler, S.; Ferrier, R. K.; Downie, I.; Bryce, S.; Burns, S.; Keith, W. N. Gastrokine 1 is abundantly and specifically expressed in superficial gastric epithelium, down-regulated in gastric carcinoma, and shows high evolutionary conservation. *J. Pathol.* **2004**, *203* (3), 789–797.

- (18) Menhenniott, T. R.; Kurklu, B.; Giraud, A. S. Gastrokines: stomach-specific proteins with putative homeostatic and tumor suppressor roles. *Am. J. Physiol. Gastrointest. Liver Physiol.* **2013**, *304* (2), G109–G121.

- (19) Cheng, T. Y.; Wu, M. S.; Lin, J. T.; Lin, M. T.; Shun, C. T.; Huang, H. Y.; Hua, K. T.; Kuo, M. L. Annexin A1 is associated with gastric cancer survival and promotes gastric cancer cell invasiveness through the formyl peptide receptor/extracellular signal-regulated kinase/integrin beta-1-binding protein 1 pathway. *Cancer* **2012**, *118* (23), 5757–5767.

- (20) Ding, T.; Yang, L.; Wang, Y.; Yuan, J.; Chen, T.; Cai, X. Down-regulation of annexin II in prostate cancer is associated with Gleason score, recurrence, metastasis and poor prognosis. *Mol. Med. Rep.* **2010**, *3* (5), 781–787.

- (21) Garcia Pedrero, J. M.; Fernandez, M. P.; Morgan, R. O.; Herrero, Z. A.; Gonzalez, M. V.; Suarez, N. C.; Rodrigo, J. P. Annexin A1 down-regulation in head and neck cancer is associated with epithelial differentiation status. *Am. J. Pathol.* **2004**, *164* (1), 73–79.

- (22) Rodrigo, J. P.; Garcia-Pedrero, J. M.; Llorente, J. L.; Fresno, M. F.; Allonca, E.; Suarez, C.; Hermsen, M. Down-regulation of annexin A1 and A2 protein expression in intestinal-type sinonasal adenocarcinomas. *Hum. Pathol.* **2011**, *42* (1), 88–94.

- (23) Liu, S. H.; Lin, C. Y.; Peng, S. Y.; Jeng, Y. M.; Pan, H. W.; Lai, P. L.; Liu, C. L.; Hsu, H. C. Down-regulation of annexin A10 in hepatocellular carcinoma is associated with vascular invasion, early recurrence, and poor prognosis in synergy with p53 mutation. *Am. J. Pathol.* **2002**, *160* (5), 1831–1837.
- (24) Emoto, K.; Sawada, H.; Yamada, Y.; Fujimoto, H.; Takahama, Y.; Ueno, M.; Takayama, T.; Uchida, H.; Kamada, K.; Naito, A.; Hirao, S.; Nakajima, Y. Annexin II overexpression is correlated with poor prognosis in human gastric carcinoma. *Anticancer Res.* **2001**, *21* (2B), 1339–1345.
- (25) Mussunoor, S.; Murray, G. I. The role of annexins in tumour development and progression. *J. Pathol.* **2008**, *216* (2), 131–140.
- (26) Alves, G. M. Glutathione S transferase mu polymorphism and gastric cancer in the Portuguese population. *Biomarkers* **1998**, *3* (6), 441–447.
- (27) Martinez, C.; Martin, F.; Fernandez, J. M.; Garcia-Martin, E.; Sastre, J.; Diaz-Rubio, M.; Agundez, J. A.; Ladero, J. M. Glutathione S-transferases mu 1, theta 1, pi 1, alpha 1 and mu 3 genetic polymorphisms and the risk of colorectal and gastric cancers in humans. *Pharmacogenomics* **2006**, *7* (5), 711–718.
- (28) Kocevar, N.; Odreman, F.; Vindigni, A.; Grazio, S. F.; Komel, R. Proteomic analysis of gastric cancer and immunoblot validation of potential biomarkers. *World J. Gastroenterol.* **2012**, *18* (11), 1216–1228.
- (29) Kearns, P. R.; Chrzanoska-Lightowlers, Z. M.; Pieters, R.; Veerman, A.; Hall, A. G. Mu class glutathione S-transferase mRNA isoform expression in acute lymphoblastic leukaemia. *Br. J. Haematol.* **2003**, *120* (1), 80–88.
- (30) Kao, W. W.; Funderburgh, J. L.; Xia, Y.; Liu, C. Y.; Conrad, G. W. Focus on molecules: lumican. *Exp. Eye Res.* **2006**, *82* (1), 3–4.
- (31) Kao, W. W.; Liu, C. Y. Roles of lumican and keratocan on corneal transparency. *Glycoconjugate J.* **2002**, *19* (4–5), 275–285.
- (32) Gu, G.; Cheng, W.; Yao, C.; Yin, J.; Tong, C.; Rao, A.; Yen, L.; Ku, M.; Rao, J. Quantitative proteomics analysis by isobaric tags for relative and absolute quantitation identified lumican as a potential marker for acute aortic dissection. *J. Biomed. Biotechnol.* **2011**, 920763.
- (33) Kelemen, L. E.; Couch, F. J.; Ahmed, S.; Dunning, A. M.; Pharoah, P. D.; Easton, D. F.; Fredericksen, Z. S.; Vierkant, R. A.; Pankratz, V. S.; Goode, E. L.; Scott, C. G.; Rider, D. N.; Wang, X.; Cerhan, J. R.; Vachon, C. M. Genetic variation in stromal proteins decorin and lumican with breast cancer: investigations in two case-control studies. *Breast Cancer Res.* **2008**, *10* (6), R98.
- (34) Yoshioka, N.; Inoue, H.; Nakanishi, K.; Oka, K.; Yutsudo, M.; Yamashita, A.; Hakura, A.; Nojima, H. Isolation of transformation suppressor genes by cDNA subtraction: lumican suppresses transformation induced by v-src and v-K-ras. *J. Virol.* **2000**, *74* (2), 1008–1013.
- (35) Naito, Z. Role of the small leucine-rich proteoglycan (SLRP) family in pathological lesions and cancer cell growth. *J. Nippon Med. Sch.* **2005**, *72* (3), 137–145.
- (36) Funderburgh, J. L.; Mitschler, R. R.; Funderburgh, M. L.; Roth, M. R.; Chapes, S. K.; Conrad, G. W. Macrophage receptors for lumican. A corneal keratan sulfate proteoglycan. *Invest. Ophthalmol. Visual Sci.* **1997**, *38* (6), 1159–1167.
- (37) Calmettes, C. Medullary cancer of the thyroid. Apropos of 20 years' experience in France. *Ann. Endocrinol. (Paris)* **1988**, *49* (1), 10–16.
- (38) Stehn, J. R.; Schevzov, G.; O'Neill, G. M.; Gunning, P. W. Specialisation of the tropomyosin composition of actin filaments provides new potential targets for chemotherapy. *Curr. Cancer Drug Targets* **2006**, *6* (3), 245–256.
- (39) He, Q. Y.; Cheung, Y. H.; Leung, S. Y.; Yuen, S. T.; Chu, K. M.; Chiu, J. F. Diverse proteomic alterations in gastric adenocarcinoma. *Proteomics* **2004**, *4* (10), 3276–3287.
- (40) Lee, H. H.; Lim, C. A.; Cheong, Y. T.; Singh, M.; Gam, L. H. Comparison of protein expression profiles of different stages of lymph nodes metastasis in breast cancer. *Int. J. Biol. Sci.* **2012**, *8* (3), 353–362.
- (41) Chan, A. O. E-cadherin in gastric cancer. *World J. Gastroenterol.* **2006**, *12* (2), 199–203.
- (42) Eberlin, L. S.; Norton, I.; Orringer, D.; Dunn, I. F.; Liu, X.; Ide, J. L.; Jarmusch, A. K.; Ligon, K. L.; Jolesz, F. A.; Golby, A. J.; Santagata, S.; Agar, N. Y.; Cooks, R. G. Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (5), 1611–1616.
- (43) Petyuk, V. A.; Qian, W. J.; Smith, R. D.; Smith, D. J. Mapping protein abundance patterns in the brain using voxelation combined with liquid chromatography and mass spectrometry. *Methods* **2010**, *50* (2), 77–84.
- (44) Petyuk, V. A.; Qian, W. J.; Chin, M. H.; Wang, H.; Livesay, E. A.; Monroe, M. E.; Adkins, J. N.; Jaitly, N.; Anderson, D. J.; Camp, D. G.; Smith, D. J.; Smith, R. D. Spatial mapping of protein abundances in the mouse brain by voxelation integrated with high-throughput liquid chromatography-mass spectrometry. *Genome Res.* **2007**, *17* (3), 328–336.
- (45) Vizcaino, J. A.; Cote, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; O'Kelly, G.; Schoenegger, A.; Ovelleiro, D.; Perez-Riverol, Y.; Reisinger, F.; Rios, D.; Wang, R.; Hermjakob, H. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **2013**, No. 41, D1063–D1069.

PepExplorer: A Similarity-driven Tool for Analyzing *de Novo* Sequencing Results[§]

Felipe V. Leprevost[‡], Richard H. Valente^{§¶}, Diogo B. Lima[‡], Jonas Perales^{§¶}, Rafael Melani^{||}, John R. Yates III^{**}, Valmir C. Barbosa^{‡‡}, Magno Junqueira^{||}, and Paulo C. Carvalho^{‡§§}

Peptide spectrum matching is the current gold standard for protein identification via mass-spectrometry-based proteomics. Peptide spectrum matching compares experimental mass spectra against theoretical spectra generated from a protein sequence database to perform identification, but protein sequences not present in a database cannot be identified unless their sequences are in part conserved. The alternative approach, *de novo* sequencing, can make it possible to infer a peptide sequence directly from a mass spectrum, but interpreting long lists of peptide sequences resulting from large-scale experiments is not trivial. With this as motivation, PepExplorer was developed to use rigorous pattern recognition to assemble a list of homologue proteins using *de novo* sequencing data coupled to sequence alignment to allow biological interpretation of the data. PepExplorer can read the output of various widely adopted *de novo* sequencing tools and converge to a list of proteins with a global false-discovery rate. To this end, it employs a radial basis function neural network that considers precursor charge states, *de novo* sequencing scores, peptide lengths, and alignment scores to select similar protein candidates, from a target-decoy database, usually obtained from phylogenetically related species. Alignments are performed using a modified Smith–Waterman algorithm tailored for the task at hand. We verified the effectiveness of our approach using a reference set of identifications gener-

ated by ProLuCID when searching for *Pyrococcus furiosus* mass spectra on the corresponding NCBI RefSeq database. We then modified the sequence database by swapping amino acids until ProLuCID was no longer capable of identifying any proteins. By searching the mass spectra using PepExplorer on the modified database, we were able to recover most of the identifications at a 1% false-discovery rate. Finally, we employed PepExplorer to disclose a comprehensive proteomic assessment of the *Bothrops jararaca* plasma, a known biological source of natural inhibitors of snake toxins. PepExplorer is integrated into the *PatternLab for Proteomics* environment, which makes available various tools for downstream data analysis, including resources for quantitative and differential proteomics. *Molecular & Cellular Proteomics* 13: 10.1074/mcp.M113.037002, 2480–2489, 2014.

Very often, groundbreaking discoveries with a significant impact on the biotechnological and biomedical fields have emerged from studying “non-canonical” organisms. For example, the study of *Thermus aquaticus* allowed us to ultimately pave the way to modern molecular biology with the characterization of that organism’s thermostable DNA polymerase (1). The characterization of the green fluorescent protein in *Aequoria victoria* led to a revolution in cellular biology and to a Nobel Prize being awarded to Osamu Shimomura, Martin Chalfie, and Roger Tsien. In Brazil, Sergio Ferreira’s work on the venom of the Brazilian poisonous snake *Bothrops jararaca* enabled the development of the first angiotensin-converting enzyme inhibitor drug (Captopril) for the treatment of hypertension (2).

In scenarios such as these, proteomics has the potential to allow a better understanding of the complexity of biological systems and the process of evolution than the study of the genetic code alone. It enables the characterization of molecular processes according to their protein content, facilitating new discoveries. In proteomics, the most frequently used strategy for protein identification is so-called peptide spectrum matching (PSM),¹ or the comparison of experimental mass spectra ob-

From the [‡]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil; [§]Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil; [¶]Instituto Nacional de Ciência e Tecnologia em Toxinas (INCTTox/CNPq), Brazil; ^{||}Proteomics Unit, Rio de Janeiro Proteomics Network, Department of Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil; ^{**}Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California; ^{‡‡}Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Received December 13, 2013, and in revised form, April 28, 2014
Published, MCP Papers in Press, May 30, 2014, DOI 10.1074/mcp.M113.037002

Author contributions: F.V.L., R.H.V., M.J., and P.C.C. designed research; F.V.L., R.H.V., D.L.B., J.P., R.M., J.R.Y., V.C.B., M.J., and P.C.C. performed research; F.V.L., R.H.V., J.P., J.R.Y., and P.C.C. contributed new reagents or analytic tools; F.V.L., R.H.V., D.L.B., J.P., R.M., V.C.B., M.J., and P.C.C. analyzed data; F.V.L., R.H.V., V.C.B., and P.C.C. wrote the paper; R.M. extensively tested the software; M.J. provided key insights making this research possible.

¹ The abbreviations used are: PSM, peptide spectrum matching; FDR, false-discovery rate; RBF-NN, radial basis function neural network; SEPro, Search Engine Processor; PAM, point accepted mutation.

tained by fragmenting peptides in a mass spectrometer to theoretical spectra generated from a sequence database. In general, the identification process follows from the sequence whose theoretical spectrum yields the highest matching score according to some empirical or probabilistic function. Examples of search engines adopting this strategy are SEQUEST (3), X!Tandem (4), and Mascot (5).

Back in the 1990s, establishment of a cutoff score for confident identification relied mostly on user experience; for example, given a specific charge state, Washburn *et al.* established cross-correlation and deltaCn cutoff values for SEQUEST in order to allow the selection of a subset of confident identifications from LCQ data. This has since been termed “the Washburn criterion.” In what followed, target-decoy databases were implemented to allow for more sophisticated refinements in filtering the data (6). In 2007, Elias and Gygi published a seminal paper on the target-decoy approach to shotgun proteomics (7) that ultimately firmed this approach as a standard and motivated the development of several statistical filters capable of converging to a list of confident identifications satisfying a user-specified false-discovery rate (FDR) with significantly more sensitivity than the conservative Washburn criterion. Such statistical filters include mixtures of probabilities (8), quadratic discriminant analysis (9), semi-supervised learning with support vector machines (10), and Bayesian logic (11) using a semi-labeled decoy analysis to account for overfitting (12). With so many advances, the PSM workflow has become the gold standard, as it is very sensitive and the least error-prone method when a database is available with the corresponding proteins. The latter factor limits the application of PSM to those organisms for which accurate sequence databases have been established. If a peptide’s sequence is not contained within the sequence database, it cannot be identified via the PSM method. However, efforts in developing error-tolerant PSM approaches such as implemented in Mascot have made it possible to handle minor sequence modifications constrained by a simple set of rules. Nevertheless, increasing the search space in the PSM approach leads to decreased sensitivity (13).

Even though the concept of computer-aided *de novo* sequencing predates that of PSM (14), advances in the quality of mass spectrometry data and the power of computer hardware have allowed it to reemerge at the heart of a highly active field. *De novo* sequencing is unbiased insofar as it is not constrained by a sequence database, and it is therefore complementary to PSM. However, it has remained the most error prone of the two methods (15). The challenges of *de novo* sequencing notwithstanding, a few recent and notable improvements in computer-aided *de novo* analysis are PepNovo (16), which combines graph theory with machine learning; pNovo+ (17), which is optimized for high-resolution HCD data; NovoHMM (18), relying on hidden Markov models for increased sensitivity; and PEAKS (19), which creates a spectrum graph model by performing dynamic programming on

the mass values regardless of the presence of an observed fragment ion. By considering the complementarities of different fragmentation strategies (*e.g.* collision induced dissociation, electron transfer dissociation (20), and electron capture dissociation (21)), computational proteomics scientists have also demonstrated significant advances in *de novo* accuracy (22). In particular, the Bandeira group has continually pushed the limits and redefined the notion of what *de novo* sequencing can do by introducing the spectral networks paradigm (23–25). Briefly, this strategy can assemble mass spectra into spectral pairs by joining overlapping spectra obtained from sample aliquots digested by different enzymes. As a consequence, it reduces noise and significantly improves protein coverage. Its latest version also combines data from different fragmentation techniques.

These algorithm developments have improved *de novo* sequencing, shifting the bottleneck to post-sequence processing of data. This is because the output of *de novo* software is a long list of highly similar full and partial peptide sequence and scores. An initial attempt to overcome these limitations consisted of a tag approach that was a hybrid of *de novo* sequencing and database searching: short sequence tags were derived from tandem mass spectra and used to search a sequence database (26). In what followed, a modified version based on the FASTA homology search tool was proposed for homology-driven proteomics (27). This strategy was implemented as part of the CIDentify tool, whose novelty was to account, in the alignment score, for limitations of mass spectrometry sequencing such as switching between leucine and isoleucine or other combinations of amino acids having the same mass. The next steps were taken mainly by the Shevchenko group through the introduction of the MS-Blast algorithm, which relies on a different set of scores and uses the PAM30MS substitution matrix, itself tailored for mass-spectrometry-based proteomics (28, 29). For a complete review of *de novo* sequencing and homology searching, we suggest Ref. 30.

The current *de novo* post-processing paradigm presents several limitations that are similar to those of the early PSM workflow. Output files generally consist of a peptide list with corresponding scores, demanding an experienced user to assess trustworthy identifications. If the same peptide is analyzed by different mass spectrometers, different scores might be generated, which makes data comparison between different groups a challenging task. In a sense, problems are similar to those encountered when adopting the early Washburn criterion. Assembling protein information from a list of peptides is not a simple task, and usually it is not performed using state-of-the-art *de novo* tools. Although there are great tools for doing this at the PSM level, there is still a lack of similar tools for *de novo* sequencing.

To tackle the aforementioned shortcomings, and in line with our strong interest in diversity-driven proteomics (29), we present a methodology for post-processing *de novo* sequenc-

ing data that allows inference of protein identification through statistical mapping of *de novo* sequencing results to a protein sequence database. Our approach begins with the use of Gotoh's version of the Smith–Waterman algorithm, based on affine gap scoring (31) for increased scalability, to align *de novo* sequences against those in a target-decoy database. Then a radial basis function neural network (RBF-NN) is used to rank results according to alignment score, *de novo* score, precursor charge state, and peptide length. Finally, a heuristic method is used to present protein identification results in a user-friendly, interactive report. The resulting algorithm was implemented as the software PepExplorer. In essence, its goal is somewhat similar to that of post-processing tools such as DTASelect (9), Percolator (10), and SEPro (11), but with an extra layer of complexity inherent from *de novo* sequencing. PepExplorer can handle the output of several widely adopted *de novo* tools, such as PepNovo, pNovo+, and PEAKS, and accepts a generic format to enable result analysis from a broader range of tools once results are run through simple parsers. Similarly, the software accepts a series of database formats for input analysis. These features are not found in other tools. PepExplorer is freely available to the scientific community and is provided with the necessary documentation.

The effectiveness of our methodology has been verified in two distinct scenarios, the first a real but controlled experiment and the other pertaining to comprehensive profiling of the plasma components of *Bothrops jararaca*, a venomous viper endemic to Brazil, southern Paraguay, and northern Argentina. The first scenario's purpose was to validate the effectiveness of the tool in analyzing a published *Pyrococcus furiosus* dataset (11). We note that this organism is recognized by the proteomics community as well suited for benchmarking, because it allows for the rigorous testing of identification algorithms at the peptide and protein levels (32, 33). We modified the *P. furiosus* sequence database in such a way that no more peptides were identified via the PSM approach or another widely adopted error-tolerant search tool, Mod-A (34). We then found that we could recover protein identifications using our tool. The *B. jararaca* scenario has allowed us to explore uncharted territory, as this organism has an incomplete sequence database and we were therefore required to rely on those of orthologous organisms. In particular, *B. jararaca* plasma was chosen because it is a main research model studied at the Laboratory of Toxinology (FIOCRUZ, Brazil), and several natural inhibitors of snake toxins have already been identified/characterized from this biological matrix (35–37).

MATERIALS AND METHODS

Bothrops jararaca Plasma Sample Preparation—*B. jararaca* plasma was supplied to the Laboratory of Toxinology (FIOCRUZ, Brazil) during the experimental procedures described in the research project, approved by the Ethics Committee of the Butantan Institute (555/2008), of the Biomedical Science Institute of the University of São Paulo (138/2009). This project was also approved by the Brazilian Institute for Environment and Renewable Natural Resources, a Bra-

zilian Ministry of the Environment's enforcement agency (IBAMA, License 01/2009). Protein concentration was determined via bicinchoninic acid assay (38), and 40 μg were processed to complete dryness via lyophilization. Next, 20 μl of a 0.25% (m/v) *RapiGest* SF (Waters) in 50 mM ammonium bicarbonate solution were added to solubilize the proteins, which were then heated for 5 min at 100 °C. Disulfide bridges were reduced with 20 mM dithiothreitol for 30 min at 60 °C and subjected to cysteine alkylation with 68 mM iodoacetamide for 15 min at room temperature in the dark. Four microliters of a 0.2- $\mu\text{g}/\mu\text{l}$ (in 50 mM acetic acid) porcine trypsin solution (catalog number V511, Promega) were added, and incubation proceeded for 22.5 h at 37 °C followed by 45 min at 56 °C. The reaction was stopped by the addition of 2.4 μl of 5% (v/v) trifluoroacetic acid in water. For *RapiGest* removal, samples were incubated for 45 min at 37 °C and centrifuged at 16,000 $\times g$ for 10 min at room temperature. The supernatant was collected and desalted/concentrated with in-house-made columns packed with Poros R2 resin (Invitrogen), eluted with 60% acetonitrile in 0.1% (v/v) trifluoroacetic acid, and completely dried using a SpeedVac (Thermo Scientific) vacuum centrifuge concentrator. Samples were resuspended in 30 μl of 1% (v/v) formic acid and submitted to a 10-min ultrasonic bath cycle before analysis via nano-LC-MS/MS.

Mass Spectrometry Analysis—The sample was analyzed in technical triplicate via LC-MS/MS. Tryptic digests were separated via reversed-phase capillary liquid chromatography coupled to nano-electrospray high-resolution mass spectrometry for identification. For each sample, 2 μl of desalted tryptic peptide digest were initially applied to a 2-cm-long (100- μm internal diameter) trap column packed with 5- μm , 200 Å Magic C18 AQ matrix (Michrom Bioresources) and then separated on a 30-cm-long (75- μm internal diameter) column that was packed with the same matrix, directly on a self-pack 15- μm PicoFrit empty column (New Objective). Chromatography was carried out on an EASY-nLC II instrument (Thermo Scientific). Samples were loaded onto the trap column at 2000 nL/min while chromatographic separation occurred at 200 nL/min. Mobile phase A consisted of 0.1% (v/v) formic acid in water, and mobile phase B consisted of 0.1% (v/v) formic acid in acetonitrile. Gradient conditions were as follows: 2% to 40% B during 162 min; up to 80% B in 4 min; and maintenance at this concentration for 2 min. Eluted peptides were directly introduced to the LTQ XL/Orbitrap mass spectrometer (Thermo, San Jose, CA) for analysis. The source voltage was set at 1.9 kV, the capillary temperature at 200 °C, and the tube lens voltage at 100 V. Fourier transform MS full and multi-stage MS automatic gain control target values were set at 500,000 and 200,000, respectively. MS1 spectra were acquired on the Orbitrap analyzer (300 to 1700 m/z) at a 60,000 resolution (for m/z 445.1200). We acquired tandem mass spectra from the 10 most intense ions by means of HCD fragmentation (minimum signal of 10,000 required; isolation width of 2.0; normalized collision energy of 45.0; and activation time of 30 s) followed by MS2 acquisition on the Orbitrap analyzer at 15,000 resolution. The dynamic exclusion option was enabled and set with the following values for each parameter: repeat count = 1; repeat duration = 30 s; exclusion list size = 500; exclusion duration = 45 s; and exclusion mass width = 10 ppm. Charge state rejection was enabled for unassigned charges and for those equal to 1.

Preparation of Sequence Databases Used for Similarity-driven Identification and PSM—Reference sequences for *P. furiosus* were obtained from UniProt, and those for *Reptilia* together with *Amphibia* are from the NCBI RefSeq; all were downloaded in June 2013. The sequences obtained from the *Reptilia* and *Amphibia* databases were merged into a single structure and then joined by 127 sequences of common mass spectrometry contaminants, as well as, for each database entry, a reversed version of the corresponding sequence

(a decoy sequence). The final *P. furiosus* and *Reptilia* plus *Amphibia* databases had 4347 and 302,287 sequences, respectively.

Three *P. furiosus* proof-of-concept databases were generated by repeatedly adding “mutations” and insertions to the sequence database. These databases are referenced as PFU_Gap25_Substitution15, PFU_Gap20_Substitution10, and PFU_Gap15_Substitution8. In the PFU_Gap25_Substitution15 database, for example, an amino acid was inserted at every 25th position, and every 15th amino acid was replaced by some other, randomly chosen amino acid. These databases provide increasing distance from the original database and thereby presented the algorithms with different levels of difficulty. As no new proteins were added to obtain any of them, each of these databases has the same number of sequences as the initial *P. furiosus* database. Our goal has been to modify the native sequences from an organism’s database to simulate the appearance of different, but phylogenetically close, organisms that would render PSM useless.

Peptide Spectrum Matches and Quality Assessment—The mass spectra were exported to the MS2 format (39) from the RAW files using PatternLab’s RawReader module. The ProLuCID (40) search engine was used to compare experimentally obtained spectra against theoretical spectra generated from a sequence database and select the most similar. Briefly, the search was limited to fully tryptic peptide candidates, as we imposed only carbamidomethylation as a fixed modification. The search engine accepted peptide candidates within a 50-ppm tolerance from the measured precursor *m/z* and 550 ppm for the MS2, and we used XCorr and ZScore as the primary and secondary search engine scores, respectively.

The validity of the peptide spectrum matches was assessed using the Search Engine Processor (SEPro) (11). Identifications were grouped by charge state ($\leq +2$ and $> +2$), resulting in two distinct subgroups. For each result, the ProLuCID XCorr, DeltaCN, and ZScore values were used to generate a Bayesian discriminator. The identifications were sorted in nondecreasing order according to the discriminator score. A cutoff score was established to accept an FDR of 1% based on the number of labeled decoys. This procedure was independently performed on each data subset, resulting in a false-positive rate that was independent of charge state. Additionally, an amino acid sequence at least six residues long was required. Results were post-processed to only accept matches with less than 5 ppm and proteins supported by at least two spectral counts. This last filter led to a 0% FDR in the search results.

De Novo Sequencing—*De novo* sequencing was performed using PEAKS Studio 6.0 (Bioinformatics Solutions Inc., ON, Canada). The parent mass error tolerance was 7 ppm, and the fragment mass error tolerance was 0.05 Da. Carbamidomethylation of cysteine was considered as a fixed modification. Acceptable results required an average local confidence score of at least 70% and a total local confidence score of at least 5 and were exported to a CSV file using the export option built into the software.

Blind Post-translational Modification Search with Mod-A—Mod-A was used to search the original and modified versions of the PFU dataset using its automatic precursor mass detection mode and allowing for arbitrary modifications in the peptides. The parameter files used by Mod-A are included in the [supplemental material](#).

PepExplorer Algorithm—The PepExplorer algorithm was coded in C# 4.5. It has a graphical user interface but can also be executed from the command prompt, which enables it to work in cluster environments. The algorithm’s workflow can be summarized in four steps: *de novo* result parsing, sequence alignment, result filtering, and result presentation (Fig. 1). Below we detail each of these steps. All parameters can be adjusted using the graphical user interface (Fig. 2).

De Novo Result Parsing—PepExplorer currently contains parsers for three widely adopted *de novo* sequencing algorithms: PepNovo,

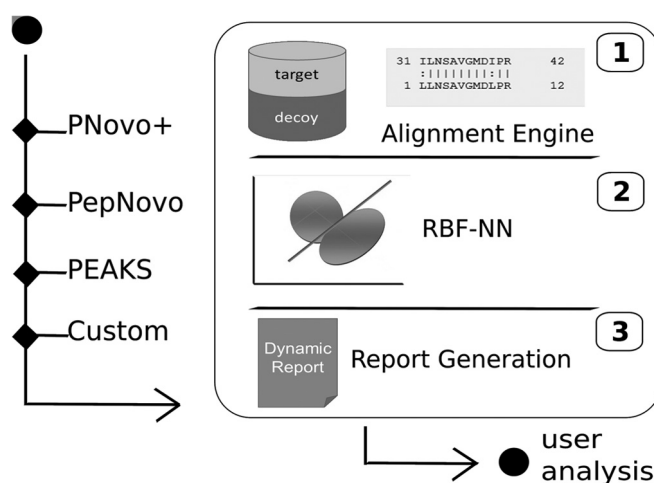


FIG. 1. A *de novo* tool is used to generate candidate sequences from mass spectra. The output from the *de novo* tool, together with a target-decoy database, serves as input to PepExplorer. PepExplorer uses the Smith–Waterman algorithm to align the *de novo* sequences against the target-decoy database (1). An RBF neural network is employed to rank the *de novo* alignments according to a confidence score that takes into account the *de novo* sequencing score, the alignment score, and the number of amino acids contained in the peptide (2). Finally, a dynamic report is generated (3).

pNovo+, and Peaks. PepExplorer treats the *de novo* algorithm with an abstraction layer that allows for the inclusion of new parsers upon request. The software also allows one to analyze a list of peptides by copying and pasting them in the corresponding text box found in the *de novo* output box (Fig. 2). However, in this scenario its neural network runs in a simplified mode and does not consider precursor charge states, scan numbers, or *de novo* score quality.

Sequence Alignment—PepExplorer relies on Gotoh’s version of the Smith–Waterman algorithm (31), built into its core for aligning peptide sequences against a target-decoy sequence database specified by the user. The user can specify several alignment parameters, such as the open gap and extend gap penalties, the number of *de novo* sequence results to be considered per spectrum, and a substitution matrix of choice. For this study these values were 13, 5, 1, and the PAM30MS matrix, respectively.

These default open gap and extend gap parameters resulted from a grid search also made available in PepExplorer through the “Advanced Analysis” menu. This enables the algorithm to explore the landscape of combinations of these two parameters within user-predefined bounds and report the combination yielding the greatest number of alignments under a user-defined FDR. For this work, we performed the grid search in the PFU dataset allowing both the open gap and the extend gap penalties to vary from 2 to 30. The grid search results are available as part of the online supplementary files in the software’s website.

Result Filtering with the RBF-NN—Each obtained sequence alignment is internally treated by PepExplorer as an alignment object containing the following properties: peptide length, *de novo* score, precursor charge, number of gaps, identifier, similarity, and alignment scores. All these parameters are used for result quality assessment, together with complementary information relevant to report assembly, such as scan number, raw file name, and details on the sequence alignment.

As a first step, these alignment objects are separated into two lists: those originating from peptide ions with charge state less than or equal to 2, and those from peptide ions with charge states greater

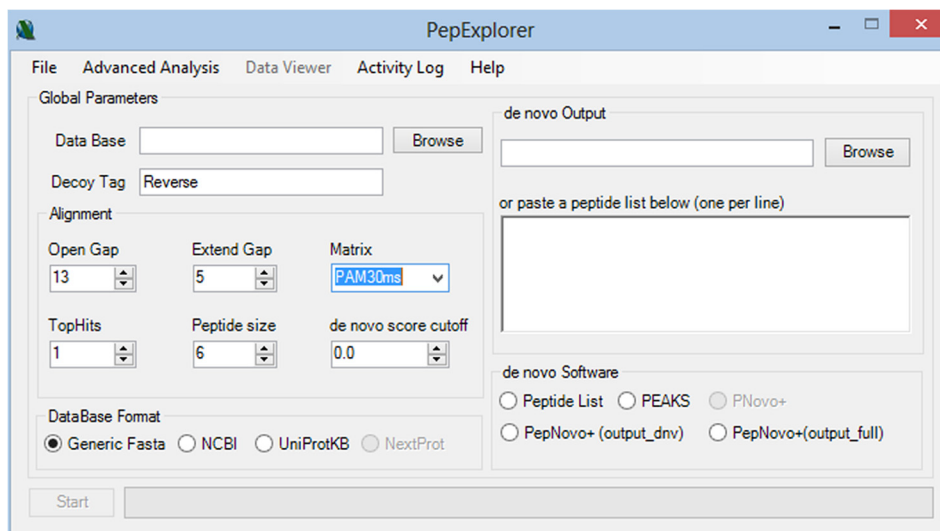


FIG. 2. The PepExplorer graphical user interface.

than 2. Each of these lists is handled by a different RBF-NN. This enables convergence to a list of alignment objects satisfying an FDR that is independent of precursor charge state.

Given a list of alignment objects, the RBF-NN is based on defining six clusters; to this end, PepExplorer relies on the k-means++ algorithm (41) applied to the normalized values (*i.e.* between 0 and 1) of the alignment score, the *de novo* score, and the peptide length of each alignment object. We note that k-means++ employs a “careful seeding” to address the NP-hard problem of minimizing the average squared distance between points in the same cluster. The “careful seeding” is performed by choosing the first cluster center randomly from among the data points to be clustered. Subsequent cluster centers are chosen from locations coinciding with the remaining data points with probability proportional to each point’s squared distance to the closest existing center. The algorithm then continues with the established k-means optimization procedure. The initial “careful seeding” is justified by the resultant faster convergence and better solutions. PepExplorer runs k-means++ 50 times in search of the best solution.

The RBF-NN is then trained to capture the nonlinear relationship between target and decoy alignment objects. The network we used was a single-hidden-layer feed-forward neural network whose three input nodes forwarded the input signal to the hidden nodes directly, with no weights. The kernel transfer function used in the j th hidden node was

$$\varphi_j(x) = \exp\left(\frac{-P_x - \mu_j P^2}{2\sigma_j^2}\right) \quad (\text{Eq. 1})$$

where μ_j is the j th cluster center determined by k-means++ and σ_j is a width parameter given by the smallest Euclidean distance between any two cluster centers. The latter is used to better capture the localness and thus the smoothness and continuity of the fitted function. The connections from the six hidden nodes to the single output node are weighted, and the value of the output node is given by

$$y(x) = \sum_{j=1}^6 w_j \varphi_j(x), \quad (\text{Eq. 2})$$

where w_j is the connection weight between the j th hidden node and the output node. During training, $y(x)$ is either +1 or -1, depending respectively on whether the alignment object in question corresponds to a target sequence or a decoy sequence (alignment objects map-

ping to sections of sequences found in both target and decoy sequences are not considered). The weights of the RBF-NN equations are determined by means of linear regression using a least-squares objective function. All identifications are sorted in a nondecreasing order according to the classification function. Finally, a cutoff score can be established to achieve an FDR based on the decoy identifications.

Result Presentation—Results are presented in the form of a dynamic, interactive report that allows the user to sort them according to a criterion of choice and interact with the report by setting parameters of interest. The report can quickly adjust to a user-specified FDR or provide a list of maximum-parsimony alignments, as all alignments are stored to enable the algorithm to quickly converge to various settings. Among the threshold parameters we highlight the global FDR, the minimum alignment count (the closest to spectral count), the maximum alignment parsimony, the use of distinct RBF-NN for precursors of different charge states, and the minimum identifier. The report is provided as two interactive panels, the upper one being related to protein information and the lower to identification data. The upper panel provides information such as protein identifier, protein length, coverage percentage, sequence count, alignment count, and description. When a protein is selected, detailed information is made available in the lower panel of all alignments that mapped to it such as the scan number, file name, *de novo* score, precursor charge state, identifier, similarity, number of gaps, alignment score, sequence found in the database, and sequence provided by the *de novo* sequencing tool (Fig. 3). When a row of interest is selected in this lower panel, a new window displaying the sequence alignment is made available. In this window, when a row is selected in the upper panel with the protein information, a graphical coverage report is displayed (Fig. 4). This report is integrated with the cloud service of PatternLab for Proteomics (42), enabling the use of the Infer Domains function to instantly access predicted on-demand protein domains inferred with HMMER3 over Pfam-A (43).

RESULTS

PFU Proof of Concept—A Venn diagram showing the overlap of the protein identifications from ProLuCID/SEPro, Mod-A, and PepExplorer on the unmodified PFU database is found in Fig. 5. We recall that only proteins having two or more spectral counts were considered.

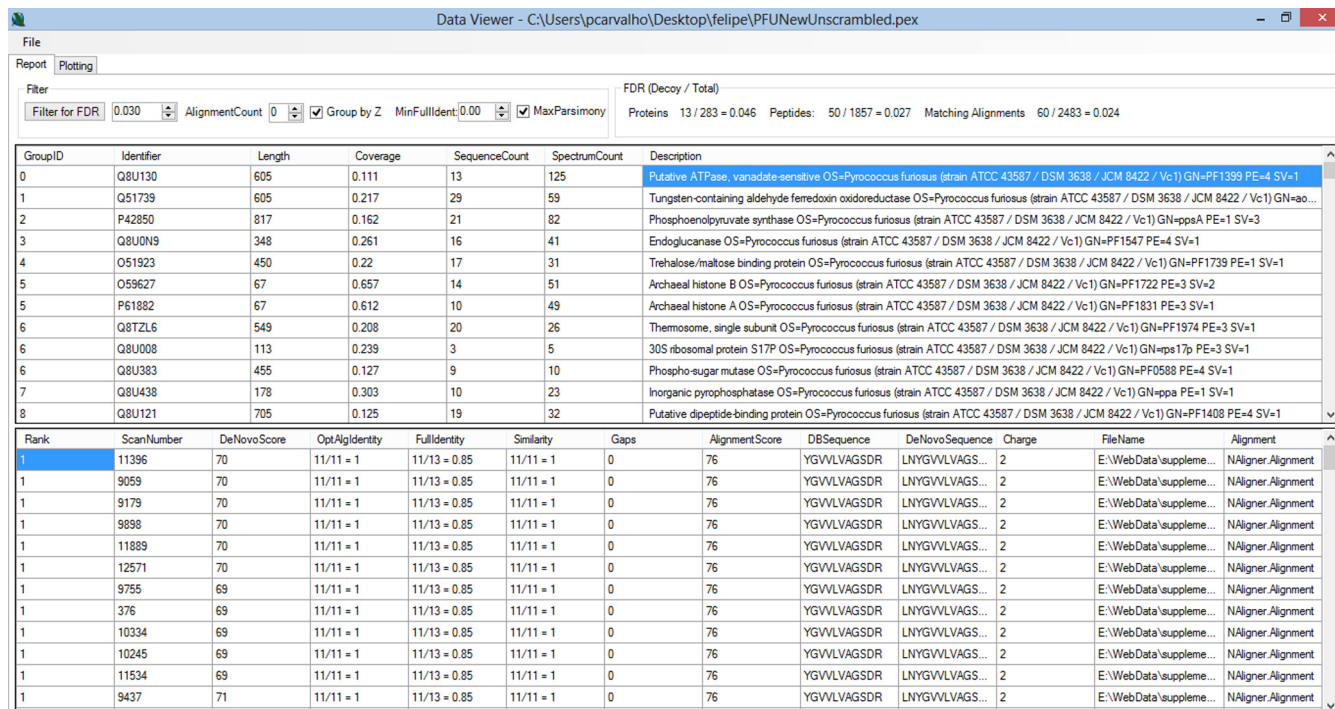


FIG. 3. Graphical user interface of the results browser. The results browser is composed of two panels. The upper panel displays information related to protein identification. When the user clicks on a protein of interest, further details on the peptides and their corresponding alignments are displayed in the lower panel.

We further manually examined the non-decoy proteins uniquely identified by PepExplorer; because we individually analyzed each case, based on spectral quality, alignment scores, and coverage, we feel comfortable in considering them as correctly identified, even though they were not found in our gold-standard search, ProLuCID.

The results of the application of these tools to the modified versions of the PFU dataset are discriminated in Table I.

All results are made available as part of the supplemental material or at the PepExplorer website.

Bothrops jararaca Plasma Proteomic Assessment—PepExplorer generated 3862 alignments (1% FDR), corresponding to 1333 peptides mapping to 199 proteins arranged into 86 protein groups. The ProLuCID/SEPro pipeline identified 349 spectra corresponding to 83 peptides mapping to 17 proteins arranged into 12 protein groups (0% FDR). All protein groups identified by ProLuCID were present in the PepExplorer results. Moreover, all but five proteins identified by ProLuCID had their identifiers contained in the PepExplorer results. These five remaining identifications shared peptides or had at least 80% identity with one protein provided by PepExplorer. The detailed lists of identifications, SEPro files, and PepExplorer files are provided in the supplemental material.

A 100% overlap between our similarity-driven approach and a PSM approach might not occur because of the convergence strategy adopted by PepExplorer, as it will opt for proteins having greater numbers of alignment mappings to

converge to a maximum-parsimony list. When we compared the average sequence coverages obtained for the same proteins identified by the PepExplorer and PSM approaches, we found an approximately 64% increase with the former approach (supplemental Table S1).

Recently, De Moraes-Zani and co-workers (44) analyzed the plasma composition of juvenile and adult *B. jararaca* snakes seeking ontogenetic variability. They used an experimental strategy consisting of two-dimensional electrophoresis separation followed by mass spectrometry analysis and protein identification by PSM, using MASCOT as the search engine. The authors were able to report eight plasma protein groups, with one of them possibly due to sample contamination during collection (β -actin). With the exception of transferrin, all plasma proteins reported in that study were also detected in our PSM approach (ProLuCID/SEPro); furthermore, we were also able to identify other proteins such as fibronectin 1, α -2-macroglobulin, apolipoprotein B100, fibrinogen β chain, and small serum protein (supplemental Table S1). One possible explanation for our extended list of PSM identifications might be our experimental approach (shotgun proteomics) as opposed to theirs (two-dimensional electrophoresis).

Finally, when we compared the PepExplorer results (for proteins displaying a sequence count greater than two) we were able to identify all the plasma protein families mentioned above and additional ones, namely, γ phospholipase inhibitor type IV, plasminogen, ceruplasmin, IgG Fc-binding protein-like, complement C4-B-like, inter- α -trypsin inhibitor heavy

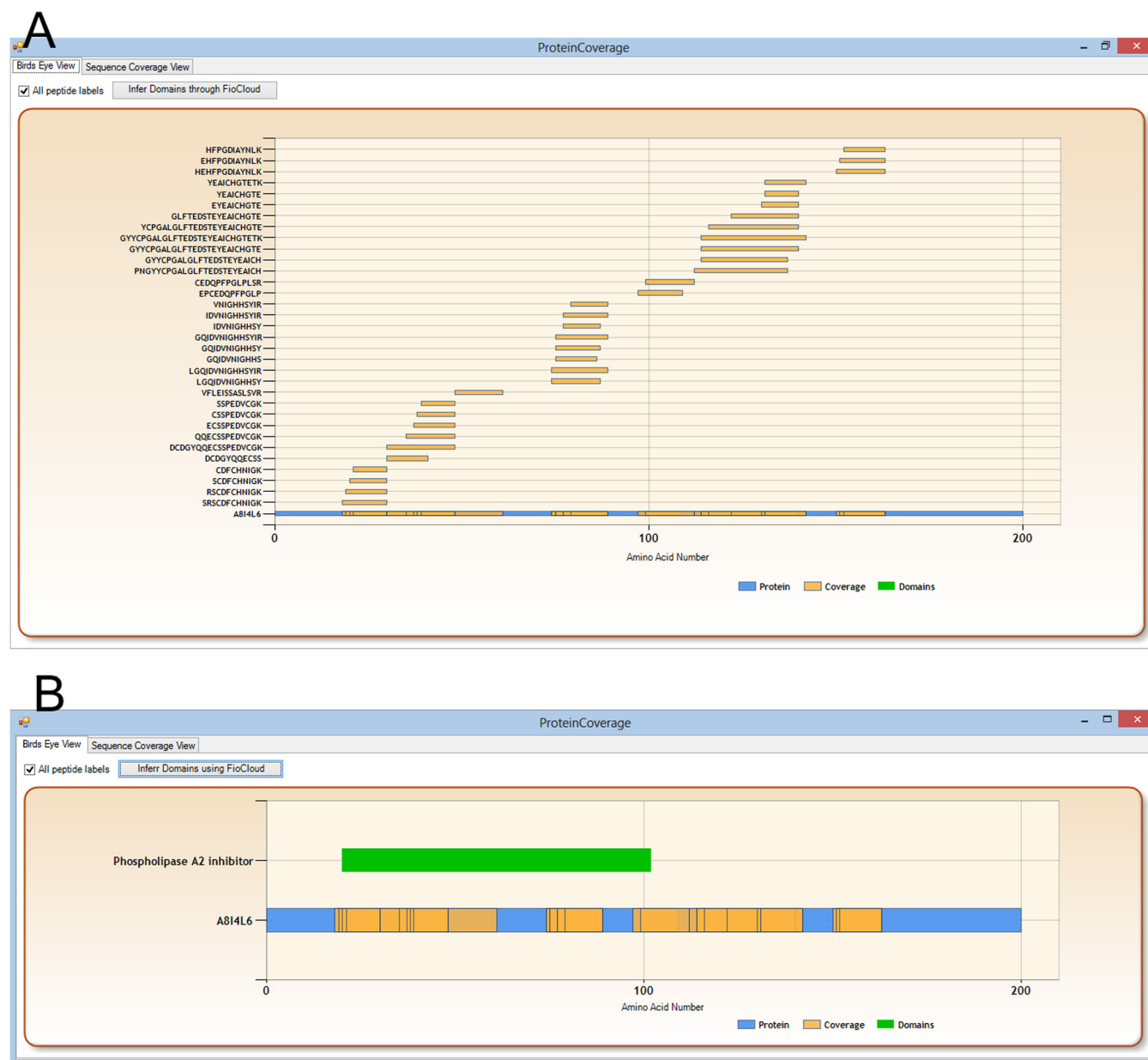


FIG. 4. Example of report provided by PepExplorer for each identified protein. A, the graphical report of the protein sequence coverage shows the extension of the area covered by predicted peptides. B, the result from the domain inferred by the cloud service running HMMER3 over Pfam-A on the fly is shown.

chains H4- and H3-like, Ig λ light chain variable region, calnexin, multiple EGF-like domains protein 6, collagen α -1(XXIV) chain, kininogen-1, anionic trypsin-like, fibrinogen α chain-like, Ig γ -1 chain C region, and heparin cofactor 2 (supplemental Table S1). Supplemental Fig. S1 exemplifies peptide identifications provided by PepExplorer that were missed in the PSM approach.

DISCUSSION

Error-tolerant, similarity-driven tools have as an ultimate goal the listing of “true” identifications. However, defining what it means for an identification to be true is far from trivial,

so in general one seeks to define trueness based on how the sequences in question differ according to some measure relating to the evolutionary distance between them. Among the first successful attempts at quantifying an evolutionary distance, we highlight the point accepted mutation (PAM) divergence, defined for two given sequences as the average number of accepted point mutations per 100 amino acids required in order to convert one sequence into the other without any insertions or deletions (45). The PAM matrices are substitution matrices that summarize an expected evolutionary change at the amino-acid level through log-odds substitution ratios. Theoretically, this approach is designed

to compare sequences that are within a known evolutionary PAM divergence in evolution. Conversely, it is common experience that PAM matrices are, in general, very effective in finding “true alignments” that reflect biological phenomena even though PAM divergences do not always correspond to true evolutionary distances. In the experiment at hand, we chose one of the so-called low-order PAM matrices (e.g. PAM30MS), which theoretically should favor “closer” sequences and therefore such true alignments. Future versions and tools should incorporate strategies for automatically selecting substitution matrices tailored for the problem at hand. This could ultimately help in determining a subset of sequences for maximizing the sensitivity of the algorithm. We argue that the current version of PepExplorer helps by showing which peptides (and ultimately proteins) can be taken into consideration confidently enough. However, selecting an adequate substitution matrix remains an issue for the user’s careful consideration.

The results provided herein can be used to compare three paradigms for performing spectral identification: PSM, an error-tolerant/blind post-translational modification approach, and a similarity-driven approach. The strategies are shown to

be complementary, each having advantages and disadvantages. For example, the PSM approach was found to be the most sensitive one on the PFU dataset. This happened because we were dealing with a model organism, and thus fully (and tightly) relying on the restrictions provided in the sequence database would yield the best sensitivity. However, its performance rapidly degraded as more distractions and modifications were inserted into the database. Although Mod-A did not outperform PSM on the original PFU database, it was able to retain significantly more identifications as more distractions were inserted in the database. Mod-A most likely did not outperform the PSM approach on the original PFU dataset because the latter takes into account many more possibilities, resulting in a larger search space and sacrificing sensitivity (13). However, it would not be surprising if Mod-A outperformed PSM with higher organisms, as it will tolerate amino acid substitutions and unanticipated post-translational modifications. Indeed, taking into account multiple post-translational modifications can also quickly degrade the performance of *de novo* tools, and for this reason Mod-A will always provide results that are complementary to those of PepExplorer. Finally, PepExplorer presented the least sensitive results on the original PFU dataset, as *de novo* approaches are known to be error prone. However, the alignment paradigm is able to effectively retain the results as distractions are included in the database.

Finally, we would like to point out some potential applications of PepExplorer. Our algorithm is used to pinpoint a subset of *de novo* results that are similar to the database at hand. Yet there can be several *de novo* results, having a very high *de novo* sequencing score, that are not included in the PepExplorer output. These results should be given special attention: what PepExplorer discards could actually turn out to be truly novel molecules, given the high confidence of the *de novo* results.

CONCLUDING REMARKS

PepExplorer is recommended for large-scale shotgun proteomic experiments, that is, those in which a considerable number of spectra are generated, as in the datasets presented. Its use is not recommended for analyzing small collections of spectra such as those obtained when analyzing a two-dimensional gel spot. In such cases MS-BLAST (28) should be used instead.

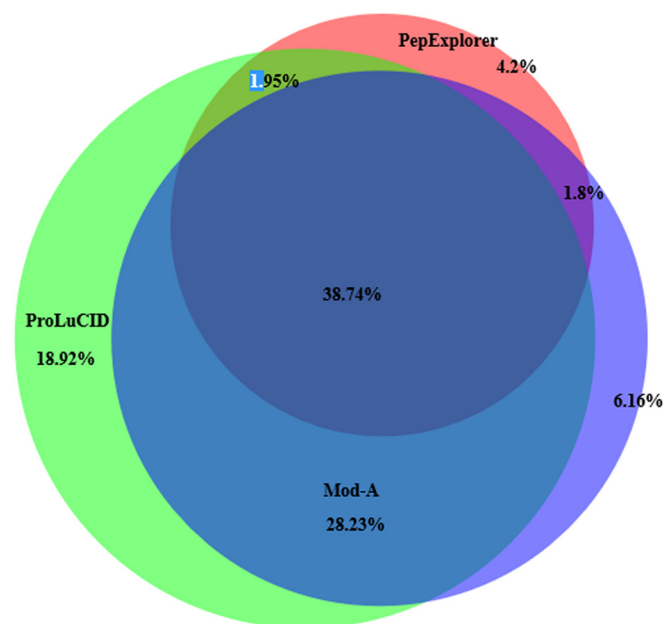


FIG. 5. A Venn diagram comparing the protein identification overlap of ProLuCID, Mod-A, and PepExplorer in the PFU dataset using the unmodified database.

TABLE I

Performance of ProLuCID, Mod-A, and PepExplorer on the four PFU datasets. The first number represents the number of proteins having at least two spectral counts identified under a 1% FDR. The number in parentheses is the average sequence coverage

	Number of proteins (Average Coverage)			
	PFU	PFU_Gap_25_Substitution_15	PFU_Gap_20_Substitution_10	PFU_Gap_15_Substitution_8
ProLuCID	585 (0.16)	45 (0.04)	7 (0.04)	0
Mod-A	499 (0.16)	63 (0.06)	45 (0.05)	0
PepExplorer	311 (0.17)	190 (0.17)	143 (0.17)	102 (0.17)

A key realization brought about by modern biotechnology has been that underneath the myriad unknown organisms lies great potential (46). Current strategies inspired by this realization include exploring extreme biomes for so-called extremophiles, a peculiar class of organisms that are generally responsible for the biosynthesis of molecular components useful for pharmaceutical or industrial applications. Perhaps one of the best examples has been the discovery of *Thermus aquaticus* and its heat-resistant polymerases, elected by *Science* in 1989 as the “molecules of the year” (47) and which have since aided in the development of biotechnology tools and ultimately facilitated the engineering of more effective drugs. The molecular characterization of venoms has also resulted in the engineering of new drugs (48). In conclusion, the literature is full of examples demonstrating the vast richness of biomolecular components and drug candidates that are naturally produced by different organisms already existing in our fauna and flora.

Recent advances in proteomic technologies are significantly impacting similarity-driven proteomics and, consequently, the exploration of novel organisms. Recently, Coon and coworkers benchmarked a new hybrid mass spectrometer, the Orbitrap Fusion (Thermo). The authors mention events in which they identified up to 19 sequences within less than a second, enabling them to achieve 90% coverage of the yeast proteome in one hour (49). Through this, the authors have raised the bar, in terms of the number of proteins identified per minute, to 70. High scanning rates coupled with ever-increasing resolving power are ingredients to boost the performance of *de novo* sequencing algorithms. As the general quality of predicted peptides is increasing, we foresee *de novo* sequencing playing a key role in the efficient handling of data from organisms with no available genomic information.

The field of genomics is also constantly going through significant advances. For example, next-generation sequencers are enabling the single-cell transcriptome (50) and personal genomics (51). Indeed, the coupling of “omics” sciences such as proteomics and metabolomics with next-generation sequencers will pave the way to true systems biology approaches, as these strategies are complementary to each other. The ever-growing amount of data on sequenced organisms, powered by next-generation sequencers, adds to similarity-driven approaches, as even more organisms will have their genomic information available. However, instrument time, expertise in data analysis, and financial resources are current bottlenecks for many groups.

Here we described a new methodology for dealing with *de novo* sequencing approaches, taking into account rigorous statistical criteria. We clearly demonstrated its efficiency in a controlled but real experiment with the PFU modified database and then presented the most comprehensive proteomic profile of *B. jararaca* plasma. Efforts such as the present work are necessary, as they expand the possibilities of what can be achieved in proteomics and in the study of organism biology.

In the near future we plan to automate the integration of data between different strategies like PSM and *de novo*, aiming at a wider perspective for mass-spectral analyses.

Availability of PepExplorer, the Raw Data, and Results—PepExplorer and supplementary files, including the *B. jararaca* raw data and all the results described in this work, are made freely available for academic purposes at our website. In order to view the full PSM results, installation of SEPro is required. PepExplorer is required for viewing results.

* We acknowledge CNPq, FAPERJ, CAPES (Grant No. 063/2010 - Edital Toxinologia), and PDTIS for financial support and use of core facilities. J.R.Y. acknowledges support from NIH Grant Nos. P41 GM103533 and R01 MH067880.

§ This article contains [supplemental material](#).

§§ To whom correspondence should be addressed.

REFERENCES

- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491
- Opie, L. H., and Kowolik, H. (1995) The discovery of captopril: from large animals to small molecules. *Cardiovasc. Res.* **30**, 18–25
- Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989
- Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinform.* **20**, 1466–1467
- Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567
- Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13**, 378–386
- Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214
- Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392
- Cociorva, D., Tabb, D., and Yates, J. R. (2007) Validation of tandem mass spectrometry database search results using DTASelect. *Curr. Protoc. Bioinform.* Chapter 13, Unit 13.4
- Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925
- Carvalho, P. C., Fischer, J. S. G., Xu, T., Cociorva, D., Balbuena, T. S., Valente, R. H., Perales, J., Yates, J. R., 3rd, and Barbosa, V. C. (2012) Search engine processor: filtering and organizing peptide spectrum matches. *Proteomics* **12**, 944–949
- Barboza, R., Cociorva, D., Xu, T., Barbosa, V. C., Perales, J., Valente, R. H., França, F. M. G., Yates, J. R., 3rd, and Carvalho, P. C. (2011) Can the false-discovery rate be misleading? *Proteomics* **11**, 4105–4108
- Borges, D., Perez-Riverol, Y., Nogueira, F. C. S., Domont, G. B., Noda, J., da Veiga Leprevost, F., Besada, V., França, F. M. G., Barbosa, V. C., Sánchez, A., and Carvalho, P. C. (2013) Effectively addressing complex proteomic search spaces with peptide spectrum matching. *Bioinform.* **29**, 1343–1344
- Biemann, K., Cone, C., Webster, B. R., and Arsenault, G. P. (1966) Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J. Am. Chem. Soc.* **88**, 5598–5606
- Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., and Zhang, X. (2006) Performance evaluation of existing *de novo* sequencing algorithms. *J. Proteome Res.* **5**, 3018–3028
- Frank, A., and Pevzner, P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973

17. Chi, H., Sun, R.-X., Yang, B., Song, C.-Q., Wang, L.-H., Liu, C., Fu, Y., Yuan, Z.-F., Wang, H.-P., He, S.-M., and Dong, M.-Q. (2010) pNovo: de novo peptide sequencing and identification using HCD spectra. *J. Proteome Res.* **9**, 2713–2724
18. Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., and Buhmann, J. M. (2005) NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.* **77**, 7265–7273
19. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342
20. Coon, J. J., Ueberheide, B., Syka, J. E. P., Dryhurst, D. D., Ausio, J., Shabanowitz, J., and Hunt, D. F. (2005) Protein identification using sequential ion/ion reactions and tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9463–9468
21. Zubarev, R. A., Horn, D. M., Fridriksson, E. K., Kelleher, N. L., Kruger, N. A., Lewis, M. A., Carpenter, B. K., and McLafferty, F. W. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* **72**, 563–573
22. Zubarev, R. A., Zubarev, A. R., and Savitski, M. M. (2008) Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet? *J. Am. Soc. Mass Spectrom.* **19**, 753–761
23. Bandeira, N. (2007) Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. *Biotechniques* **42**, 687, 689, 691 passim
24. Guthals, A., and Bandeira, N. (2012) Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Mol. Cell. Proteomics* **11**, 550–557
25. Guthals, A., Clauser, K. R., and Bandeira, N. (2012) Shotgun protein sequencing with meta-contig assembly. *Mol. Cell. Proteomics* **11**, 1084–1096
26. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399
27. Taylor, J. A., and Johnson, R. S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11**, 1067–1075
28. Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., and Standing, K. G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73**, 1917–1926
29. Junqueira, M., and Carvalho, P. C. (2012) Tools and challenges for diversity-driven proteomics in Brazil. *Proteomics* **12**, 2601–2606
30. Ma, B., and Johnson, R. (2012) De novo sequencing and homology searching. *Mol. Cell. Proteomics* **11**, O111.014902
31. Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–708
32. Vaudel, M., Burkhart, J. M., Breiter, D., Zahedi, R. P., Sickmann, A., and Martens, L. (2012) A complex standard for protein identification, designed by evolution. *J. Proteome Res.* **11**, 5065–5071
33. Yates, J. R., 3rd, Park, S. K. R., Delahunty, C. M., Xu, T., Savas, J. N., Cociorva, D., and Carvalho, P. C. (2012) Toward objective evaluation of proteomic algorithms. *Nat. Methods* **9**, 455–456
34. Na, S., Bandeira, N., and Paek, E. (2012) Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell. Proteomics* **11**, M111.010199
35. Estevão-Costa, M. I., Rocha, B. C., de Alvarenga Mudado, M., Redondo, R., Franco, G. R., and Fortes-Dias, C. L. (2008) Prospection, structural analysis and phylogenetic relationships of endogenous gamma-phospholipase A(2) inhibitors in Brazilian Bothrops snakes (Viperidae, Crotalinae). *Toxicon* **52**, 122–129
36. Tanaka-Azevedo, A. M., Tanaka, A. S., and Sano-Martins, I. S. (2003) A new blood coagulation inhibitor from the snake Bothrops jararaca plasma: isolation and characterization. *Biochem. Biophys. Res. Commun.* **308**, 706–712
37. Valente, R. H., Dragulev, B., Perales, J., Fox, J. W., and Domont, G. B. (2001) BJ46a, a snake venom metalloproteinase inhibitor. Isolation, characterization, cloning and insights into its mechanism of action. *Eur. J. Biochem. FEBS* **268**, 3042–3052
38. Smith, P. K., Krohn, R. I., Hermanson, G. T., Mallia, A. K., Gartner, F. H., Provenzano, M. D., Fujimoto, E. K., Goeke, N. M., Olson, B. J., and Klenk, D. C. (1985) Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150**, 76–85
39. McDonald, W. H., Tabb, D. L., Sadygov, R. G., MacCoss, M. J., Venable, J., Graumann, J., Johnson, J. R., Cociorva, D., and Yates, J. R., 3rd (2004) MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **18**, 2162–2168
40. Xu, T., Venable, J. D., Park, S., Cociorva, D., Lu, B., Liao, L., Wohlschlegel, J., Hewel, J., and Yates, J. R. (2006) ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol. Cell. Proteomics* **5**, S174
41. Arthur, D., and Vassilvitskii, S. (2007) in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, pp. 1027–1035, Society for Industrial and Applied Mathematics, Philadelphia, PA
42. Leprevost, F. V., Lima, D. B., Crestani, J., Perez-Riverol, Y., Zanchin, N., Barbosa, V. C., and Carvalho, P. C. (2013) Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for Proteomics. *J. Proteomics* **89**, 179–182
43. Eddy, S. R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform. Int. Conf. Genome Inform.* **23**, 205–211
44. De Moraes-Zani, K., Grego, K. F., Tanaka, A. S., and Tanaka-Azevedo, A. M. (2013) Proteomic analysis of the ontogenetic variability in plasma composition of juvenile and adult Bothrops jararaca snakes. *Int. J. Proteomics* **2013**, 135709
45. Dayhoff, M. O. (1979) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington, D.C.
46. Brock, T. D. (1997) The value of basic research: discovery of *Thermus aquaticus* and other extreme thermophiles. *Genetics* **146**, 1207–1210
47. Guyer, R. L., and Koshland, D. E., Jr. (1989) The Molecule of the Year. *Science* **246**, 1543–1546
48. Fox, J. W., and Serrano, S. M. T. (2007) Approaching the golden age of natural product pharmaceuticals from venom libraries: an overview of toxins and toxin-derivatives currently involved in therapeutic or diagnostic applications. *Curr. Pharm. Des.* **13**, 2927–2934
49. Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., and Coon, J. J. (2014) The one hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–347
50. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382
51. Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., Stamataki, J., and Akey, J. M. (2012) Personal and population genomics of human regulatory variation. *Genome Res.* **22**, 1689–1697

RESEARCH ARTICLE

Open Access

The influence of iron on the proteomic profile of *Chromobacterium violaceum*

Daniel C Lima¹, Fábio T Duarte², Viviane KS Medeiros³, Diogo B Lima⁴, Paulo C Carvalho⁴, Diego Bonatto⁵ and Silvia R Batistuzzo de Medeiros^{1*}

Abstract

Background: *Chromobacterium violaceum* is a bacterium commonly found in tropical and subtropical regions and is associated with important pharmacological and industrial attributes such as producing substances with therapeutic properties and synthesizing biodegradable polymers. Its genome was sequenced, however, approximately 40% of its genes still remain with unknown functions. Although *C. violaceum* is known by its versatile capacity of living in a wide range of environments, little is known on how it achieves such success. Here, we investigated the proteomic profile of *C. violaceum* cultivated in the absence and presence of high iron concentration, describing some proteins of unknown function that might play an important role in iron homeostasis, amongst others.

Results: Briefly, *C. violaceum* was cultivated in the absence and in the presence of 9 mM of iron during four hours. Total proteins were identified by LC-MS and through the PatternLab pipeline. Our proteomic analysis indicates major changes in the energetic metabolism, and alterations in the synthesis of key transport and stress proteins. In addition, it may suggest the presence of a yet unidentified operon that could be related to oxidative stress, together with a set of other proteins with unknown function. The protein-protein interaction network also pinpointed the importance of energetic metabolism proteins to the acclimatation of *C. violaceum* in high concentration of iron.

Conclusions: This is the first proteomic analysis of the opportunistic pathogen *C. violaceum* in the presence of high iron concentration. Our data allowed us to identify a yet undescribed operon that might have a role in oxidative stress defense. Our work provides new data that will contribute to understand how this bacterium achieve its capacity of surviving in harsh conditions as well as to open a way to explore the yet little availed biotechnological characteristics of this bacterium with the further exploring of the proteins of unknown function that we showed to be up-regulated in high iron concentration.

Keywords: Energetic metabolism, Coordinated adaptation, Sod

Background

Chromobacterium violaceum is a mobile bacillus, a facultatively aerobic and free-living organism that is associated with opportunistic infections in immunosuppressed individuals. Its ability to withstand various antibiotics poses this organism as responsible to a considerable number of deaths [1,2]. Research on pharmaceutical applications of *C. violaceum* have been carried out since the 1970's with reported production of antitumor peptides, compounds with

analgesic and antibiotic capability [3-5], and synthesis of biodegradable polymers [6]. *C. violaceum's* hallmark is the production of the purple pigment violacein that has innumerable therapeutic properties such as anti-tumoral activity [7].

Pharmacological and industrial attributes motivated the scientific community to sequence its approximately 4.7 Mbp genome [8]. To date, approximately 40% of its ORFs remain with unknown functions which make this organism a target for prospecting genes with biotechnological properties. Previous reports have described several key aspects of *C. violaceum*, such as having great metabolic versatility presenting several ORFs related to osmotic stress, response to heat, oxidative stress, DNA

* Correspondence: sbatistu@cb.ufrn.br

¹Laboratório de Biologia Molecular e Genômica, Centro de Biociências, Universidade Federal do Rio Grande do Norte (UFRN), Campus Universitário s/n, Lagoa Nova, Natal, RN 59072-970, Brazil

Full list of author information is available at the end of the article

repair, and a large number of proteins involved in iron metabolism [9,10].

In general, microorganisms need iron to survive; it is essential in key biological processes such as cellular respiration, electron transfer, gene regulation, DNA synthesis [11,12] and is known to play key roles in metabolism of various pathogenic microorganisms [13-16]. Despite its importance in catalysis of biological processes, an excess of iron can damage cells through the Fenton reaction in which ferrous ions react with hydrogen peroxide and produce hydroxyl radical [17]. Thus, high concentrations of this metal can lead to oxidative stress which causes damage to biomolecules and ultimately the death of the organism or cell [17].

The aim of this study was to compare the proteomic profiles of *C. violaceum* cultivated in absence and presence of iron. Taken together, a further comprehension on the metabolism of this metal in pathogenic organisms can aid in the discovery of therapeutic targets and further understanding in the biology of its adaptation mechanisms. Briefly, our results showed an increased synthesis of proteins mainly related with the tricarboxylic acid cycle (TCA), transportation, oxidative stress, and lists key proteins possibly linked to its adaptation in an iron rich medium. In addition, under the light of proteomic and RT-PCR data, we identified a new non-characterized *operon* possibly related to oxidative stress, composed with *sodB2* and two other genes of unknown function.

Methods

Culture conditions and treatment with FeSO₄

Isolated colonies of *Chromobacterium violaceum*, ATCC 12472 strain, were inoculated in liquid Luria Bertani (LB) medium for 16–18 hours at 28°C and shaken at 200 rpm.

The treatment was performed in an Erlenmeyer flask diluting the pre-cultivated bacteria with liquid LB medium (1:10) in a final volume of 100 mL. The FeSO₄ solution was prepared at an initial concentration of 500 mM and was further filtered with 0.22 µm GHP membrane disc filters [Acrodisc] and added at a final concentration of 9 mM. This concentration was chosen after a screening of treatment concentrations in which at only 9 mM of iron we observed different patterns of proteins synthesis at 1D-SDS-PAGE, in contrast to the negative control as well as previous results of our group that have demonstrated the high resistance of *C. violaceum* to iron (data not shown). The negative control consisted of *C. violaceum* grown in a LB broth. The treatment exposed the culture for four hours under the same conditions as described above. A measurement of *C. violaceum* growth was performed between 0–4 hours and the optical density was read at 600 nm in a 96-well spectrophotometer. The

negative control and the experimental conditions were performed in biological triplicate.

Total intracellular iron measurement

Total intracellular iron concentration was estimated as described at Barbehenn et al. [18]. First, *C. violaceum* cultures (100 mL) were grown until reaching 0.4–0.5 OD when the treatment began. After the four-hour treatment, the samples were centrifuged at 4°C, 2700 g for 15 min. Then, samples were washed with 50 mM EDTA, followed by another wash with ultrapure water. Finally, the phenanthroline assay was performed as described by Barbehenn et al. [18]. The negative control and the experimental condition were performed in biological triplicate.

Protein extraction

After four hours of treatment, the samples were centrifuged at 2880 g and 4°C for 20 minutes. The supernatant of all samples (including negative control) were discarded and then the samples were washed in 50 mM EDTA pH 8.0 and centrifuged under the same conditions previously mentioned. Then the samples were washed once more in 10 mM Tris-HCl, pH 8.5 solution and centrifuged at 2880 g, 4°C for 25 minutes. Cells were lysed in a 300 µL – 400 µL extraction buffer containing 7 M urea, thiourea 1 M, 50 mM DTT, 0.5% CHAPS, and 30 mM Tris pH 8.5. Proteins were precipitated by adding 1 mL of acetone, vortexing the samples and then centrifuging at 10,000 g, 4°C for 3 minutes. This process was performed a second time, and the sample was centrifuged again for 5 minutes. Finally, the proteins were solubilized in the same extraction buffer (300 µL–900 µL).

Antioxidant activity evaluation of *Chromobacterium violaceum*

In order to evaluate if the iron treatment promotes oxidative stress in *Chromobacterium violaceum*, antioxidant enzymes catalase (CAT) and superoxide dismutase (SOD) activities were measured on total protein extracts. Catalase Assay Kit 707002 and Superoxide Dismutase Assay Kit 706002 (Cayman Chemical, Ann Arbor, MI) were used according to manufacturer's recommendations to quantify the catalase and superoxide dismutase activity level, respectively. The total antioxidant activity was evaluated using the Antioxidant Assay kit from Sigma-Aldrich (CS0790) according to manufacturer's instructions. The protein concentration values were used to normalize the enzyme activity. The experiment was performed in biological triplicate.

SDS-PAGE

The total protein extracts were quantified by the Bradford method [19] and 20 µg of protein from each sample was resolved on a polyacrylamide gel under denaturing conditions

(SDS-PAGE) at 12%. The marker used was the Precision Plus Proteins™ WesternC™ Standard from Bio-Rad. The gel was stained using the Coomassie Colloidal from Sigma-Aldrich.

In gel digestion and peptides extraction

Following electrophoresis, each lane was excised in nine fragments according to the protein's density. All the proteins from each lane were excised and each group of proteins was analyzed independently by mass spectrometry (see below). Proteins were extracted and digested from these fragments as in accordance to the revised Shevchenko et al. protocol [20]. The dye and the SDS were quickly removed by washing the fragments three times in 50% acetonitrile solution (ACN) and 10 mM ammonium bicarbonate. Then the bands were dehydrated in ACN at 100%, reduced with 10 mM dithiothreitol (DTT) at room temperature and alkylated with 50 mM iodacetamide (IAA) in a dark environment. Then, the bands were washed again with 100 mM ammonium bicarbonate, dehydrated with 100% ACN, and rehydrated with 100 mM ammonium bicarbonate; this was done twice. The bands were then hydrated in a trypsin solution (Trypsin Gold, Mass Spectrometry Grade from Promega [v5280]) prepared according to manufacturer's instructions. 35–50 μ L of trypsin at 20 μ g/mL was added to the samples kept on ice. Then we added 50 mM of ammonium bicarbonate, sufficient to cover the bands during incubation at 37°C for 16–18 hours.

For peptide extraction, 10–30 μ L of 5% formic acid were added to the bands. After 10 minutes of incubation at room temperature, the supernatant containing peptides was transferred to another tube. Then a second extraction solution (5% formic acid and 50% ACN) was added in enough volume to cover the bands. The supernatant was transferred to a previously prepared tube containing the already extracted peptides. The extraction of the peptides performed with the second solution was repeated once again. Finally, the solution containing the digested peptides was concentrated in Eppendorf's *Concentrator Plus*.

Mass spectrometry data acquisition

After the *in gel* digestion, the samples were loaded onto the liquid chromatography NanoAcquity UPLC system (Waters) connected with an ESI-Q-ToF *premier* (Waters) mass spectrometer. The tryptic peptides from each sample (4.5 μ L) were separated on a BEH130-C18 column (100 μ m \times 100 mm) at a 600 nL/min flow rate. The gradient varied from 2 to 98% ACN in 0.1% formic acid for 45 minutes. The instrument's data acquisition mode was set to a data dependent "top three" and controlled using MassLynx v.4.1.

Protein identification

The ProLuCID search engine v 1.3 [21] was used to compare experimental spectra against those theoretically generated from *C. violaceum* ATCC 12472 sequence downloaded from Uniprot in January 2013, plus those from 127 common contaminants to proteomic experiments (e.g., Keratin, BSA, etc.). The search was limited to tryptic and semi-tryptic peptide candidates; carbamidomethylation and oxidation of methionine were imposed as fixed and variable modifications, respectively. The search engine accepted peptide candidates within a 50-ppm tolerance from the measured precursor m/z and used the XCorr and Z-Score as the primary and secondary search engine scores, respectively.

The validity of the Peptide Sequence Matches (PSMs) was assessed using the Search Engine Processor (SEPro) v.2.2.0.1 [22]. Briefly, identifications were grouped by charge state (+2 and \geq +3) and then by tryptic status (tryptic or semi-tryptic), resulting in four distinct subgroups. For each group, the XCorr, ZScore, DeltaCN, and DeltaMass values were used to generate a Bayesian discriminant function. The identifications were sorted in a non-decreasing order according to the discriminator score. A cutoff score was established to accept a false-discovery rate (FDR) of 1% based on the number of decoys. This procedure was independently performed on each data subset, resulting in a false-positive rate that was independent of tryptic status or charge state. Additionally, a minimum sequence length of six amino acid residues was required. Results were post processed to only accept PSMs with less than 10 ppm and proteins supported by two or more independent evidences (e.g., identification of a peptide with different charge states, a modified and a non-modified version of the same peptide, or two different peptides). This last filter led to a 0% FDR in all search results for all our sample analyses.

Differential proteomics and functional analysis

The PatternLab's updated ACFold module was employed to pinpoint differentially expressed proteins between the control and iron exposed condition [23,24]. The revised ACFold module presents increased sensitivity under the Benjamini-Hochberg q-value [25] bound by applying a variable fold-change that varies with the AC-test p-value as a power law [24].

Proteins uniquely identified in one condition (control or iron) were pinpointed according to PatternLab's Approximate Area Proportional Venn Diagram module. To better cope with the limitations from undersampling, we only considered proteins identified in two replicates of each condition, and not found in any replicates of the other condition.

The functional categorization of the proteins was assessed using PatternLab's Gene Ontology Explorer

(GOEx) module [26]. Our data analysis used the gene ontology database (OBO v1.2 - http://www.geneontology.org/GO.format.obo-1_2.shtml), downloaded February 16th, 2013 and the *C. violaceum* gene ontology annotation in the Uniprot text file format of its protein sequences.

Functional domain analysis

The protein sequences were compared [27] against the protein database from NCBI for automatic annotation and functional domains were predicted using the Conserved Domain Database (CDD) [28].

The structural prediction of hypothetical ORFs was performed using the server for protein homology detection HHpred (data not shown) [29]. FASTA sequences from each hypothetical protein were submitted to the server and the PDB database from February 23rd of 2013 was used. No specific organism proteome was selected, the method for multiple sequence alignment (MSA) generation was HHblits [30], up to 3 MSA generation iterations was used, the secondary structure score was applied, and the alignment mode was set to local.

Protein-protein interaction network analysis

The metasearch tool STITCH 3.1 (<http://stitch.embl.de>) was used to estimate protein-protein interaction networks related to iron response. STITCH is a tool used to explore known and predicted interactions between proteins, and chemical or physical agents. These agents interconnected by evidences derived from experiments, databases, and literature. One network was generated composed of proteins that were identified exclusively in the iron treatment or having an increased expression after being exposed to the metal. As the interaction targets were derived from experimental data, a high confidence index (0.700) was used to generate the networks. All prediction methods were activated: neighborhood, gene fusion, co-occurrence, co-expression, experiments, and databases text mining.

The assembled network was exported to be subsequently analyzed in Cytoscape 2.8.2 and Cytoscape 3.0 [31]. The former was used to calculate the centrality indexes and the later for the remaining analysis. The most relevant proteins sub-networks were selected using the Cytoscape MCODE v. 1.4.0 plugin (Molecular Complex Detection) [32]. The MCODE analysis parameters were: loops inclusion; degree cutoff of 2; haircut option enabled (which leads to deletion of nodes cluster with a single connection); fluff option enabled, node score cutoff at 0.2; K-core at 2, and maximum depth of 100. The only clusters used were those in which the MCODE index score was greater or equal to 2.5.

The network centrality calculations were computed from local networks and topologies. The network's bottlenecks were identified through a Betweenness vs. Node Degree chart with the values generated by the CestiScaPe 1.21

plugin installed at Cytoscape 2.8.2. Betweenness indicates the extent to which a particular node is among all other nodes in a network and usually shows the influence of this node on information propagation within the network [33,34]. Node Degree corresponds to the number of connections that a particular node has with other adjacent nodes. High node degree levels are called "hubs" [35]. Thus, a particular node with high node degree and betweenness values represents a bottleneck, or a protein that interconnects many biological processes [34].

Total RNA extraction from *C. violaceum* and cDNA synthesis

Isolated colonies of *C. violaceum* were cultured in absence and the presence (9 mM) of iron in the previously mentioned conditions. Further, 2 mL of the culture was used to extract and purify total RNA using the *RNAspin Mini Isolation* kit according to manufacturer's instructions (GE, catalog number 25-0500-72).

Once extracted, the total RNA was used to synthesize cDNA with random primers with the *High capacity cDNA reverse transcriptase* kit according to manufacturer's instructions (Applied Biosystems, catalog number 4368814).

RT-PCR analysis

The Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) was used to validate the expression of the hypothesized operon as a unique transcript. As such, RT-PCR was performed to provide the transcript evidence of regions encompassing the ORFs CV_0869 (the first) and CV_0867 (the last). Primers for the above mentioned genes were designed using Primer 3 (v. 0.4.0) software.

RT-PCR was performed using AmpliTaq Gold® 360 Master Mix (Catalog Number 4398881, Applied Biosystems). One microliter of cDNA was used and a final volume of 25 µL was used according to the manufacturer's instructions. Then, the PCR steps were the following: initial denaturation at 98°C for 5 min, 40 cycles of 98°C for 30 sec, 59.8°C for 30 sec, 72°C for 90 sec and a final extension at 72°C for 7 min were followed. Reactions using water and RNA instead of cDNA were carried out as negative controls. To verify the size of the amplicons, 5 µL of the PCR reaction was loaded on 1% agarose gel and submitted to electrophoresis. The DNA ladder of 1 Kb (Promega) was used as reference. An additional file (Additional file 1) shows the sequence of the primers used in this work.

Validation of expression by Real-Time quantitative PCR

The quantitative real-time PCR was used to validate the proteome analysis by verifying the expression of the genes CV_0868 and CV_0867, both comprising what we hypothesized as a newly described operon. Seven nanograms of cDNA (produced as mentioned above, from the iron-culture and control) and a final concentration

of 0.25 nM of each primer were applied in a 10 μ L reaction using Power SYBR[®] Green PCR Master Mix (Applied Biosystems, 4367659) in an One-Step cycling with the following conditions: 95°C for 3 min, 40 cycles of denaturation at 95°C for 3 sec, annealing/extension at 60°C for 15 sec. The 16S rRNA was used as endogenous control and to normalize the expression of the other two genes. The relative quantification was assessed by Δ Ct comparative analysis. The primers were designed using Primer 3 (v. 0.4.0). Primers sequences are described in Additional file 1. Statistical analysis was performed according to the t-test. Results were considered significant for $p < 0.05$. Two biological replicates were used.

Results

C. violaceum growth and intracellular iron estimation

The phenanthroline assay showed that the treatment is, as expected, leading to a significant increase in the total intracellular iron in *C. violaceum* (Figure 1) when compared with the bacterium grown in the absence of the metal. The OD measurement (Additional file 2) showed that the iron treatment leads to a growth arrest in *C. violaceum* although, from this assay, we cannot see a death tendency caused by the experimental condition, suggesting that this bacterium has a mechanism to withstand this elevated iron exposure.

C. violaceum antioxidant profile

We evaluated if the iron concentration used in the experiment can induce oxidative stress by assessing the catalase and superoxide dismutase enzymatic activities and the total antioxidant activity from the *C. violaceum* protein extract. The catalase activity (Figure 2A), the superoxide dismutase activity (Figure 2B), and the total antioxidant activity (Figure 2C) increased significantly during the treatment performed with 9 mM iron, which

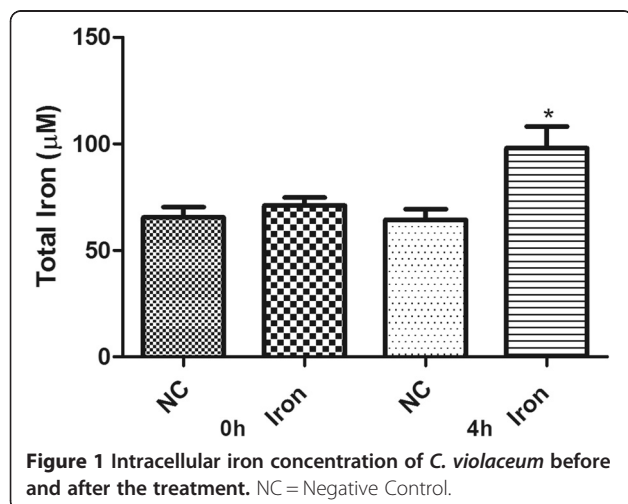


Figure 1 Intracellular iron concentration of *C. violaceum* before and after the treatment. NC = Negative Control.

suggests that the presence of metal is inducing an oxidative stress scenario in *C. violaceum*.

LC-MS/MS analysis

PatternLab's analysis allowed us to identify 230 proteins found in at least two biological replicates and in both biological conditions of which 28 were only identified in the control and 9 in the iron-responsive proteome. We note that identifying a protein exclusively in a condition does not exclude the hypothesis that it is present in the other condition; it can be below the detection sensitivity or not identified given the random sampling nature and undersampling of our data acquisition strategy. Nevertheless, for the experimental design at hand, identifying proteins on a single condition for two technical replicates suggests a differential expression [36].

The differentially expressed proteins found in the Control versus Iron treatment were accessed according to the updated ACFold module; 45 proteins were pointed as differentially expressed (Blue dots in Additional file 3) ($q < 0.05$). Table 1 lists the differentially expressed proteins while the Additional files 4 and 5 list the proteins identified only in the iron-responsive proteome or control condition, respectively.

The PatternLab's Gene Ontology Explorer module shows the most representative class of proteins according to the Gene Ontology classification. As can be noted in Figure 3, the majority of the proteins were mapped was the "Organic substance metabolic process" (11.2%) GO term which includes proteins belonging to many different biological processes. Other worth-mentioning categories are: transferase activity (5.9%), ion binding (5.9%), and oxidoreductase activity (4.9%). For the sake of completeness, we used more specific classifications retrieved from Uniprot server to describe each class of proteins in the tables.

Protein-protein interaction network analysis

We created interaction networks from the proteomic data using the STITCH tool to look for proteins candidates coordinating the response of *C. violaceum* in response to the high concentration of iron. For this, proteins identified exclusively in the iron-responsive proteome and those found to be up-regulated served as input for the interactomes tools.

Cytoscape's interactome analysis revealed a network comprised of 27 nodes and 44 edges (Figure 4). The MCODE plug-in identified two main clusters in the network. The first (Red nodes in Figure 4), corresponds to proteins belonging to energetic metabolism; the second is composed of ribosomal proteins (Blue nodes in Figure 4). Indeed, these were two classes of proteins significantly abundant in our data; a possible role of these proteins in response to iron in *C. violaceum* will be discussed further.

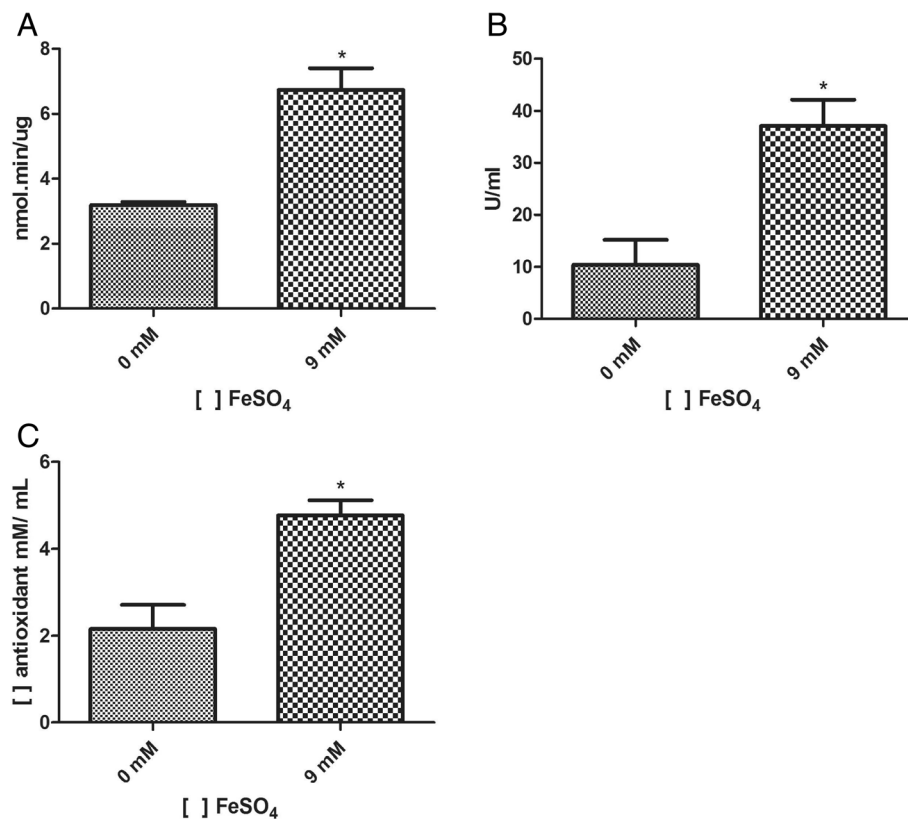


Figure 2 Antioxidant profile of *Chromobacterium violaceum*. **A)** Catalase activity assay. **B)** Superoxide dismutase activity assay. **C)** Total antioxidant activity assay of the total proteic extract of *Chromobacterium violaceum*.

Finally, the centrality indexes provided by the CentisScape allowed the identification of the central nodes controlling the communication between the biological processes (Additional file 6). The nodes with high betweenness were the Fumarate Hidratase (AspA) and α -oxoglutarate dehydrogenase (SucA). This indicates the importance of the energetic metabolism in the iron-responsive proteome of *C. violaceum*.

Proteins with unknown function

Possible roles for the proteins with unknown function from the iron-responsive proteome were investigated using BLAST and CDD (Conserved Domain Database). The identification of conserved domains may be the only clue to estimate the location and function of some proteins, as their domains have a high similarity with previously characterized proteins [28].

Of the five hypothetical ORFs that were up-regulated in the presence of iron, we were unable to identify any conserved domain in two of them (CV_1472 and CV_3099). The ORF CV_4300 contains the USP-like functional domain that is related to stress defense. According to the CDD, the synthesis of proteins in the USP family is increased when the bacteria is exposed to

stressful agents, increasing its survival rate during exposure to the agent.

CV_0868 has a domain of unknown function (DUF1842). This protein was previously detected in *C. violaceum* in another proteomic work from our group that studied this bacterium when submitted to hydrogen peroxide (personal communication). Thus, we inferred this protein to be related to a stress response. The HHpred analysis showed that the primary structure of CV_0868 to be associated to the tertiary structure of a Cu-Zn Superoxide Dismutase from different organisms (Additional file 7). Analyzing the genomic context of CV_0868, we observed that this ORF is adjacent to two other ORFs, one encoding a putative SodB (CV_0867) and the other encoding a protein of unknown function (CV_0869). Thus, we hypothesized *C. violaceum* possess a yet non-described operon related to oxidative stress.

To further investigate our assumption, we performed a RT-PCR of the possible new operon. The results showed the presence of an individual transcript comprising the three mentioned ORFs, leading us to suggest that this is, indeed, a new operon (Additional file 8). We further verified this hypothesis by Real Time PCR for the two ORFs comprising the operon (CV_0868 and CV_0869).

Table 1 Differential expressed proteins distributed in functional categories

Uniprot identifier	Fold change	Coverage	Description	Spectral count
Up-regulated				
Energetic metabolism				
Q7NZ60	2.01267122691591	0.3865	Malate dehydrogenase (Mdh)	76
Q7NY63	2.12401648386714	0.4413	Formate C-acetyltransferase (PflB)	139
Q7NZ52	2.52323856021444	0.1701	Citrate synthase (GtlA)	30
Q7NQM5	3.90695002871913	0.2484	Fumarate hydratase class II (AspA)	35
Q7NZ50	3.90695002871913	0.2054	Dihydrolipoamide succinyltransferase E2 component (SucB)	21
Proteins of unknown function				
Q7NZQ3	3.32090752441126	0.3605	Putative uncharacterized protein (CV_0868)	26
Q7NQ40	3.71160252728317	0.4051	Putative uncharacterized protein (CV_4300)	20
Oxidative metabolism				
Q7NWH0	15.6278001148765	0.2869	Probable aldehyde dehydrogenase	15
Q7NUH0	1.73292138370606	0.3369	Probable alcohol dehydrogenase	61
Stress response				
Q7P1C4	10.7441125789776	0.2365	Glutathione S-transferase family protein	10
Translation				
Q7NQG5	12.6975875933372	ND	30S ribosomal protein S14	ND
Q7NVZ4	1.65034958109687	0.4815	30S ribosomal protein S2	60
Transport				
Q7NZ25	2.4154454569446	0.5273	Probable binding protein component of ABC dipeptide transporter	212
Q7NQ13	3.6278821695249	0.1415	Probable oligopeptide ABC transporter system, substrate-binding protein	24
Q7NQN4	11.7208500861574	0.1116	Outer membrane protein W	11
Q7NSK0	2.05114876507754	0.6229	Porin signal peptide protein	347
Q7NXT7	4.23252919777905	0.2015	Probable amino acid ABC transporter	12
Others				
Q7NYA8	1.7992533026996	0.6141	Probable phasin	46
Q7NYB1	2.27905418341949	0.4851	Probable trans-acting regulatory HvrA protein	30
Q7NX40	28.3253877082137	0.2368	Protein kinase	30
Down-regulated				
Energetic metabolism				
Q7P0K7	-10.238165245516	0.241	Glyceraldehyde-3-phosphate dehydrogenase (GapA)	18
Q7NX09	-3.75399392335588	0.1019	Probable ribonuclease E	13
Stress response				
Q7NXP2	-9.21434872096442	0.3333	Thioredoxin	7
Q7NXI3	-2.18117433491428	0.2897	Chaperone protein DnaK	50
Q7NYF6	-13.3096148191708	0.1268	Chaperone protein HtpG	12
Q7NQ87	-13.3096148191708	0.3355	DNA-binding stress protein	15
Translation				
Q7NQH5	-9.21434872096442	0.2214	30S ribosomal protein S11	8
Q7NQH1	-9.21434872096442	0.3916	50S ribosomal protein L15	16
Q7NRL5	-4.35122022934431	0.4286	30S ribosomal protein S21	17
Q7NQF6	-2.27514783233689	0.3696	30S ribosomal protein S19	20
Q7NQE5	-2.13295109281584	0.374	50S ribosomal protein L7/L12	27

Table 1 Differential expressed proteins distributed in functional categories (Continued)

Q7NQF0	-1.87507970226867	0.4355	Elongation factor G	178
Q7NQH7	-1.82506423941807	0.3639	DNA-directed RNA polymerase subunit alpha	51
Q7NQE6	-1.62879901633209	0.182	DNA-directed RNA polymerase subunit beta	36
Q7NQF3	-1.59971331961188	0.6311	50S ribosomal protein L4	99
Q7NQF4	-13.3096148191708	0.3922	50S ribosomal protein L23	23
P60100	-12.2857982946192	0.2168	50S ribosomal protein L11	16
Q7M7F1	-1.04957291510636	0.6061	Elongation factor Tu	524
Q7NQG9	-2.0476330491032	0.4477	30S ribosomal protein S5	36
Q7NRPO	-10.238165245516	0.1005	Aspartate-tRNA ligase	9
Proteins of unknown function				
Q7NQ36	-11.2619817700676	0.3807	Putative uncharacterized protein (CV_4304)	10
Others				
Q7NXX7	-9.21434872096442	0.0967	N-succinylglutamate 5-semialdehyde dehydrogenase	8
Q7NXU5	-2.51300419662666	0.2443	Acetate kinase	26

All the proteins listed below has a q value <0.05.

The expression analysis (Additional file 9) of these two genes showed that after four hours of treatment, an up-regulation of these ORFs could be observed and therefore strongly suggests these proteins to be a response of *C. violaceum* to the iron treatment.

Discussion

Energetic metabolism

Energetic Metabolism describes one of the most expressive groups containing up-regulated or exclusively identified proteins in the presence of iron (Table 1 and Additional file 4). Many enzymes from this metabolic pathway have an Fe-S cluster in its active center, which helps to explain

the metabolic exchange to pathways in which enzymes do not use iron in their catalytic core. Supporting our results, Nwugo *et al.* [37] observed a boost in the expression of TCA cycle related proteins when exposing *Acinetobacter baumannii* to an iron rich medium.

Four proteins that were up-regulated after the iron treatment belong to the tricarboxylic acid cycle (SucB, AspA, GtA, and Mdh). Moreover, the protein Formate c-acetyltransferase (PflB) was up-regulated in the experimental condition. This protein is not known to be directly related to the tricarboxylic acid cycle but could be indirectly playing a role through an anaplerotic pathway. Figure 5 summarizes the energetic metabolism of

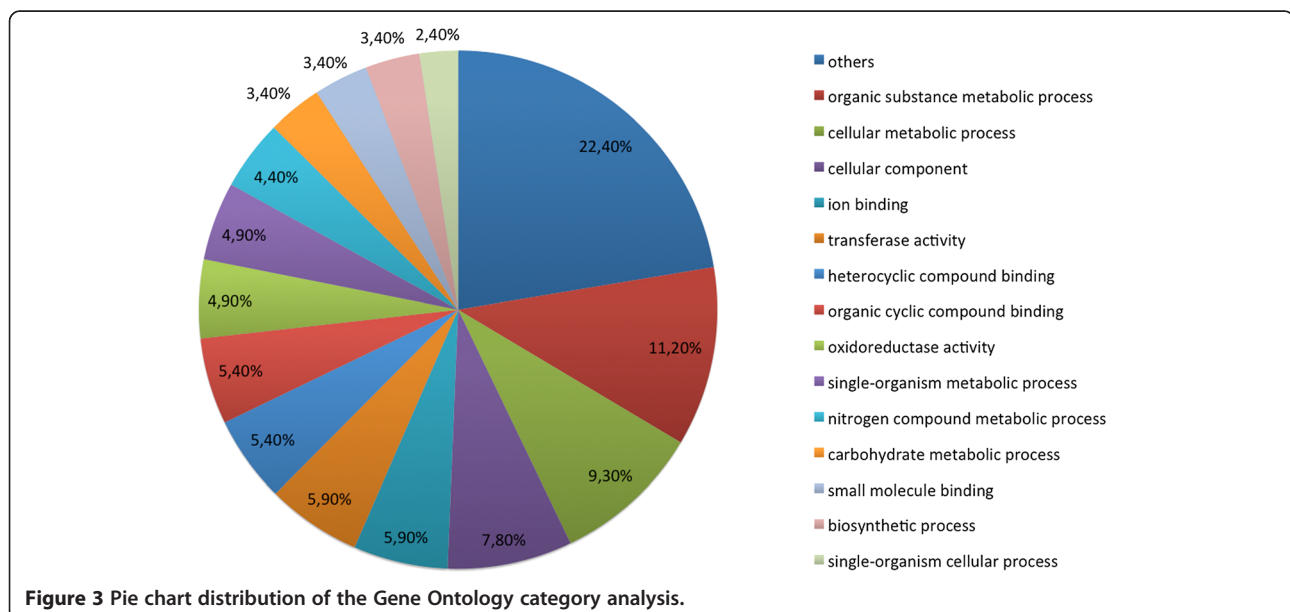


Figure 3 Pie chart distribution of the Gene Ontology category analysis.

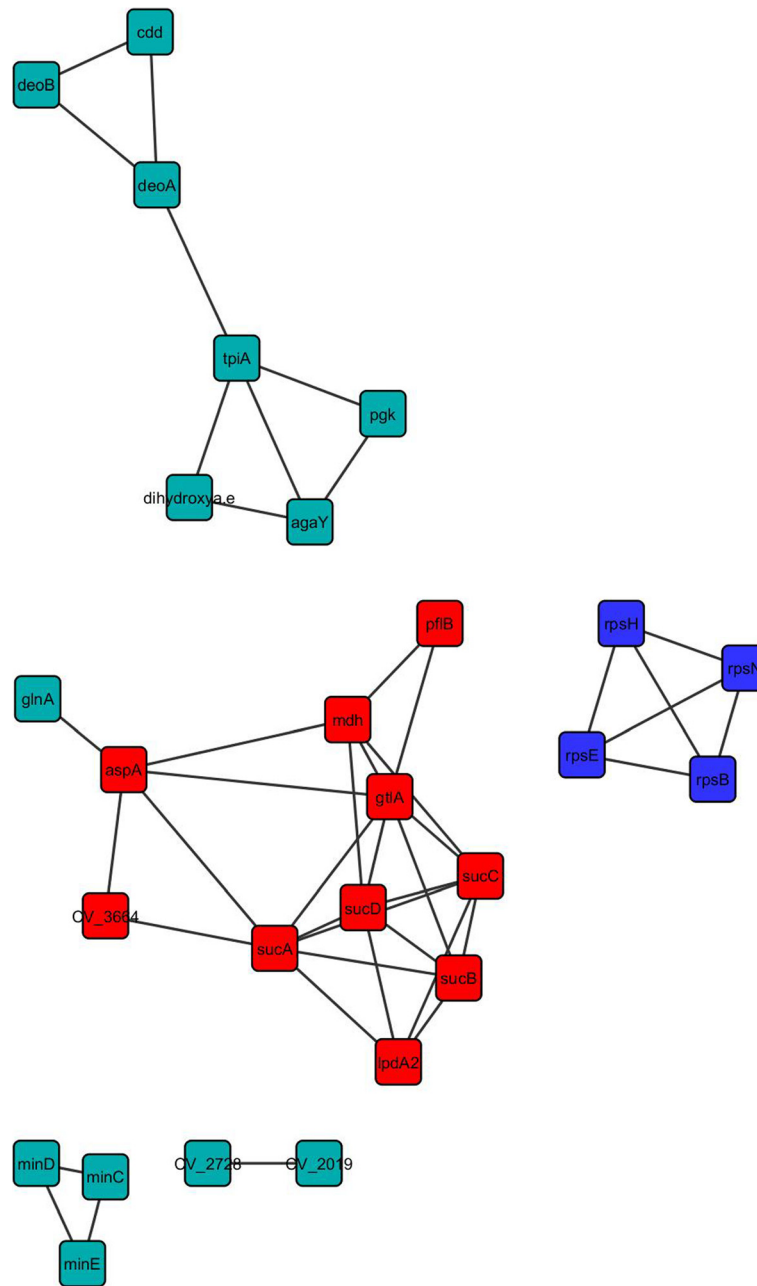


Figure 4 Protein interaction network of the iron-responsive proteome. The protein input was the data obtained from the proteomics analysis. Only the proteins that were up-regulated or exclusively identified in the iron responsive proteome were used in the construction of the protein-protein interaction network. The red and blue nodes represents the two clusters obtained with the MCODE plug-in, in which the red ones are the proteins related to energetic metabolism and the blue dots are ribosomal proteins.

C. violaceum in response to the high concentration of iron used in the culture.

Stress response

The relationship between iron and oxidative stress is well established. Although iron cannot directly damage biomolecules, it enables the Fenton Reaction, leading to the production of Reactive Oxygen Species (ROS), which

harms molecules such as DNA and proteins [38]. The total protein extract from *C. violaceum* was submitted to TAA assay and we observed an increase in the antioxidant activity in the protein extract from the iron treatment (Figure 2C). Additionally, the activity of the antioxidant enzymes Kat and Sod was also increased when the bacterium was cultured in the presence of the metal (Figure 2A and B). Although these assays showed the generation of

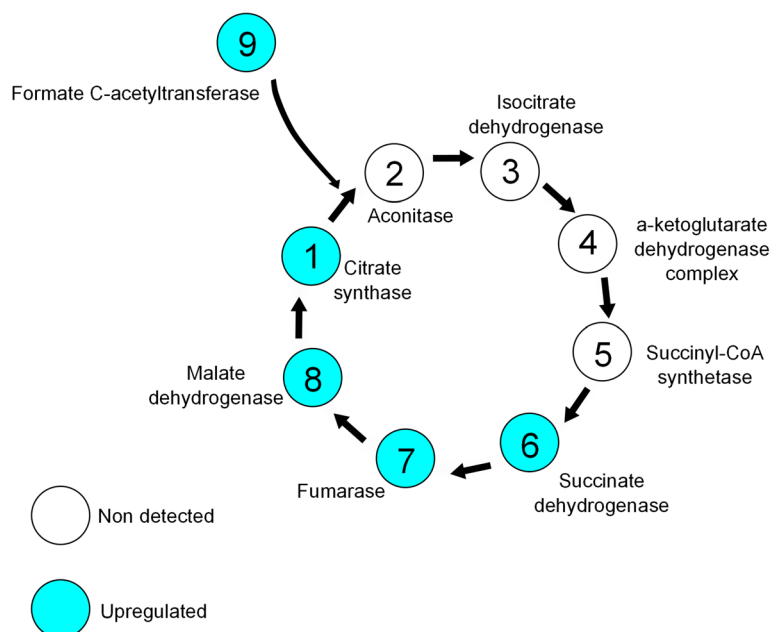


Figure 5 Integration of energetic metabolism in *C. violaceum*. The up-regulation of enzymes with iron in its catalytic center is leading to a boost of the TCA.

oxidative stress in the experimental condition, we were unable to detect these enzymes in our mass spectrometry analysis. In fact, we observed the down-regulation of many enzymes clearly related to antioxidant defense (DnaK, HtpG, Dps, and Thioredoxin) (Table 1). Nonetheless, the down-regulation of these enzymes is not necessarily correlated to a decrease in their activities. On the other side, two Glutathione S-transferase proteins were up-regulated in our iron-responsive proteome which suggests that the iron treatment induced an oxidative stress in *C. violaceum*.

Two other proteins that were up-regulated in the iron treatment were a Phasin and an outer membrane protein (OmpW). The former is a granule-associated protein that affects the synthesis and accumulation of Polyhydroxyalkanoates (PHAs) [39]. These polymers contribute to the redox balance and are known to be accumulated when the organism is subject to unfavorable conditions [40]. Thus, to compensate the down-regulation of the usual stress-related proteins, *C. violaceum* might be enhancing the expression of PHAs synthesis-related proteins as Phasin.

OmpW is a porin widely distributed among Gram-negative bacteria and is associated to stress resistance [41]. For example, Gil et al. [42] showed that OmpW becomes up-regulated when *Salmonella enterica* is submitted to Paraquat, a generator of superoxide, promoting the efflux of this ROS. The superoxide anion is yielded as a by-product of electron transporter chain and is a harmful reactive oxygen species. In this way, *C. violaceum* may be up-regulating the synthesis of OmpW to counteract

the harmful effects of superoxide, as the main antagonist enzyme of this component (Superoxide dismutase) was not detected in our work.

General metabolism

General metabolism proteins (replication, translation, cell cycle) were the proteins most abundantly detected in this study, especially between those that were down-regulated in the presence of iron. Indeed, in other proteomics studies, ribosomal proteins, proteins related to the biomolecules synthesis, such as tRNAs, have the greatest representations [43-45]. As one can note from Additional file 2, the iron is inducing a growth arrest in *C. violaceum*. The fact that many proteins from general metabolism were down-regulated in the iron responsive proteome could explain this growth halt.

Protein-protein interaction network analysis

The bottlenecks from the protein-protein interaction network are key nodes as they represent proteins that connect various functional clusters [35]. Yu and colleagues [35] provide an interesting discussion on importance of the bottleneck proteins in the maintenance of biological systems and that the deletion of some of these proteins may lead to a disruption of signal cascades, and ultimately to cell death. The protein-protein interaction study pinpointed two proteins with high degree of betweenness and node degree inside the iron-responsive network; the AspA and SucA. Both proteins are part of the energetic metabolism, more specifically the Tricarboxylic Acid Cycle,

suggesting that the great connectivity between these proteins with the whole iron-responsive network is part of the response of *C. violaceum* to the treatment. The importance of the energetic metabolism to adaptation of many microorganisms in the presence (and absence) of iron has been well established in previous works [14,46].

An oxidative stress related operon candidate

Our data suggests the identification of a new operon that encompasses the ORFs CV_0867, CV_0868, and CV_0869. The RT-PCR results indicated that all three genes are transcribed as a single transcript. We hypothesize that this operon could be related to oxidative stress mainly because our HHPred analysis indicated CV_0868 as being a Superoxide Dismutase (Sod). Moreover, CV_0868 was up-regulated in the iron-responsive proteome. Further functional characterization experimentation is required to confirm if the protein CV_0868 is a Superoxide Dismutase.

Conclusions

This is the first proteome study of *Chromobacterium violaceum* in response to a high concentration of iron. The analysis reveals the importance of energetic metabolism and stress proteins to the adaptation of the bacterium to the iron-repleted environment. Most importantly, we identified a new operon, encompassing the ORFs CV_0867, CV_0868 and CV_0869, that is probably related to oxidative stress response. Our data analysis supports the hypothetical protein CV_0868 as a Type C Superoxide Dismutase (SodC). Biochemical studies in our laboratory are being carried out to confirm the dismutation capacity of this protein. The other proteins with unknown function are good candidates to have their role determined in the involvement of iron homeostasis.

Data availability

All the .raw, .sqz and .sepr files from this work are available for download at <http://proteomics.fiocruz.br/dchaves/2014-1>.

Additional files

Additional file 1: List of all primers used in this work.

Additional file 2: Growth curve of *C. violaceum* cultivated in the absence and in the presence of 9 mM iron.

Additional file 3: ACFold analysis comparing the proteins expressed in the iron-responsive proteome and our control condition. Each protein is represented as a dot in the chart. Red dots are proteins that satisfy neither the variable fold-change cutoff nor the FDR cutoff $\alpha = 0.05$. Green dots are those that satisfy the fold-change cutoff but not α . Orange dots are those that satisfy both the fold-change cutoff and α but are lowly abundant proteins and therefore most likely have their quantitations compromised and can yield artificially low p-values. At the end, blue dots are those that satisfy all statistical filters. Dots in the upper part of the plot correspond to proteins over-expressed in the resection margin while the dots in the bottom are proteins down-regulated after exposition to iron.

Additional file 4: Proteins exclusively expressed in the iron responsive proteome. All the proteins were detected in at least two biological replicates.

Additional file 5: Proteins exclusively expressed in the control condition. All the proteins were detected in at least two biological replicates.

Additional file 6: Chart of the centrality analysis showing Betweenness \times Node Degree values. The proteins with higher betweenness and node degree are the ones predicted as the bottlenecks once they interconnect more pathways.

Additional file 7: Result of the HHPred analysis made with the protein sequence of CV_0868. As one can see, its probably structure matches with Type C Superoxide Dismutase.

Additional file 8: Electrophoresis in 1% Agarose of the RT-PCR amplicon. The result suggests the amplification of a single transcript, proving that the ORFs CV_0869, CV_0868, and CV_0867 encompass an operon. C1 – Negative Control (Nuclease Free Water); C2 – Negative Control (RNA that originates the cDNA was used); O – amplicon representing the operon (RT-PCR performed with cDNA).

Additional file 9: qPCR of the genes CV_0868 (A) and the ORF CV_0867 (B), a sod that is in operon with the former.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DCL contributed to the design, conduct and analyses of experiments, and the writing and preparation of the manuscript. FTD contributed to experiment design, data interpretation and manuscript revision. VKSM contributed to experiment design, data interpretation and manuscript revision. DBL participated in the proteomic data interpretation. PC participated in the proteomic data interpretation and manuscript elaboration. DB was involved the protein-protein network analysis. SRB contributed to the study conception and design, writing of the manuscript and overall supervision. All authors read and approved the final manuscript.

Acknowledgements

We thank the Laboratory of Mass Spectrometry of the Biosciences National Laboratory, CNPEM-ABTLuS, Campinas, SP, for the assistance in mass spectrometry data acquisition. We also thank Jana Dara Freires de Queiroz for her help with the antioxidant analysis. This work had financial support from CNPq, INEspaço, Fiocruz – PDTIS, FAPERJ, and CAPES.

Author details

¹Laboratório de Biologia Molecular e Genômica, Centro de Biociências, Universidade Federal do Rio Grande do Norte (UFRN), Campus Universitário s/n, Lagoa Nova, Natal, RN 59072-970, Brazil. ²Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, São Gonçalo do Amarante, RN, Brazil. ³Universidade Federal do Vale do São Francisco, Petrolina, PE, Brazil. ⁴Laboratório de Proteômica e Engenharia de Proteínas, Instituto Carlos Chagas, Fiocruz, PR, Brazil. ⁵Departamento de Biologia Molecular e Biotecnologia, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul (UFRGS), Avenida Bento Gonçalves, 9500-Prédio 43421, Caixa Postal 15005, 91509-900 Porto Alegre, RS, Brazil.

Received: 23 April 2014 Accepted: 10 October 2014

Published online: 20 October 2014

References

1. Yang C-H, Li Y-H: **Chromobacterium violaceum** infection: a clinical review of an important but neglected infection. *J Chin Med Assoc* 2011, **74**:435–441.
2. Vijayan AP, Anand MR, Remesh P: **Chromobacterium violaceum** Sepsis in an Infant. *Indian Pediatr* 2009, **46**:721–722.
3. Nakajima H, Kim YB, Terano H, Yoshida M, Horinouchi S: **FR901228, a potent antitumor antibiotic, is a novel histone deacetylase inhibitor.** *Exp Cell Res* 1998, **241**:126–133.
4. Durán N, Menck CF: **Chromobacterium violaceum: a review of pharmacological and industrial perspectives.** *Crit Rev Microbiol* 2001, **27**:201–222.

5. Durán M, Faljoni-Alario A, Durán N: **Chromobacterium violaceum and its important metabolites—review.** *Folia Microbiol (Praha)* 2010, **55**:535–547.
6. Kolibachuk D, Miller A, Dennis D: **Cloning, molecular analysis, and expression of the polyhydroxyalkanoic acid synthase (phaC) gene from Chromobacterium violaceum.** *Appl Environ Microbiol* 1999, **65**:3561–3565.
7. Durán N, Justo GZ, Ferreira CV, Melo PS, Cordi L, Martins D: **Violacein: Properties and Biological Activities.** *Biotechnol. Appl. Biochem* 2007, **48**:127–133.
8. Brazilian National Genome Project Consortium: **The complete genome sequence of Chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability.** *Proc Natl Acad Sci U S A* 2003, **100**:11660–11665.
9. Duarte FT, Carvalho FM, Bezerra e Silva U, Scortecci KC, Blaha CA, Agnez-Lima LF, Batistuzzo de Medeiros SR: **DNA repair in Chromobacterium violaceum.** *Genet Mol Res* 2004, **3**:167–180.
10. Hungria M, Gomes EA, Tereza A, De Vasconcelos R: **Tolerance to stress and environmental adaptability of Chromobacterium violaceum.** *Genet Mol Res* 2004, **3**:102–116.
11. Bou-Abdallah F: **The iron redox and hydrolysis chemistry of the ferritins.** *Biochim Biophys Acta* 1800, **2010**:719–731.
12. Haas H, Eisendle M, Turgeon BG: **Siderophores in fungal physiology and virulence.** *Annu Rev Phytopathol* 2008, **46**:149–187.
13. Doherty CP: **Host-pathogen interactions: the role of iron.** *J Nutr* 2007, **137**:1341–1344.
14. Miyamoto K, Kosakai K, Ikebayashi S, Tsuchiya T, Yamamoto S, Tsujibo H: **Proteomic analysis of Vibrio vulnificus M2799 grown under iron-repleted and iron-depleted conditions.** *Microb Pathog* 2009, **46**:171–177.
15. Mercier A, Labbé S: **Iron-dependent remodeling of fungal metabolic pathways associated with ferrichrome biosynthesis.** *Appl Environ Microbiol* 2010, **76**:3806–3817.
16. Crestani J, Carvalho PC, Han X, Seixas A, Broetto L, Fischer JDSG, Staats CC, Schrank A, Yates JR, Vainstein MH: **Proteomic profiling of the influence of iron availability on Cryptococcus gattii.** *J Proteome Res* 2012, **11**:189–205.
17. Touati D: **Iron and oxidative stress in bacteria.** *Arch Biochem Biophys* 2000, **373**:1–6.
18. Barbehenn R, Dodick T, Poopat U, Spencer B: **Fenton-type reactions and iron concentrations in the midgut fluids of tree-feeding caterpillars.** *Arch Insect Biochem Physiol* 2005, **60**:32–43.
19. Bradford MM: **A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.** *Anal Biochem* 1976, **72**:248–254.
20. Shevchenko A, Wilm M, Vorm O, Mann M: **Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels.** *Anal Chem* 1996, **68**:850–858.
21. Xu T, Venable JD, Park SK, Cociorva D, Lu B, Liao L, Wohlschlegel J, Hewel J, Yates JR III: **ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program.** *Mol Cell Proteomics* 2006, **5**:S174.
22. Carvalho PC, Fischer JSG, Xu T, Cociorva D, Balbuena TS, Valente RH, Perales J, Yates JR, Barbosa VC: **Search Engine Processor: filtering and organizing PSMs.** *Proteomics* 2012, **12**:944–949.
23. Carvalho PC, Fischer JSG, Chen EI, Yates JR, Barbosa VC: **PatternLab for proteomics: a tool for differential shotgun proteomics.** *BMC Bioinformatics* 2008, **9**:316.
24. Carvalho PC, Yates JR, Barbosa VC: **Improving the TFold test for differential shotgun proteomics.** *Bioinformatics* 2012, **28**:1652–1654.
25. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57**:289–300.
26. Carvalho PC, Fischer JSG, Chen EI, Domont GB, Carvalho MGC, Degraeve WM, Yates JR, Barbosa VC: **GO Explorer: a gene-ontology tool to aid in the interpretation of shotgun proteomics data.** *Proteome Sci* 2009, **7**:6.
27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
28. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH: **CDD: a Conserved Domain Database for the functional annotation of proteins.** *Nucleic Acids Res* 2011, **39**(Database issue):D225–D229.
29. Söding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W244–W248.
30. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.** *Nat Methods* 2012, **9**:173–175.
31. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 2011, **27**:431–432.
32. Bader G, Hogue C: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4**:2.
33. Newman MEJ: **A measure of betweenness centrality based on random walks.** *Soc Networks* 2005, **27**:39–54.
34. Feltes BC, Faria Poloni J, Bonatto D: **The developmental aging and origins of health and disease hypotheses explained by different protein networks.** *Biogerontology* 2011, **12**:293–308.
35. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M: **The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.** *PLoS Comput Biol* 2007, **3**:e59.
36. Carvalho PC, Fischer JSG, Perales J, Yates JR, Barbosa VC, Bareinboim E: **Analyzing marginal cases in differential shotgun proteomics.** *Bioinformatics* 2011, **27**:275–276.
37. Nwugo CC, Gaddy JA, Zimmler DL, Actis LA: **Deciphering the iron response in Acinetobacter baumannii: a proteomics approach.** *J Proteomics* 2011, **74**:44–58.
38. Wall SB, Oh J-Y, Diers AR, Landar A: **Oxidative modification of proteins: an emerging mechanism of cell signaling.** *Front Physiol* 2012, **3**(September):369.
39. Dawes EA, Senior PJ: **The role and regulation of energy reserve polymers in micro-organisms.** *Adv Microb Physiol* 1973, **10**:135–266.
40. Ruiz JA, Fernández RO, Nikel PI, Méndez BS, Pettinari MJ: **Dye (arc) mutants: insights into an unexplained phenotype and its suppression by the synthesis of poly (3-hydroxybutyrate) in Escherichia coli recombinants.** *FEMS Microbiol Lett* 2006, **258**:55–60.
41. Nandi B, Nandy RK, Sarkar A, Ghose AC: **Structural features, properties and regulation of the outer-membrane protein W (OmpW) of Vibrio cholerae.** *Microbiology* 2005, **151**:2975–2986.
42. Gil F, Ipinza F, Fuentes J, Fumeron R, Villarreal JM, Aspée A, Mora GC, Vásquez CC, Saavedra C: **The ompW (porin) gene mediates methyl viologen (paraquat) efflux in Salmonella enterica serovar Typhimurium.** *Res Microbiol* 2007, **158**:529–536.
43. Anderson DC, Campbell EL, Meeks JC: **A soluble 3D LC/MS/MS proteome of the filamentous cyanobacterium Nostoc punctiforme.** *J Proteome Res* 2006, **5**:3096–3104.
44. Sixt BS, Heinz C, Pichler P, Heinz E, Montanaro J, Op den Camp HJM, Ammerer G, Mechtler K, Wagner M, Horn M: **Proteomic analysis reveals a virtually complete set of proteins for translation and energy generation in elementary bodies of the amoeba symbiont Protochlamydia amoebophila.** *Proteomics* 2011, **11**:1868–1892.
45. Gomes DF, da Silva Batista JS, Torres AR, de Souza AD, Galli-Terasawa LV, Hungria M: **Two-dimensional proteome reference map of Rhizobium tropici PRF 81 reveals several symbiotic determinants and strong resemblance with agrobacteria.** *Proteomics* 2012, **12**:859–863.
46. Friedman DB, Stauff DL, Pishchany G, Whitwell CW, Torres VJ, Skaar EP: **Staphylococcus aureus redirects central metabolism to increase iron availability.** *PLoS Pathog* 2006, **2**:e87.

doi:10.1186/s12866-014-0267-6

Cite this article as: Lima et al.: The influence of iron on the proteomic profile of Chromobacterium violaceum. *BMC Microbiology* 2014 **14**:267.

Comparative Proteomic Analysis of the Aging Soleus and Extensor Digitorum Longus Rat Muscles Using TMT Labeling and Mass Spectrometry

Daniela F. S. Chaves,^{*,†} Paulo C. Carvalho,[‡] Diogo B. Lima,[‡] Humberto Nicastro,[†] Fábio M. Lorenzetti,[†] Mário Siqueira-Filho,[§] Sandro M. Hirabara,^{§,#} Paulo H. M. Alves,^{||} James J. Moresco,[⊥] John R. Yates, III,^{*,⊥} and Antonio H. Lancha, Jr.[†]

[†]Laboratory of Applied Nutrition and Metabolism, School of Physical Education and Sports, University of São Paulo, Av. Prof. Mello Moraes, 65, 05508-900 São Paulo, SP, Brazil

[‡]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, 81350-010 Fiocruz, Paraná, Brazil

[§]Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil

^{||}Department of Anatomy, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil

[⊥]Department of Chemical Physiology, The Scripps Research Institute, San Diego, California 92121, United States

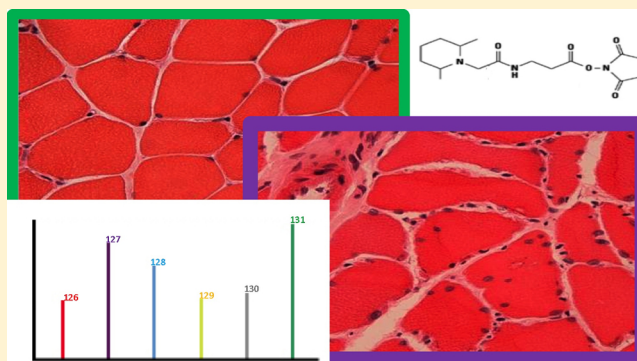
[#]Institute of Physical Activity Sciences and Sports, Cruzeiro do Sul University, Sao Paulo, SP, Brazil

S Supporting Information

ABSTRACT: Sarcopenia describes an age-related decline in skeletal muscle mass, strength, and function that ultimately impairs metabolism and leads to poor balance, frequent falling, limited mobility, and a reduction in quality of life. Here we investigate the pathogenesis of sarcopenia through a proteomic shotgun approach. In brief, we employed tandem mass tags to quantitate and compare the protein profiles obtained from young versus old rat slow-twitch type of muscle (soleus) and a fast-twitch type of muscle (extensor digitorum longus, EDL). Our results disclose 3452 and 1848 proteins identified from soleus and EDL muscles samples, of which 78 and 174 were found to be differentially expressed, respectively. In general, most of the proteins were structural related and involved in energy metabolism, oxidative stress, detoxification, or transport.

Aging affected soleus and EDL muscles differently, and several proteins were regulated in opposite ways. For example, pyruvate kinase had its expression and activity different in both soleus and EDL muscles. We were able to verify with existing literature many of our differentially expressed proteins as candidate aging biomarkers and, most importantly, disclose several new candidate biomarkers such as the glioblastoma amplified sequence, zero β -globin, and prolargin.

KEYWORDS: aging, EDL, proteome, rat, sarcopenia, soleus, TMT



INTRODUCTION

Sarcopenia describes an age-related decline in skeletal muscle mass, strength, and function^{1–4} that ultimately results in impaired metabolism, poor balance, frequent falling, limited mobility, and reduction in quality of life.^{5,6} Therefore, the identification of new disease markers will aid in understanding the biology of this disease and therefore pave the way to more effective diagnostic methods and treatments to stop or at least delay age-related muscle wasting.

Sarcopenia is a multifactorial disease that is highly correlated with aging. Alterations in cells have been found to occur such as increased oxidative stress, increased apoptosis, metabolic abnormalities, alterations in hormone response, decreased protein synthesis of myofibrillar components, and impaired

signal transduction.^{3,7–10} Although contractile proteins are lost during atrophy, alterations can occur in noncontractile proteins, increasing the complexity of such alterations.

Aging affects the metabolic capacity of skeletal muscle and changes the activity of specific enzymes. In general, there is a decline in the oxidative capacity of skeletal muscles of aged rats as well as a decline in energy-rich molecules such as ATP and creatine phosphate (CrP).^{11,12} However, the effect on enzyme activity and expression cannot be generalized because data suggest that distinct muscles may respond differently to aging.

Received: June 30, 2013

Published: September 3, 2013

During aging there is a progressive loss of skeletal muscle proteins. Fractional synthesis rates of myofibrillar proteins are reduced in the elderly, indicating that the old muscles have reduced capacity to synthesize new proteins. Furthermore, decline in mitochondrial proteins has also been reported (reviewed by Carmeli et al.¹³).

Oxidative stress appears to be a key component of sarcopenia, and it has been suggested that the deleterious effects of reactive oxygen species (ROS) are at least partially responsible for the aging process.¹⁴ Free radicals are also responsible for dysfunctional mitochondria, which play a major role in muscle function decline. Alterations in mitochondrial volume, increased oxidative stress, reduced oxidative capacity, and an increase in mitochondrially mediated apoptosis have been described (reviewed by Peterson et al.¹⁵).

Thirty-month old wistar rats represent an established animal model of skeletal muscle aging, and a variety of studies have been performed using this animal strain.^{16–18} Over the past decade, several muscle proteomic studies, following such biological model, have been completed using different techniques such as two-dimensional gel electrophoresis (2DE)^{19–21} and shotgun-based proteomics^{18,22–24} to catalogue muscle skeletal proteins. Within the 2DE universe, we note that several staining methods have been evaluated, such as Coomassie blue,²² fluorescent Deep Purple labeling,²³ and fluorescent difference in-gel electrophoresis (DIGE).²⁴ Taken together, these results describe various differentially expressed proteins related to structure or metabolic enzymes or involved in ion homeostasis and stress response.²⁵ Previous reports have also investigated post-translational modifications such as phosphorylation,^{26,27} glycosylation,^{28–30} carbonylation,³¹ and nitrosylation.¹⁷

Because soleus and extensor digitorum longus (EDL) muscles have different fiber composition, function, and metabolic features, we argue that an in-depth proteomic analysis of these tissues could unveil complementary aspects in the biology of aging. Skeletal muscle is composed of two major types of muscle fibers, the slow- and the fast-twitch fibers, which can be distinguished by the enzymatic characteristics of myosin ATPase and species of myosin heavy chain isoforms. Slow-twitch fibers express type I, while type II predominates in the fast-twitch fibers.³² Among skeletal muscles in the rat hind limb, the soleus muscle is a typical muscle composed mainly of slow-twitch fibers (>90% of type I fibers), which allow them prolonged, steady contractions. The EDL muscle is composed mainly of fast-twitch fibers (>90% of type II fibers), which ensures rapid and delicate movements.³³

Therefore, the aim of our study was to evaluate the effect of aging in the protein expression patterns of soleus and EDL muscles. To accomplish this, we employed state-of-the-art proteomic techniques such as multidimensional protein identification technology (MudPIT),³⁴ high-resolution mass spectrometry with an Orbitrap Velos (Thermo, San José) and an isobaric labeling quantitation approach using the tandem mass tags (TMTs).³⁵

■ EXPERIMENTAL PROCEDURES

Animals

This research was approved by the Local Ethical Committee, and the experiments were conducted in accordance with the National Research Council's Guidelines for the Care and Use of Laboratory Animals. Male Wistar rats were housed under

controlled environmental conditions (temperature, 22 °C; 12 h dark period starting at 6:00 p.m.). Young rats had an average weight of 0.370 kg ± 0.01 g (3 months of age), and old rats had an average weight of 0.500 kg ± 0.08 g (24 months of age).

Serum and Muscle Oxidative Stress

Serum oxidative stress was determined using ferrous oxidation–xylenol orange (FOX) assay, as previously described.³⁶

Creatine Kinase and Pyruvate Kinase Enzyme Activity

Samples were homogenized in extraction buffer (0.05 M Imidazole-HCl, 0.12 M KCl, and 0.062 M MgSO₄; pH 7.6) at 1:10 dilution (w/v). Homogenates were centrifuged at 7000g for 10 min at 4 °C, and the supernatants were used for the further assays. An aliquot of each homogenate was used for determining total protein content using BSA a standard (Bradford et al., 1979), which was used for normalization of the results. Evaluation of CK and PK activities was performed by the methodologies described by Ainsworth and MacFarlane³⁷ and Dinovo et al.,³⁸ respectively. Results are shown as μmol/min/mg protein.

Histological Analysis

Muscles were embedded in tissue tek, cooled in isopentane, frozen in liquid nitrogen, and sectioned with a cryostat. The resulting 10 μm transverse sections were stained with hematoxylin and eosin (HE) for morphological analysis and NADH-TR. Fibers from soleus muscle were classified as slow-oxidative (SO) and fast-oxidative-glycolytic (FOG); fibers from EDL muscle were classified as SO, FOG, and fast-glycolytic (FG). Cross-sectional areas (CSAs) of ~400 muscle fibers of a muscle from each rat were measured using the Image Pro-Plus software.

Trypsin Digestion

One hundred micrograms of protein were precipitated using the methanol/chloroform precipitation method. In brief, 4 volumes of methanol and 1 volume of chloroform was added, and the samples were mixed, centrifuged, and then washed once with 4 volumes of methanol. The samples were then dissolved in 45 μL of TEAB 200 mM pH 8.0 and 5 μL of 2% SDS, and ultrapure water was added to complete 100 μL. Reduction was performed with 10 μL TCEP 200 mM for 60 min at 55 °C, followed by 5 μL of 375 mM iodoacetamide for 30 min in the dark at RT. The sample was precipitated in cold acetone (1:6) at –20 °C overnight and dissolved in 45 μL of TEAB 200 mM pH 8.0, 2.5 μL of SDS 2%, and ultrapure water. Trypsin digestion was carried out overnight at 37 °C with 2.5 μg of enzyme.

TMT Labeling

Six-plex TMT labeling (Thermo Scientific) was performed according to the manufacturer's instructions to isobarically label primary amino groups and thus allow us to simultaneously compare six samples. In brief, each tube (0.8 mg) was reconstituted with 41 μL of anhydrous acetonitrile at RT, and the reagents were dissolved by vortexing for 5 min. Each sample was labeled by adding 10 μL of tag, followed by incubation for 1 h at RT. The reaction was quenched with the addition of 5% hydroxylamine for 15 min of incubation at RT. Samples were then pooled and stored at –80 °C until further analysis. Young muscle samples were labeled with 126, 127, and 128, while old muscle samples were labeled with 129, 130, and 131.

Table 1. Animal Characteristics

	rat	body weight (g)	soleus weight (g)	EDL weight (g)	SI (soleus) ^a	SI (EDL) ^b
young	1	0.37	0.23	0.17	0.62	0.45
	2	0.36	0.20	0.17	0.55	0.47
	3	0.37	0.20	0.17	0.54	0.45
	4	0.39	0.21	0.18	0.53	0.45
	5	0.38	0.20	0.16	0.52	0.42
	6	0.33	0.17	0.16	0.51	0.48
	7	0.37	0.18	0.16	0.51	0.45
	8	0.36	0.18	0.15	0.51	0.42
	9	0.36	0.18	0.19	0.50	0.53
MEAN ± SEM		0.369 ± 0.00574	0.196 ± 0.00599	0.168 ± 0.00409	0.532 ± 0.0117	0.458 ± 0.0117
old	1	0.52	0.16	0.19	0.31	0.37
	2	0.41	0.13	0.10	0.33	0.26
	3	0.55	0.25	0.23	0.46	0.43
	4	0.52	0.22	0.15	0.43	0.30
	5	0.54	0.23	0.31	0.44	0.57
	6	0.61	0.29	0.19	0.48	0.32
	7	0.35	0.20	0.16	0.42	0.46
	8	0.49	0.18	0.18	0.38	0.38
	9	0.51	0.21	0.19	0.41	0.37
MEAN ± SEM		0.502 ± 0.0250	0.210 ± 0.0158	0.193 ± 0.0187	0.404 ± 0.0189	0.384 ± 0.0313

^aSI = sarcopenia index = soleus weight/body weight. ^bSI = EDL weight/body weight.

Multidimensional Protein Identification Technology Analysis (MudPIT)

Analysis was performed on an LTQ Orbitrap Velos (Thermo Scientific, San Jose, CA) interfaced with a quaternary HP 1100 series HPLC pump (Agilent Technology, Santa Clara, CA). The analytical column was a 100 μ m diameter fused-silica capillary (J/W Scientific, Agilent Technology), pulled with a P-2000 laser (Sutter Instrument, Novato, CA), and packed with 12 cm of 5 μ m C18 resin (Aqua, Phenomenex, Torrance, CA). The biphasic microcapillary trapping column (5 cm of 250 μ m diameter) consisted of a fritted capillary with Kasil 1624, packed with 2.5 cm reversed-phase C18 (Aqua, Phenomenex) and 2.5 cm of strong cation exchange (5 μ m Partisphere, Whatman, Maidstone, Kent, U.K.) packing material. The biphasic column was loaded offline with sample using a pressure pump at \sim 800 psi.

An automated 11-step chromatographic run was performed on each sample using three mobile phases consisting of buffer A (5% acetonitrile (ACN); 0.1% formic acid (FA)), buffer B (80% ACN, 0.1% FA), and buffer C (500 mM ammonium acetate, 5% ACN, 0.1% FA). Step one consisted of a linear gradient from 0 to 100% buffer B (120 min). Steps 2–9 had the following profile: 1 min of 100% buffer A; 4 min of (100 – X)% buffer A, X% buffer C; 80 min of 100% buffer A to 50% buffer A and 50% buffer B; 10 min of 50% to 100% buffer B; 1 min of 100% buffer B to 100% A; 10 min of 100% buffer A. For buffer C, X% was, respectively, 10, 20, 30, 40, 50, 60, 70, and 100%. Steps 10 and 11 had the following profile: 2 min of 100% buffer A; 4 min of 10% buffer B, 90% buffer C; 45 min of 100% buffer A to 50% buffer A and 50% buffer B; 10 min of 50% to 100% buffer B; 1 min of 100% buffer B to 100% A; and 10 min of 100% buffer A.

The application of the distal voltage of 2.5 kV electrospayed the eluted peptides directly into LTQ Orbitrap Velos (Thermo Scientific). A cycle of one full scan was applied (300–2000 m/z , resolution 30 000), followed by 10 data-dependent HCD dual MS/MS and repeated continuously through each MudPIT step. Full scans and higher energy collisional activated

dissociation (HCD) scans were in Orbitrap, at resolutions of 30 000 and 7500, respectively. The normalized collision energy in the HCD was of 45% in HCD. The following parameters for dynamic exclusion were applied: 1 repeat count, 30 s repeat duration, 180 exclusion list size, 60 s exclusion duration, and a dynamic exclusion list of 60 s and an HCD fragmentation energy of 45. The number of microscans for ms1 and ms2 was 1 and a 2 m/z isolation window. The mass spectrometer and HPLC were controlled by the Xcalibur data system (Thermo, San Jose, CA).

Database Searching

Tandem mass spectra were extracted from the raw files using RawExtract 1.9.3.³⁹ (The RAW data, .sqt files, and SEPro files are available via Internet at <http://proteomics.fiocruz.br/daniela/jpr2013-1>.) Database searching for protein identification was performed using the ProLuCID algorithm v.1.3.1,⁴⁰ with the following parameters: carbamidomethylation of cysteines (57.021464 C) and TMT-labeled in the primary amino group (229.1629 K) as static modifications; trypsin as enzyme (KR) allowing for semispecific peptides; tolerance of 50 ppm; and the ProLuCID isotopic peaks parameter was set to 3.⁴¹ The search was performed against all *Rattus* sequences available for download from Uniprot on April 1, 2013; these comprised 41 772 sequences plus another 127 sequences that we added because they are known to be common mass spectrometry contaminants. For each sequence, a reversed decoy sequence was included, doubling the number of sequences in the final database.

Assessment of PSMs

The validity of the Peptide Sequence Matches (PSMs) was assessed using the Search Engine Processor (SEPro).⁴² In brief, identifications were grouped by charge state (+2 and +3) and then by tryptic status (fully tryptic, semitryptic), resulting in four distinct subgroups. For each group, the ProLuCID XCorr, DeltaCN, δ ppm, number of peaks matched, and ZScore values were used to generate a Bayesian discriminator. The identifications were sorted in a nondecreasing order according

to the discriminator score. A cutoff score was established to accept a false-discovery rate (FDR) of 1% based on the number of decoys. This procedure was independently performed on each data subset, resulting in a false-positive rate that was independent of tryptic status or charge state. Additionally, a minimum sequence length of six amino acid residues was required. Results were postprocessed to only accept PSMs with <8 ppm.

Quantitation and Pinpointing Differentially Expressed Proteins

The quantitation of confident identifications was assessed using SEPro's quantitation module SEProQ41, which is another module from PatternLab for proteomics suite. In brief, this module reads the intensity corresponding to each of the six TMT reporter ions (i.e., 126, 127, 128, 129, 130, and 131), of which the first three and the last three correspond to samples from different biological conditions. The data treatment options selected for this analysis were signal normalization by channel sum and impurity correction. For the latter, we included the impurity and isotopic overlap information provided together with the TMT kit. SEProQ uses a peptide centric approach to pinpoint differentially expressed proteins. For this, it relies on the paired Student's *t* test to pinpoint differentially expressed proteins by considering, for each mass spectrum associated with a given protein, the differences in average of the reporter ions associated with each biological state. Finally, after associating a differential expression *p* value to each protein, SEProQ applies the Benjamini–Hochberg⁴³ to control the theoretical FDR to $q < 0.05$.

RESULTS

Animal Characteristics

The body weight, soleus muscle wet weight, EDL muscle wet weight, and the muscle wet weight to body ratio (sarcopenia index (SI)) are shown in Table 1. The mean body weight for the young animals was $0.369 \text{ kg} \pm 0.00574$, and for the old group it was $0.502 \text{ kg} \pm 0.0250$ ($p < 0.0001$). Despite differences in body weight, soleus wet weight ($0.196 \text{ g} \pm 0.00599$ and $0.210 \text{ g} \pm 0.0158$; $p = 0.415$ for young and old, respectively) and EDL wet weight ($0.168 \text{ g} \pm 0.00409$ and $0.193 \text{ g} \pm 0.0187$; $p = 0.2243$ for young and old, respectively) were not statistically different between the groups.

The SI is defined as the relation between body weight and muscle weight²⁵ and was calculated for both soleus and EDL muscles. For young and old soleus, the SIs were 0.532 ± 0.0117 and 0.4044 ± 0.0189 , respectively, with $p < 0.0001$. For EDL, the young and old SIs were 0.458 ± 0.00117 and 0.384 ± 0.031 , respectively, with $p < 0.04$ (Figure 1). A drop in the SI, as observed for both muscles, but more significantly in soleus, indicates muscle wasting.

Histological Analysis

In soleus muscles, both SO and FOG fiber areas were significantly reduced in the old group when compared with the young group (Supplementary Figure 1A in the Supporting Information; 1418 ± 334.7 vs $115.2 \pm 7.4 \text{ mm}^2$ of SO fibers and 1719 ± 310.2 vs $136.0 \pm 2.7 \text{ mm}^2$ of FOG fiber in young and old groups, respectively; $p = 0.0065$ and 0.0013). Furthermore, soleus CSA was significantly decreased in the old group when compared with the young group (Supplementary Figure 1C in the Supporting Information; 3137 ± 640.2 vs $251.2 \pm 9.0 \text{ mm}^2$ in young and old groups,

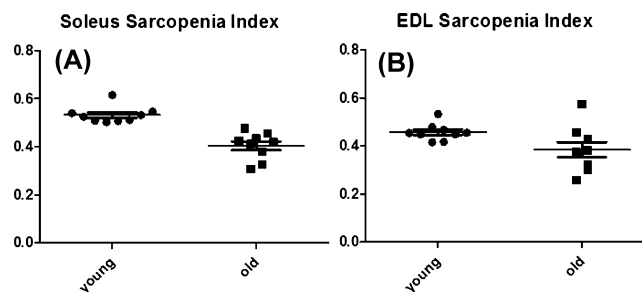


Figure 1. Relation between body weight and soleus and EDL muscles is shown as the sarcopenia index (SI, muscle wet weight over whole body weight). In aged animals, there is a loss of muscle weight, which results in a drop in the sarcopenia index. (A) young soleus SI (0.5323 ± 0.01169) versus old (0.4044 ± 0.01892) $p < 0.0001$; (B) young EDL SI (0.4576 ± 0.001169) versus old (0.3845 ± 0.03129) $p < 0.04$.

respectively; $p = 0.0028$). Additionally, soleus muscles presented an increased number of centrally located nuclei, extensive variability in diameter, and increased connective tissue (Supplementary Figure 2B in the Supporting Information).

In contrast with soleus, EDL muscles presented no significant differences in fiber areas (Supplementary Figure 1B in the Supporting Information; 890.4 ± 150.9 vs $642.2 \pm 185.3 \text{ mm}^2$ of SO fibers, 1595 ± 281.1 vs $1015 \pm 303.3 \text{ mm}^2$ of FOG fibers, and 2605 ± 533.9 vs $1489 \pm 458.3 \text{ mm}^2$ of FG fibers in young and old groups, respectively; $p > 0.05$) and in CSA area (Supplementary Figure 1C in the Supporting Information; 5091 ± 959.9 vs $3146 \pm 932.7 \text{ mm}^2$ in young and old groups, respectively; $p > 0.05$). Likewise, EDL muscles presented an increased number of centrally located nuclei, extensive variability in diameter, and increased connective tissue (Supplementary Figure 2A in the Supporting Information).

Differentially Expressed Proteins

Our soleus muscle shotgun proteomic analysis identified 3452 proteins (FDR at protein level = 0.98%) and a total of 6756 peptides (FDR at peptide level = 0.33%). PatternLab for proteomics^{41,44} provides an additional report indicating that these 3452 proteins can be reduced to 2027 when applying the bipartite maximum parsimony algorithm to converge to a maximum parsimony protein list⁴⁵ (i.e., a reduced list of proteins that explains all identified peptides). We note that 1006 of these proteins had at least one unique peptide. In our EDL muscle analysis, we identified 1848 proteins (FDR = 0.97%, 1075 with maximum parsimony of which 523 had at least one unique peptide) and a total of 4255 peptides (FDR at peptide level = 0.19%). To determine which proteins were differentially regulated, we used a fold change cutoff of 1.3 and a *q* value of 0.05 (viz., for the experiment at hand, a corrected *p* value cutoff of 0.03). Overall, 78 proteins had a change in abundance in soleus muscles, of which 57 were down-regulated in old muscles and 21 up-regulated (considering maximum parsimony) (Table 2). Accordingly, for EDL muscles, the number of up- and down-regulated proteins in the old muscles were 145 and 29, respectively, totaling 174 (Table 3).

Most of the differentially expressed proteins are involved in metabolism (glycolysis, oxidative metabolism, and cellular processes), contractile apparatus, cellular stress response, and detoxification.

Energy Metabolism

Most of the differentially expressed proteins in soleus muscles were down-regulated in the old tissue (Table 2). Glycolytic

Table 2. Differentially Expressed Proteins in Soleus Muscles

protein accession no.	p value	$-\text{Log}_2$ fold change ^a	description
Energy Metabolism			
Q6IMX3	0.017	-0.8	acetyl-coenzyme A dehydrogenase, short chain, isoform CRA_a
P13221	<0.001	-0.8	aspartate aminotransferase, cytoplasmic
P13221	0.001	-1.1	aspartate aminotransferase, mitochondrial
F1LP05	<0.001	-0.9	ATP synthase subunit α
P15429	<0.001	-0.6	β -enolase
Q68FY0	0.003	-0.7	cytochrome <i>b-c1</i> complex subunit 1, mitochondrial
P32551	<0.001	-0.7	cytochrome <i>b-c1</i> complex subunit 2, mitochondrial
P10888	<0.001	-1.3	cytochrome <i>c</i> oxidase subunit 4 isoform 1, mitochondrial
Q68FU3	0.005	-1.0	electron transfer flavoprotein subunit β
Q66HF3	0.019	-0.4	electron transfer flavoprotein-ubiquinone oxidoreductase, mitochondrial
P42123	<0.001	0.6	L-lactate dehydrogenase B chain
G3 V6H5	0.001	0.4	mitochondrial 2-oxoglutarate/malate carrier protein
P04636	<0.001	-1.2	mitochondrial 2-oxoglutarate/malate carrier protein
P16617	<0.001	-0.9	phosphoglycerate kinase 1
P25113	<0.001	-1.4	phosphoglycerate mutase 1
F1LPA6	0.001	3.7	protein-arginine deiminase type-2 (fragment)
Q6P7S0	0.001	-0.9	pyruvate kinase
P15651	0.017	-0.8	short-chain specific acyl-CoA dehydrogenase, mitochondrial
Q5RK08	0.010	-22.1	glioblastoma amplified sequence
Q64428	0.005	-0.7	trifunctional enzyme subunit α , mitochondrial
Structure and Cell Motility			
F1LYK7	0.001	0.4	protein Cfl2 (cofilin 2)
P85972	0.002	-0.5	vinculin
F1LMC6	0.016	17.8	troponin I, slow skeletal muscle (fragment)
G3 V885	<0.001	1.0	myosin-6 (Myh6)
G3 V8B0	<0.001	1.0	myosin-7 (Myh7)
Q9EQP5	0.018	1.0	prolargin
F1LRV9	0.006	0.9	myosin-4 (Myh4)
F1M789	<0.001	0.7	protein Myh13 (fragment)
G3 V6E1	0.002	0.6	myosin-4 (Myh4)
Q62920	0.015	-0.4	PDZ and LIM domain protein 5
F1LNH3	0.001	-0.5	protein Col6a2 (fragment)
D3ZCV0	<0.001	-0.7	protein Actn2
D3ZHA0	<0.001	-0.8	protein Flnc
F2Z3T2	0.003	-0.9	tropomyosin α -3 chain
P08733	<0.001	-0.9	myosin regulatory light chain 2, ventricular/cardiac muscle isoform
P68136	0.004	-0.9	actin, α skeletal muscle
Q63781	0.002	-1.0	myosin regulatory light chain
Q8K551	0.003	-1.1	truncated α -actinin
P04466	0.009	-1.5	myosin regulatory light chain 2, skeletal muscle isoform
P04692	<0.001	-2.0	tropomyosin α -1 chain
P58775	<0.001	-18.2	tropomyosin β chain
P09495	<0.001	-18.4	tropomyosin α -4 chain
Oxidative Stress, Detoxification and Degradation			
Q2TA66	0.018	2.6	ferritin (fragment) GN=Fth1 PE=2 SV=1
F1LVC6	0.002	1.2	glutathione <i>S</i> -transferase (fragment)
B6DYQ2	0.011	1.1	glutathione <i>S</i> -transferase mu 2
P63018	<0.001	-0.4	heat shock cognate 71 kDa protein
Q66HD0	0.003	-1.5	endoplasmic
P97541	<0.001	-2.8	heat shock protein β -6
P35565	0.010	-0.5	calnexin
D4ACB8	0.013	-0.7	chaperonin subunit 8 (theta) (predicted), isoform CRA_a
P11598	0.003	-0.7	protein disulfide-isomerase A3 GN=Pdia3 PE=1 SV=2
G3V8T4	0.011	-1.1	DNA damage-binding protein 1 GN=Ddb1 PE=4 SV=1
P0CC09	0.018	-1.2	histone H2A type 2-A
Transport and Catabolism			
P02770	<0.001	-1.0	serum albumin
Q63011	0.006	-16.9	zero β -globin (Fragment)
Q6MG90	0.013	-0.4	complement component 4, gene 2
P11442	0.004	-0.4	clathrin heavy chain 1

Table 2. continued

protein accession no.	p value	$-\text{Log}_2$ fold change ^a	description
Transport and Catabolism			
Q4V8H8	0.014	-0.9	EH domain-containing protein 2
Additional Differentially Expressed Proteins			
Q5M7V3	0.013	1.8	LOC367586 protein
P08932	0.002	1.4	T-kininogen 2
F1LPQ6	<0.001	0.9	uncharacterized protein (Fragment)
B2RZB2	0.005	0.7	putative uncharacterized protein
Q63041	<0.001	0.6	α -1-macroglobulin
B0BNM1	0.013	0.6	NAD(P)H-hydrate epimerase
A0JPJ7	0.001	0.4	Obg-like ATPase 1
O08557	0.002	-0.4	N(G),N(G)-dimethylarginine dimethylaminohydrolase 1
Q8R490	0.015	-0.5	cadherin 13
D4ABR6	0.002	-0.5	annexin
P14046	0.017	-0.9	α -1-inhibitor 3
Q9WTT7	0.015	-1.1	basic leucine zipper and W2 domain-containing protein 2
D3ZHA7	0.001	-1.2	protein LOC100359980
D3ZCI9	<0.001	-1.5	uncharacterized protein
D3ZH98	<0.001	-21.1	uncharacterized protein
P05197	<0.001	-0.9	elongation factor 2
P11762	<0.001	-1.1	galectin-1 GN=Lgals1 PE=1 SV=2
Q62881	0.003	-1.2	nucleolar protein 3
Q7TP38	0.016	-1.3	Ab2-371
O35814	0.003	-1.0	stress-induced-phosphoprotein 1

^aFold change = young/old.

enzymes such as β -enolase, phosphoglycerate kinase, and pyruvate kinase (PK) were down-regulated during aging. PK mediates one of the rate-limiting steps catalyzing the final stage of the glycolytic pathway. Decreased expression of PK during aging were previously described;^{23,46} however, some authors have found increased levels of this protein with aging.^{47,48} Proteins involved in oxidative metabolism, such as ATP synthase (subunit α), electron transfer flavoprotein and mitochondrial cytochrome complex subunits 1 and 2 were also down-regulated.

Among the up-regulated proteins were protein-arginine deiminase type-2, L-lactate dehydrogenase β chain (LDH), and mitochondrial 2-oxoglutarate/malate carrier protein.

The glioblastoma amplified sequence (GAS), also known as NIPSNAP2, is most abundant in brain, heart and skeletal muscle.^{49,50} Interestingly, this protein was one of the most down-regulated proteins in our study ($-\text{Log}_2$ fold change = -22.10), and it may be at least partially responsible for the reduction in oxidative capacity that is found in sedentary elderly adults.⁵¹ To the best of our knowledge, this is the first report of altered levels of this protein with aging and its specific role in the process requires further studies.

Most of the differentially regulated proteins identified in EDL muscles are metabolic enzymes, and in contrast with what was observed in soleus muscles, the majority of the proteins (56) were up-regulated in the aged muscle, while only 10 proteins were down-regulated in this condition (Table 3). Glycolytic enzymes such as β -enolase, α enolase, γ -enolase, PK, glyceraldehyde-3-phosphate dehydrogenase, fructose-biphosphate aldolase, phosphoglycerate kinase, and phosphoglycerate mutase were up-regulated during aging. Several proteins involved in oxidative metabolism were also up-regulated in the old EDL muscles. Among them were subunits of ATP synthase (subunit α , β , and δ), NADH-ubiquinone oxidoreductase, malate dehydrogenase, isocitrate dehydrogenase,

and citrate synthase. Increased levels of the mitochondrial enzymes such as ATP synthase, malate dehydrogenase, and isocitrate dehydrogenase have also been found during the aging of rat gastrocnemius muscles.^{46,48,52,53}

Structural Proteins

Several structural proteins were up-regulated in old soleus muscles (Table 2). Among them were troponin I, several isoforms of myosin heavy chain (4, 6, 7, and 13), and prolargin. We hypothesize that troponin (Tn) may be strongly correlated with scarponia. Recently, Zhang et al.⁵⁴ have demonstrated a role for Tn in nuclear imbalance and muscle aging, and overexpression of TnT fragments resulted in defects in nuclear shape and caused high levels of apoptosis.

Among the down-regulated proteins were actin, myosin regulatory light chain, Actn2, Col6a2, Fln C, and different chains of tropomyosin (Tm) (α -1, α -3, and α -4). Other proteins that are involved in myofibrils assembly such as PDZ and LIM domain protein 5 were also down-regulated.

In EDL muscles, among the up-regulated proteins were two α chains of type-VI collagen (Col6a2 and Col6a3), FlnC (filamin C), myosin-binding protein C, α -actinin 1, tropomyosin β chain, Actn2 (actinin, α 2), and several chains of myosin regulatory light chain (light chain 1/3, light chain 2, and light chain 4) (Table 3).

Collagen proteins play a role in maintaining the integrity of various tissues. Age-related studies of collagen indicate that collagen content varies with age and type of muscle, and a 40% increase in total collagen has been found in fast-twitch type of muscles.⁵⁵ Col6a2 and prolargin were the two proteins with the greatest increase in expression ($-\text{Log}_2$ fold change = +23.10 and +23.5, respectively).

The down-regulated group of proteins contained myosin heavy chain (chains 4, 7, and 8), cytoskeletal type II keratin, collagen α , and vimentin. Vimentin is a constituent of

Table 3. Differentially Expressed Proteins in EDL Muscles

protein accession no.	p value	$-\log_2$ fold change ^a	description
Energy Metabolism			
Q52KS1	<0.001	0.6	6-phosphofructokinase
G3V796	0.005	-0.4	acetyl-coenzyme A dehydrogenase, medium chain
Q9ER34	0.014	0.9	aconitate hydratase, mitochondrial
P10760	0.009	1.4	adenosylhomocysteinase
F1LN88	<0.001	1.6	aldehyde dehydrogenase, mitochondrial
Q63041	<0.001	0.8	α -1-macroglobulin
P04764	<0.001	1.1	α -enolase
D4AEH9	<0.001	0.6	amylol-1,6-glucosidase, 4- α -glucanotransferase, isoform CRA_a
P13221	<0.001	1.3	aspartate aminotransferase, cytoplasmic
P00507	<0.001	1.1	aspartate aminotransferase, mitochondrial
P15999	<0.001	0.8	ATP synthase subunit α , mitochondrial
P19511	<0.001	0.7	ATP synthase subunit b, mitochondrial
P10719	<0.001	0.9	ATP synthase subunit β , mitochondrial
G3V7Y3	0.001	0.6	ATP synthase subunit δ , mitochondrial
Q06647	<0.001	-0.4	ATP synthase subunit O, mitochondrial
Q7TP35	0.002	0.7	ATPase family AAA domain-containing protein 1
B4F7E5	<0.001	-0.4	ATPase, Ca ⁺⁺ transporting, cardiac muscle, fast twitch 1
P15429	<0.001	1.2	β -enolase
G3V936	<0.001	1.3	citrate synthase
B0BNC0	<0.001	0.7	Ckmt2 protein (creatine kinase, mitochondrial)
P32551	<0.001	0.7	cytochrome b-c1 complex subunit 2, mitochondrial
P10888	<0.001	0.7	cytochrome c oxidase subunit 4 isoform 1, mitochondrial
P11951	<0.001	-0.5	cytochrome c oxidase subunit 6C-2
P62898	0.003	0.8	cytochrome c, somatic
D3ZFAQ8	0.011	-0.6	cytochrome c-1 (predicted), isoform CRA_c
P11348	0.002	-26.4	dihydropteridine reductase
Q68FU3	<0.001	0.4	electron transfer flavoprotein subunit β
D3ZJT8	0.007	1.0	enolase
Q5BJ93	<0.001	1.1	enolase 1, (α)
G3V900	<0.001	0.7	fructose-bisphosphate aldolase
P09117	0.004	1.0	fructose-bisphosphate aldolase C
Q5M964	<0.001	0.7	fumarate hydratase 1
P07323	<0.001	1.0	gamma-enolase
Q6P6V0	<0.001	0.7	glucose-6-phosphate isomerase
D4A6J7	<0.001	0.8	glyceraldehyde-3-phosphate dehydrogenase
D4A3W5	<0.001	0.6	glyceraldehyde-3-phosphate dehydrogenase
P04797	<0.001	0.5	glyceraldehyde-3-phosphate dehydrogenase
O35077	<0.001	0.5	glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic
P41565	<0.001	-0.4	isocitrate dehydrogenase [NAD] subunit gamma 1, mitochondrial
P56574	<0.001	0.5	isocitrate dehydrogenase [NADP], mitochondrial
P04642	<0.001	1.9	L-lactate dehydrogenase A chain
P42123	<0.001	0.8	L-lactate dehydrogenase B chain
P15650	<0.001	1.3	long-chain specific acyl-CoA dehydrogenase, mitochondrial
O88989	<0.001	1.5	malate dehydrogenase, cytoplasmic
P04636	<0.001	1.0	malate dehydrogenase, mitochondrial
P08503	0.005	-0.4	medium-chain specific acyl-CoA dehydrogenase, mitochondrial
Q02253	0.007	0.5	methylmalonate-semialdehyde dehydrogenase [acylating], mitochondrial
G3V6H5	<0.001	-0.6	mitochondrial 2-oxoglutarate/malate carrier protein
Q5RJN0	<0.001	-0.4	NADH dehydrogenase (ubiquinone) Fe-S protein 7
B0BNE6	0.002	0.7	NADH dehydrogenase (ubiquinone) Fe-S protein 8 (predicted), isoform CRA_a
Q641Y2	0.006	1.6	NADH dehydrogenase [ubiquinone] iron-sulfur protein 2, mitochondrial
Q66HF1	0.013	1.2	NADH-ubiquinone oxidoreductase 75 kDa subunit, mitochondrial
A1ASL2	<0.001	1.2	Pgm1 protein (Fragment)
Q499Q4	<0.001	1.2	phosphoglucomutase 1
P16617	<0.001	0.8	phosphoglycerate kinase 1
P25113	<0.001	0.8	phosphoglycerate mutase 1
P16290	<0.001	0.7	phosphoglycerate mutase 2
D3ZAP9	0.004	0.7	protein Gpd1l (glycerol-3-phosphate dehydrogenase 1-like)
D4ADX5	<0.001	-0.4	protein Ndufs7(NADH dehydrogenase (ubiquinone)Fe-S protein 7)

Table 3. continued

protein accession no.	p value	$-\text{Log}_2$ fold change ^a	description
Energy Metabolism			
P49432	0.001	5.3	pyruvate dehydrogenase E1 component subunit β , mitochondrial
Q6P7S0	<0.001	0.8	pyruvate kinase
P11980	<0.001	0.8	pyruvate kinase isozymes M1/M2
F8WG21	<0.001	1.3	succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial
Q64428	<0.001	1.0	trifunctional enzyme subunit α , mitochondrial
Q60587	0.009	0.7	trifunctional enzyme subunit β , mitochondrial
P48500	0.001	1.1	triosephosphate isomerase
Structure and Cell Motility			
P51886	<0.001	1.2	lumican
D4A6M0	0.003	1.1	mitsugumin 29 (predicted)
D4A111	<0.001	1.5	procollagen, type VI, α 3 (predicted), isoform CRA_a
P45592	0.014	1.2	cofilin-1
P50609	<0.001	0.8	fibromodulin
D3ZVD7	<0.001	-0.9	keratocan (predicted)
D3ZVB7	<0.001	1.2	osteoglycin (predicted)
Q9EQP5	<0.001	23.5	prolargin
F1LYK7	<0.001	1.2	protein Cfl2 (fragment)
F1M7Q6	<0.001	-1.9	protein Myh13 (fragment) (myosin heavy chain 13)
F2Z3S8	<0.001	1.2	protein Tnnc2 (fragment) (troponin C type 2)
F1LNH3	0.003	23.1	protein Col6a2 (fragment)
F2Z3T2	0.019	1.6	tropomyosin α -3 chain
D4A115	<0.001	1.6	protein Col6a3
D3ZHA0	0.015	1.4	protein Flnc
D4A2S4	<0.001	1.3	myosin-binding protein C, slow-type
Q9Z1P2	0.006	1.2	α -actinin-1
P60711	<0.001	1.2	actin, cytoplasmic 1
Q63518	<0.001	1.2	myosin-binding protein C, slow-type (Fragment)
P68136	0.013	1.2	actin, α skeletal muscle
D3ZCV0	<0.001	1.0	protein Actn2
F1LP83	<0.001	0.6	tropomyosin β chain
P27768	<0.001	0.6	troponin I, fast skeletal muscle
P17209	<0.001	0.5	myosin light chain 4
P04466	<0.001	0.5	myosin regulatory light chain 2, skeletal muscle isoform
Q63781	0.001	0.5	myosin regulatory light chain
Q5FVG5	<0.001	0.4	similar to tropomyosin 1, embryonic fibroblast-rat, isoform CRA_c
G3V7K1	<0.001	0.4	myomesin 2
P02600	<0.001	0.4	myosin light chain 1/3, skeletal muscle isoform
G3V8B0	<0.001	-0.5	myosin-7 (Myh7)
G3V6E1	<0.001	-0.7	myosin-4 (Myh4)
F1M8F6	<0.001	-0.7	myosin-8 (Myh8)
F1LRV9	<0.001	-0.8	myosin-4 (Myh4)
Q4FZU2	0.010	-0.9	keratin, type II cytoskeletal 6A
Q6P6Q2	0.014	-1.0	keratin, type II cytoskeletal 5
P02454	<0.001	-1.1	collagen α -1(I) chain
P31000	0.006	-1.2	vimentin
F1LMU0	<0.001	-1.3	myosin-4 (Myh4)
Oxidative Stress, Detoxification, and Degradation			
P15865	<0.001	1.2	histone H1.4
D3Z8U0	<0.001	1.7	histone H2B
Q6PDW8	0.028	26.8	glutathione peroxidase
P04785	<0.001	0.7	protein disulfide-isomerase
P08009	<0.001	4.2	glutathione S-transferase Yb-3
P08010	0.001	3.3	glutathione S-transferase Mu 2
P07895	<0.001	1.5	superoxide dismutase [Mn], mitochondrial (SOD 2)
B6DYQ7	<0.001	1.5	glutathione S-transferase pi
P63018	<0.001	1.2	heat shock cognate 71 kDa protein
P07632	0.002	0.9	superoxide dismutase [Cu-Zn] (SOD 1)
P34058	<0.001	0.7	heat shock protein HSP 90- β

Table 3. continued

protein accession no.	p value	$-\text{Log}_2$ fold change ^a	description
Transport and Catabolism			
P02770	0.016	2.8	serum albumin
Q7TP52	0.026	2.7	carboxymethylenebutenolidase homologue
Q62669	<0.001	1.1	protein Hbb-b1
Q05962	<0.001	1.0	ADP/ATP translocase 1
P12346	<0.001	0.9	serotransferrin
Q9ER30	<0.001	0.9	Kelch repeat and BTB domain-containing protein 10
P02767	<0.001	-0.5	transthyretin
Additional Differentially Expressed Proteins			
P35213	<0.001	0.4	14-3-3 protein β/α
P62260	<0.001	1.0	14-3-3 protein epsilon
P68511	0.018	-1.5	14-3-3 protein ETA
P63102	0.015	22.8	14-3-3 protein ζ/δ
P39069	<0.001	1.0	adenylate kinase isoenzyme 1
P14046	<0.001	1.2	α -1-inhibitor 3
P08461	0.002	2.1	dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex, mitochondrial
P06399	0.002	0.8	fibrinogen α chain
P06866	<0.001	0.7	haptoglobin
Q5I0J0	0.004	1.8	immunoglobulin heavy chain (gamma polypeptide)
F1MAN9	<0.001	1.2	myelin protein P0
P19804	<0.001	1.2	nucleoside diphosphate kinase B
P02625	<0.001	0.7	parvalbumin α
G3 V6L9	0.017	-0.4	peptidyl-prolyl cis-trans isomerase
P49744	0.008	1.1	thrombospondin-4
Q63581	<0.001	3.9	rat T-kininogen (T-KG)
Q63654	<0.001	2.8	polyubiquitin (fragment)
P14141	<0.001	2.5	carbonic anhydrase 3
F1LPQ6	0.004	1.7	uncharacterized protein (fragment)
P08932	<0.001	1.6	T-kininogen 2
P31044	<0.001	1.4	phosphatidylethanolamine-binding protein 1
Q5BJZ2	<0.001	1.4	LOC367586 protein
P20717	0.005	1.3	protein-arginine deiminase type-2
P07943	0.001	1.3	aldose reductase
F1LNBS	<0.001	1.2	murinoglobulin-2
Q01129	<0.001	1.2	decorin
F1LX07	0.002	1.2	protein LOC100360985 (fragment)
P11517	<0.001	1.3	hemoglobin subunit β -2
P02650	<0.001	0.6	apolipoprotein E
P62632	<0.001	1.0	elongation factor 1- α 2
P05197	0.004	1.0	elongation factor 2
F8WG42	0.002	-0.4	eukaryotic translation initiation factor 3 subunit J (fragment)
Q3T1J1	<0.001	0.7	eukaryotic translation initiation factor 5A-1
B1H216	<0.001	1.3	hemoglobin α , adult chain 2
P02091	<0.001	1.9	hemoglobin subunit β -1
P20059	<0.001	1.0	hemopexin
F1LNF1	<0.001	0.7	heterogeneous nuclear ribonucleoproteins A2/B1 (fragment)
P85125	0.004	1.3	polymerase I and transcript release factor
F1LWG8	<0.001	1.1	uncharacterized protein (fragment)
P19633	<0.001	1.0	calsequestrin-1
D3ZYG6	0.009	0.9	protein Rtn2 (reticulum 2)
Q6WN19	0.009	0.9	RTN2-B
Q6IMZ3	<0.001	0.9	annexin
Q07936	0.004	0.8	annexin A2
D3ZQT0	<0.001	0.8	uncharacterized protein
Q9ERS3	0.030	0.8	voltage-dependent calcium channel subunit α -2/ δ -1
P04276	0.004	0.6	vitamin D-binding protein
D4AAC3	<0.001	0.6	protein Sh3bgr
D4A3D2	<0.001	0.4	protein Smyd1
Q64649	0.024	-0.5	phosphorylase b kinase regulatory subunit α , skeletal muscle isoform
Q6IRH6	<0.001	-0.6	phosphate carrier protein, mitochondrial

Table 3. continued

protein accession no.	p value	$-\text{Log}_2$ fold change ^a	description
Additional Differentially Expressed Proteins			
D3Z8G0	0.019	-0.6	phosphorylase b kinase regulatory subunit α , skeletal muscle isoform

^aFold change = young/old.

cytoskeletal filaments, and in gastrocnemius rat muscle it is increased with age.⁴⁷

Oxidative Stress, Detoxification, and Degradation

In the up-regulated proteins of the soleus muscle, we identified two isoforms of glutathione S-transferase (GST) and ferritin (Table 2). Increased levels of GST suggest increased detoxification of cytotoxic products, while higher levels of ferritin indicate an altered iron metabolism.

Some heat shock proteins such as Hsp90, heat shock 71 kDa protein, Hspb6, endoplasmic, and chaperonin subunit 8 were down-regulated. Deficits in chaperone function have been reported in several age-related diseases, and literature also supports that synthesis of heat shock proteins is impaired in aging.^{56–58} The DNA damage-binding protein 1 is responsible for repair of UV-damaged DNA;⁵⁹ therefore, reduced levels of this protein may leave the DNA more susceptible to damage.

In EDL muscles, proteins involved in scavenging of ROS as well as enzymes involved in the detoxification of cytotoxic products were found to be up-regulated in old muscles (Table 3). Among them were several isoforms of GSTs, glutathione peroxidase (GPx), heat shock proteins (71 kDa protein and HSP 90- β), and the antioxidant enzymes superoxide dismutase (SOD) [Cu/Zn] and SOD [Mn]. The up-regulation of glutathione transferase suggests increased detoxification of cytotoxic products, while the up-regulation of GPx may be a counterbalance for increased levels of oxidative stress. Higher levels of chaperones may be needed to fold an increased number of misfolded proteins in aged muscles.

Transport and Catabolism

In the transport and catabolism class, all of the identified proteins were down-regulated in aged soleus muscles (Table 2). These proteins are clathrin heavy chain 1, complement component 4, the EH domain-containing protein 2, serum albumin, and zero β -globin.

The complement component 4 (C4) is involved in classical complement activation, and there is evidence in the literature suggesting that the immune system deteriorates with age, rendering old animals less able to mount an immune response.⁶¹ C4 prevents early stage autoimmune diseases,⁶² and mice with a disrupted C4 locus show impaired immune response.⁶³

In EDL muscles an ADP/ATP translocase 1, BSA, carboxymethylenebutenolidase, kelch repeat and BTB domain-containing protein, protein Bgg-b1, serum albumin, and serotransferrin were up-regulated with aging, while transthyretin was down-regulated (Table 3).

Serum albumin was found to be increased, suggesting a shift to more aerobic-oxidative metabolism in aged fibers.²³ Serum albumin plays a crucial role in maintaining the osmotic blood pressure and apparently also plays a role as radical and heme scavenger.⁶⁴ Serotransferrin was also increased, and it is involved in the control of stress and iron levels by increasing iron uptake.

Additional Differentially Expressed Proteins

In soleus muscles (Table 2), among the up-regulated proteins present were putative uncharacterized proteins: T-kininogen 2, α -1-macroglobulin, NAD(P)H-hydrate epimerase, and Obg-like ATPase 1. T-kininogens (T-KG) expression has been shown to be considerably increased during aging in the liver of Sprague–Dawley rats⁶⁵ and serum of Fisher rats. Some authors^{66,67} have suggested that T-KG may be a reliable biomarker for senescence in rats.^{66,67} This protein may be involved in the deterioration of the immune system that occurs with aging. Annexin, cadherin 13, α -1-inhibitor 3, N(G),N(G)-dimethylarginine dimethylaminohydrolase 1, basic leucine zipper, W2 domain-containing protein 2, GBAS, and two uncharacterized proteins were present at reduced levels.

Several proteins were up-regulated in old EDL muscles (Table 3). Among them, T-KG was also up-regulated in soleus muscles, presenting a greater increase in EDL ($-\text{Log}_2$ fold change = 1.8 and 3.9 for soleus and EDL, respectively). Calsequestrin and parvalbumin (PV) are Ca^{2+} binding proteins that were up-regulated in old EDL muscles. Ca^{2+} translocation from the myofibril to the sarcoplasmic reticulum (SR) is facilitated by PV in the fast-twitch skeletal muscle. The energy-dependent Ca^{2+} uptake into the SR is mediated by the SR ATPase, which is regulated by both Ca^{2+} - and CaM-dependent phosphorylation. The Ca^{2+} cycle is completed by binding of Ca^{2+} to the high-capacity, low-affinity Ca^{2+} -binding protein calsequestrin. Therefore, altered levels of both PV and calsequestrin may be related to the age-related impairment of intrinsic SR function and influence the speed of contraction in old fast-twitch motor units.⁶⁸

A voltage-dependent calcium channel (Q9ERS3) was also up-regulated in old muscles. Up-regulation of anion-selective channel proteins in aging has already been reported,⁴⁷ and it may be a way to facilitate the access of kinases to ATP, therefore bypassing the restriction exerted by the mitochondrial outer membrane on the permeability for metabolites.⁶⁹

Carbonic anhydrase III (CAIII) is a cytosolic zinc-containing enzyme that facilitates the transport of CO_2 in skeletal muscle by catalyzing the reversible hydration of CO_2 . This activity is probably involved in the maintenance of ionic balance and acid–base homeostasis within the muscle tissue. This enzyme is detected in large amounts in red slow-twitch muscles such as soleus^{70,71} but is virtually absent from adult white fast-twitch muscle, such as rodent anterior tibialis (AT) and EDL.⁷⁰ The age-dependent expression of this protein is still controversial in the literature. Whereas some authors have demonstrated down regulation of CA3 in slow-type muscles such as gastrocnemius,^{24,46,60} others have shown up-regulation^{47,72} in the same muscle type.

Oxidative Stress in Plasma and Muscle

Taking in consideration the increased levels of antioxidant and enzymes involved in detoxification such as GST, SOD [Mn], and SOD [Cu/Zn] that were up-regulated in both soleus and EDL muscles, we decided to verify whether there were

increased levels of oxidative stress in plasma and muscle of the old animals.

We determined oxidative stress levels in plasma and muscle of young and old animals using the FOX assay.³⁶ For the experiment at hand, we were unable to show statistical differences in the plasma of young versus old animals (1.08 ± 0.10 mmol·mg⁻¹ and 1.12 ± 0.07 mmol·mg⁻¹, $p = 0.33$, respectively). However, muscle oxidative stress was higher in old animals for both soleus and EDL muscles (17.71 ± 1.42 mmol·mg⁻¹ and 110.69 ± 16.21 mmol·mg⁻¹, $p < 0.0001$, for young and old soleus, respectively, and 16.54 ± 1.13 mmol·mg⁻¹ and 131.89 ± 26.06 mmol·mg⁻¹, $p < 0.0001$ for young and old EDL, respectively) (Figure 2).

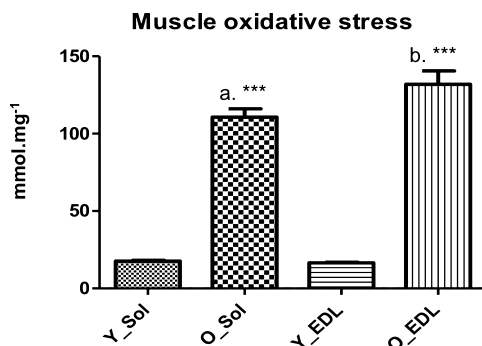


Figure 2. Soleus and EDL FOX values in the young and old group. Y_Sol = young group, soleus; O_Sol = old group, soleus; Y_EDL = young group, EDL; O_EDL = old group, EDL. *a* = $p < 0.0001$ vs Y_Sol; *b* = $p < 0.0001$ vs Y_EDL; *c* = $p < 0.0001$ vs YH. No statistical differences were found between O_Sol × O_EDL and Y_Sol × Y_EDL.

Creatine Kinase and Pyruvate Kinase Activity

Creatine kinase catalyzes the reversible transfer of phosphate between ATP and creatine. The effect of aging on the levels of muscle CK has been controversial. We found that creatine kinase was up-regulated in old EDL muscles, while there were no changes in expression for soleus muscles. Capitanio et al.⁴⁷ and O'Connell et al.²³ found decreased levels of this enzyme in gastrocnemius muscles, while Doran et al.⁴⁶ found a two-fold increase in the same muscle type and Donoghue et al.⁴⁸ reported differential effects of CK in 2D-gels of gastrocnemius muscles.

Therefore, we evaluated CK activity to investigate whether an altered expression also resulted in different activities. When comparing the different muscles, we found that CK activity was higher in EDL muscles for both young and old (506.88 ± 145.62 and 159.74 ± 63.22 , $p < 0.0001$ for young EDL and young soleus, respectively, and 347.52 ± 194.31 and 128.52 ± 78.52 , $p = 0.0022$ for old EDL and old soleus muscles, respectively) (Figure 3).

When we compared young versus old soleus muscles, we found that CK activity was not different (159.74 ± 63.22 and 128.52 ± 78.52 , $p = 0.25$ for young old muscles, respectively) (Figure 3). On the other hand, CK activity was lower in old EDL muscles (506.88 ± 145.62 and 347.52 ± 194.31 , $p = 0.01$ for young and old muscles, respectively) (Figure 3). Therefore, increased levels of CK may be a physiological response to counterbalance the reduced activity of this enzyme with aging. An age-related decline in this enzyme's activity has been shown to occur in humans and rodents.^{73,74}

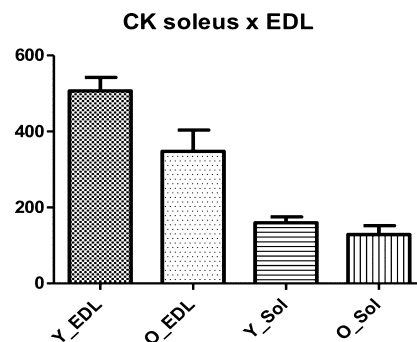


Figure 3. Soleus and EDL creatine kinase (CK) activity in the young and old group. Y_Sol = young group, soleus; O_Sol = old group, soleus; Y_EDL = young group, EDL; O_EDL = old group, EDL. (a) Young EDL × young soleus, $p < 0.0001$; (b) old EDL × old soleus, $p = 0.0022$; (c) young EDL × old EDL, $p = 0.01$.

PK, a key enzyme in the glycolytic pathway, was differentially regulated in both soleus and EDL muscles. In EDL, it was up-regulated in the old muscles, while in soleus muscles, the result was the opposite. We evaluated PK activity and found that it was higher for EDL when compared with soleus (7.43 ± 1.09 and 3.30 ± 0.42 , $p < 0.0001$ for young EDL and young soleus muscles, respectively, and 16.55 ± 9.25 and 2.47 ± 0.79 , $p < 0.0001$ for old EDL and old soleus, respectively) (Figure 4).

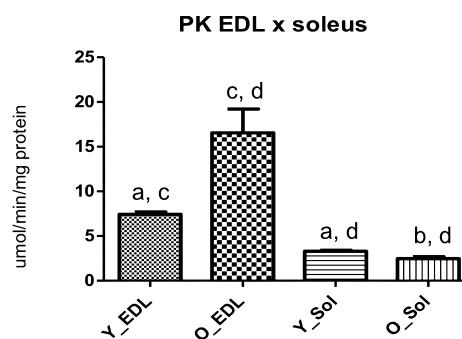


Figure 4. Soleus and EDL pyruvate kinase (PK) activity in the young and old group. Y_Sol = young group, soleus; O_Sol = old group, soleus; Y_EDL = young group, EDL; O_EDL = old group, EDL. (a) Young EDL × young soleus, $p < 0.0001$; (b) old EDL × old soleus, $p < 0.0001$; (c) young EDL × old EDL, $p < 0.0001$; and (d) young soleus × old soleus, $p = 0.0018$.

Furthermore, in EDL old muscles, there was an increased activity for this enzyme (7.43 ± 1.09 and 16.55 ± 9.25 , $p < 0.0001$ for young and old, respectively) (Figure 4), while for soleus muscles, PK activity decreased with age (3.30 ± 0.42 and 2.47 ± 0.79 , $p = 0.0018$ for young and old, respectively) (Figure 4).

DISCUSSION

Over the past few years several proteomic studies have investigated changes in the protein complement of aging skeletal muscles in humans and animal considered as aging models (reviewed by Doran et al.²⁴). Most of these studies used 2DE and mass spectrometry as the main tool to identify differentially regulated proteins (reviewed by Doran et al.²⁴).

As far as we know, the use of TMT coupled to LC/LC/MS/MS mass-spectrometry-based proteomics allowed us to identify and quantify a larger number of proteins from both soleus and

Table 4. Metabolic Enzymes Regulated in Opposite Ways in Soleus and EDL Muscles

protein accession no.	description	SOLEUS		EDL	
		p value	−Log ₂ fold change	p value	−Log ₂ fold change
P13221	aspartate aminotransferase, cytoplasmic	<0.001	−0.8	<0.001	1.3
P13221	aspartate aminotransferase, mitochondrial	0.001	−1.1	<0.001	1.1
F1LP05	ATP synthase subunit α	<0.001	−0.9	<0.001	0.8
P15429	β -enolase	<0.001	−0.6	<0.001	1.2
P10888	cytochrome c oxidase subunit 4 isoform 1, mitochondrial	<0.001	−1.3	<0.001	0.7
Q68FU3	electron transfer flavoprotein subunit β	0.005	−1.0	<0.001	0.4
P42123	L-lactate dehydrogenase B chain	<0.001	0.6	<0.001	0.8
P16617	phosphoglycerate kinase 1	<0.001	−0.9	<0.001	0.8
P25113	phosphoglycerate mutase 1	<0.001	−1.4	<0.001	0.8
Q6P7S0	pyruvate kinase	0.001	−0.9	<0.001	0.8
Q64428	trifunctional enzyme subunit α , mitochondrial	0.005	−0.7	<0.001	1.0

EDL muscles than any preceding report (total of 5300 proteins, out of which 252 were differentially regulated in aging).

We chose to analyze two different muscles, one in which fast-twitch type-II fibers predominate (EDL) and the other in which the type-I, slow-twitch fibers are more abundant (soleus) to have a comprehensive overview of the effects of aging on different types of muscles. Our investigation revealed many differences in the protein expression pattern of young versus old muscles, and there were considerable differences between soleus and EDL muscles as well.

Histological analysis of soleus and EDL showed striking features of aged muscles which include extensive variability in fiber diameter, a higher frequency of longitudinal splitting, an increased number of centrally located nuclei and thickening of the endomysium,^{25,75} which indicate that the old rats used in this study (24 months old) did show signs of sarcopenia.

Furthermore, the SI, which indicates the degree of muscle wasting, was decreased for both muscles, and the decrease was more pronounced in the soleus than in the EDL muscle. Although sarcopenia is widely considered to preferentially impact fast twitch muscles,⁷⁶ this notion may not be applied at more advanced ages. Hagen et al.⁷⁷ have shown that the relative protection of the slow twitch soleus muscle from age-related atrophy is present only until middle age with a great degree of atrophy present thereafter.

Our results are in agreement with those of Carter et al.,⁷⁸ which have shown that slow twitch soleus muscles undergo large phenotypic alterations in very old age and also that, for several measures, it is of greater magnitude than fast twitch muscle. Furthermore, there are several reports of atrophy, force decline, altered MHC expression, and myofiber loss in aged rat soleus muscles.^{77,79,80}

The identification of several metabolic enzymes among the differentially regulated proteins suggests perturbations in the energy metabolism of old skeletal muscles. Only one enzyme, L-lactate dehydrogenase (LDH), was up-regulated in both soleus and EDL muscles. The effect of aging in the expression levels of this enzyme is still controversial in the literature. While Doran et al.⁴⁶ and Donoghue et al.⁴⁸ found LDH to be one of the most drastically up-regulated proteins in old gastrocnemius muscle, Capitanio et al.⁴⁷ showed decreased levels of this enzyme in the same muscle type.

The mitochondria is the major cellular site for ATP production. During aging, the respiratory chain function (RCF) falls considerably in humans between ages 17 and 90 accompanied by an impaired function of cytochrome c oxidase in old muscles.⁸¹

Down-regulation of several components of the respiratory chain was observed for soleus muscles, while this trend was not true for EDL muscles, in which some components were up while others were down-regulated. Soleus muscles may be more susceptible to oxidative stress that plays a role in mitochondrial dysfunction and may therefore contribute to decreased expression of mitochondrial enzymes.

Interestingly, 11 metabolic enzymes were regulated in opposite ways in both muscles (down-regulated in soleus and up-regulated in EDL) (Table 4). These enzymes are involved in amino-acid metabolism (aspartate aminotransferase, both the cytoplasmic and mitochondrial isoforms), glycolysis/gluconeogenesis (β -enolase, phosphoglycerate kinase, phosphoglycerate mutase, and pyruvate kinase), and oxidative metabolism (ATP synthase subunit β , cytochrome c oxidase, and electron-transfer flavoprotein).

The GAS, according to our quantitative strategy, showed to be the most down-regulated protein (Log₂ fold change = −22.1), along with a yet uncharacterized protein (D3ZH98, Log₂ fold change = −21.1). Zero β -globin, a protein that participates in oxygen transport, was also drastically decreased (Log₂ fold change = −16.9). Therefore, the down-regulation of both GAS and zero β -globin may contribute for the reduction of oxidative capacity in the elderly.

In muscle, the interaction between myosin and actin filaments is responsible for force production and several of the differentially expressed proteins belong to the myosin-actin complex, suggesting that force production may be compromised in old rats due to a more disorganized structure (some of the components were down- and other components were up-regulated). In soleus muscles, the structural proteins Tm (α and β chains) were among the most down-regulated proteins (Log₂ fold change = −18.4 and −18.2, respectively), while troponin I was among the most up-regulated (Log₂ fold change = 17.8).

The thin filament regulatory proteins Tn and Tm are essential for contraction of striated muscle (skeletal and cardiac), which is regulated by the concentration of intracellular calcium.^{82,83} Such a drastic decrease in Tm proteins may be responsible for a lower force contraction and muscle velocity in the elderly. The up-regulation of Tn confirms previous studies that have described this protein as a potential new biomarker for sarcopenia.⁴⁵

During aging there is an increase in the production of ROS due to the functional deterioration of mitochondria, and the increase in free radicals generation may contribute to several age-related pathologies.^{84–86} Enhanced ROS production may contribute to an accumulation of mtDNA damage and

increased apoptosis of muscles fibers, which may represent a key mechanism underlying sarcopenia.⁸⁷ GST was up-regulated in both muscles, while two of the main antioxidant enzymes, GPx, which controls the levels of hydrogen peroxide, and SOD, were up-regulated only in EDL muscles. In EDL muscles, GPx was among the most up-regulated (Log2 fold change = 26.8). Up-regulation of GPx has been reported in Duchenne muscular dystrophy (DMD), a progressive muscle wasting disorder that impairs muscle function and ultimately results in muscle degeneration and death.⁸⁸ Therefore, an increase in GPx expression is most likely a compensatory attempt to counterbalance increased levels of oxidative stress.

We also performed an assay (FOX) to determine the levels of oxidative stress in plasma and muscles. We confirmed that old muscles had significantly higher oxidative stress than young muscles, which suggests that a higher level of antioxidant enzymes may be an attempt to counterbalance oxidative stress. However, we were unable to find a convincing statistical difference between soleus and EDL. Nevertheless, we encourage future studies to consider complementary features such as the activity of specific antioxidant enzymes or the content of reduced glutathione to further investigate the difference between these muscles. The soleus muscle is known to be more susceptible to oxidative stress than EDL due to its higher oxidative potential when compared with a fast-twitch muscle.⁸⁹

There is still controversy in the literature to whether aging increases or decreases the enzymatic antioxidant defenses in the cell. Studies conducted in both humans and rats have indicated that although the antioxidant system undergoes significant changes, both up- and down-regulation or no change of the antioxidant enzymes may occur, suggesting that the results may be dependent on muscle fiber composition and other factors such as age and sex.^{48,60,87,90}

Our results show that aging affects slow-twitch and fast-twitch muscles in different ways and that the most significant differences were found for metabolic enzymes. Furthermore, EDL muscles had a higher number of differentially expressed proteins (16.2% of the total identified proteins had a change in abundance against 3.8% of the total proteins for soleus muscles), suggesting that aging affects protein expression more drastically in a fast-twitch type of muscle.

Compared with previous age-related muscle proteome studies,²⁴ we identified novel muscle proteins with a change in abundance during aging. Among these were the GAS, with a marked down-regulated in old soleus muscles, the uncharacterized protein D3ZH98, zero β -globin, and prolargin.

Therefore, the confirmation of many already identified proteins involved with aging and the identification of new candidate biomarkers will help in the establishment of a more comprehensive database for the study of sarcopenia and aid in the development of new diagnostic methods and treatment for age-associated muscular disorders.

■ ASSOCIATED CONTENT

📄 Supporting Information

Muscle fiber and cross-sectional area and hematoxylin and eosin (HE) staining of EDL and soleus muscles. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Authors

*(D.F.S.C.) E-mail: dseixas@usp.br.
*(J.R.Y.) E-mail: jyates@scripps.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by FAPESP, grants 2009/52022-2, 2010/10852-6 and 2012/07319-0. J.J.M. and J.R.Y. were supported by the National Center for Research Resources (P41RR011823), National Institute of General Medical Sciences (P41GM103533), and National Institute on Aging (AG027463). P.C.C. was supported by Fiocruz – PDTIS. D.B.L. was supported by IBMP and CAPES.

■ REFERENCES

- (1) Melton, L. J., III; Khosla, S.; Crowson, C. S.; O'Connor, M. K.; O'Fallon, W. M.; Riggs, B. L. Epidemiology of sarcopenia. *J. Am. Geriatr. Soc.* **2000**, *48* (6), 625–630.
- (2) Morley, J. E.; Baumgartner, R. N.; Roubenoff, R.; Mayer, J.; Nair, K. S. Sarcopenia. *J. Lab. Clin. Med.* **2001**, *137* (4), 231–243.
- (3) Greenlund, L. J.; Nair, K. S. Sarcopenia—consequences, mechanisms, and potential therapies. *Mech. Ageing Dev.* **2003**, *124* (3), 287–299.
- (4) Vandervoort, A. A.; Symons, T. B. Functional and metabolic consequences of sarcopenia. *Can. J. Appl. Physiol.* **2001**, *26* (1), 90–101.
- (5) Janssen, I.; Heymsfield, S. B.; Ross, R. Low relative skeletal muscle mass (sarcopenia) in older persons is associated with functional impairment and physical disability. *J. Am. Geriatr. Soc.* **2002**, *50* (5), 889–896.
- (6) Doherty, T. J. Invited review: Aging and sarcopenia. *J. Appl. Physiol.* **2003**, *95* (4), 1717–1727.
- (7) Dirks, A.; Leeuwenburgh, C. Apoptosis in skeletal muscle with aging. *Am. J. Physiol.: Regul., Integr. Comp. Physiol.* **2002**, *282* (2), R519–R527.
- (8) Renault, V.; Thornell, L. E.; Eriksson, P. O.; Butler-Browne, G.; Mouly, V. Regenerative potential of human skeletal muscle during aging. *Ageing Cell* **2002**, *1* (2), 132–139.
- (9) Vandervoort, A. A. Aging of the human neuromuscular system. *Muscle Nerve* **2002**, *25* (1), 17–25.
- (10) Balagopal, P.; Rooyackers, O. E.; Adey, D. B.; Ades, P. A.; Nair, K. S. Effects of aging on in vivo synthesis of skeletal muscle myosin heavy-chain and sarcoplasmic protein in humans. *Am. J. Physiol.* **1997**, *273* (4 Pt 1), E790–E800.
- (11) Pastoris, O.; Boschi, F.; Verri, M.; Baiardi, P.; Felzani, G.; Vecchiet, J.; Dossena, M.; Catapano, M. The effects of aging on enzyme activities and metabolite concentrations in skeletal muscle from sedentary male and female subjects. *Exp. Gerontol.* **2000**, *35* (1), 95–104.
- (12) Taylor, D. J.; Kemp, G. J.; Thompson, C. H.; Radda, G. K. Ageing: effects on oxidative function of skeletal muscle in vivo. *Mol. Cell. Biochem.* **1997**, *174* (1–2), 321–324.
- (13) Carmeli, E.; Coleman, R.; Reznick, A. Z. The biochemistry of aging muscle. *Exp. Gerontol.* **2002**, *37* (4), 477–489.
- (14) Harman, D. The aging process. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78* (11), 7124–7128.
- (15) Peterson, C. M.; Johannsen, D. L.; Ravussin, E. Skeletal Muscle Mitochondria and Aging: A Review. *J. Aging Res.* **2012**, 1–20.
- (16) Kanski, J.; Alterman, M. A.; Schoneich, C. Proteomic identification of age-dependent protein nitration in rat skeletal muscle. *Free Radicals Biol. Med.* **2003**, *35* (10), 1229–1239.
- (17) Kanski, J.; Hong, S. J.; Schoneich, C. Proteomic analysis of protein nitration in aging skeletal muscle and identification of nitrotyrosine-containing sequences in vivo by nano-electrospray

- ionization tandem mass spectrometry. *J. Biol. Chem.* **2005**, *280* (25), 24261–24266.
- (18) Picc, I.; Listrat, A.; Alliot, J.; Chambon, C.; Taylor, R. G.; Bechet, D. Differential proteome analysis of aging in rat skeletal muscle. *FASEB J.* **2005**, *19* (9), 1143–1145.
- (19) Isfort, R. J. Proteomic analysis of striated muscle. *J. Chromatogr. B* **2002**, *771* (1–2), 155–165.
- (20) Doran, P.; Donoghue, P.; O'Connell, K.; Gannon, J.; Ohlendieck, K. Proteomic profiling of pathological and aged skeletal muscle fibres by peptide mass fingerprinting (Review). *Int. J. Mol. Med.* **2007**, *19* (4), 547–564.
- (21) Doran, P.; Gannon, J.; O'Connell, K.; Ohlendieck, K. Proteomic profiling of animal models mimicking skeletal muscle disorders. *Proteomics: Clin. Appl.* **2007**, *1* (9), 1169–1184.
- (22) Gelfi, C.; Viganò, A.; Ripamonti, M.; Pontoglio, A.; Begum, S.; Pellegrino, M. A.; Grassi, B.; Bottinelli, R.; Wait, R.; Cerretelli, P. The human muscle proteome in aging. *J. Proteome Res.* **2006**, *5* (6), 1344–1353.
- (23) O'Connell, K.; Gannon, J.; Doran, P.; Ohlendieck, K. Proteomic profiling reveals a severely perturbed protein expression pattern in aged skeletal muscle. *Int. J. Mol. Med.* **2007**, *20* (2), 145–153.
- (24) Doran, P.; Donoghue, P.; O'Connell, K.; Gannon, J.; Ohlendieck, K. Proteomics of skeletal muscle aging. *Proteomics* **2009**, *9* (4), 989–1003.
- (25) Edstrom, E.; Altun, M.; Bergman, E.; Johnson, H.; Kullberg, S.; Ramirez-Leon, V.; Ulfhake, B. Factors contributing to neuromuscular impairment and sarcopenia during aging. *Physiol. Behav.* **2007**, *92* (1–2), 129–135.
- (26) Gannon, J.; Staunton, L.; O'Connell, K.; Doran, P.; Ohlendieck, K. Phosphoproteomic analysis of aged skeletal muscle. *Int. J. Mol. Med.* **2008**, *22* (1), 33–42.
- (27) Hojlund, K.; Bowen, B. P.; Hwang, H.; Flynn, C. R.; Madireddy, L.; Geetha, T.; Langlais, P.; Meyer, C.; Mandarino, L. J.; Yi, Z. In vivo phosphoproteome of human skeletal muscle revealed by phosphopeptide enrichment and HPLC-ESI-MS/MS. *J. Proteome Res.* **2009**, *8* (11), 4954–4965.
- (28) Cieniewski-Bernard, C.; Bastide, B.; Lefebvre, T.; Lemoine, J.; Mounier, Y.; Michalski, J. C. Identification of O-linked N-acetylglucosamine proteins in rat skeletal muscle using two-dimensional gel electrophoresis and mass spectrometry. *Mol. Cell. Proteomics* **2004**, *3* (6), 577–585.
- (29) O'Connell, K.; Doran, P.; Gannon, J.; Ohlendieck, K. Lectin-based proteomic profiling of aged skeletal muscle: decreased pyruvate kinase isozyme M1 exhibits drastically increased levels of N-glycosylation. *Eur. J. Cell Biol.* **2008**, *87* (10), 793–805.
- (30) Hedou, J.; Bastide, B.; Page, A.; Michalski, J. C.; Morelle, W. Mapping of O-linked beta-N-acetylglucosamine modification sites in key contractile proteins of rat skeletal muscle. *Proteomics* **2009**, *9* (8), 2139–2148.
- (31) Feng, J.; Xie, H.; Meany, D. L.; Thompson, L. V.; Arriaga, E. A.; Griffin, T. J. Quantitative proteomic profiling of muscle type-dependent and age-dependent protein carbonylation in rat skeletal muscle mitochondria. *J. Gerontol., Ser. A* **2008**, *63* (11), 1137–1152.
- (32) Rivero, J. L.; Talmadge, R. J.; Edgerton, V. R. Correlation between myofibrillar ATPase activity and myosin heavy chain composition in equine skeletal muscle and the influence of training. *Anat. Rec.* **1996**, *246* (2), 195–207.
- (33) Gupta, R. C.; Misulis, K. E.; Dettbarn, W. D. Activity dependent characteristics of fast and slow muscle: biochemical and histochemical considerations. *Neurochem. Res.* **1989**, *14* (7), 647–655.
- (34) Washburn, M. P.; Wolters, D.; Yates, J. R., III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–247.
- (35) Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895–1904.
- (36) Nourooz-Zadeh, J. Ferrous ion oxidation in presence of xylenol orange for detection of lipid hydroperoxides in plasma. *Methods Enzymol.* **1999**, *300*, 58–62.
- (37) Ainsworth, S.; MacFarlane, N. A kinetic study of rabbit muscle pyruvate kinase. *Biochem. J.* **1973**, *131* (2), 223–236.
- (38) Dinovo, E. C.; Miyada, D. S.; Nakamura, R. M. Evaluation of direct and indirect coupled enzyme assay systems for measurement of creatine kinase activity. *Clin. Chem.* **1973**, *19* (9), 994–997.
- (39) McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., III. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **2004**, *18* (18), 2162–2168.
- (40) Xu, T.; Venable, J. D.; Park, S. K.; Cociorva, D.; Liao, L.; Wohlschlegel, J.; Hewel, J.; Yates, J. R. ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol. Cell. Proteomics* **2006**, *5*, S174.
- (41) Carvalho, P. C.; Fischer, J. S.; Xu, T.; Yates, J. R.; Barbosa, V. C. PatternLab: from Mass Spectra to Label-Free Differential Shotgun Proteomics. In *Current Protocols in Bioinformatics*; Wiley: New York, 2012.
- (42) Carvalho, P. C.; Fischer, J. S.; Xu, T.; Cociorva, D.; Balbuena, T. S.; Valente, R. H.; Perales, J.; Yates, J. R., III; Barbosa, V. C. Search engine processor: Filtering and organizing peptide spectrum matches. *Proteomics* **2012**, *12* (7), 944–949.
- (43) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. 1. *J. R. Stat. Soc. B* **1995**, *57*, 289–300.
- (44) Carvalho, P. C.; Fischer, J. S.; Chen, E. I.; Yates, J. R., III; Barbosa, V. C. PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinf.* **2008**, *9*, 316.
- (45) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **2007**, *6* (9), 3549–3557.
- (46) Doran, P.; O'Connell, K.; Gannon, J.; Kavanagh, M.; Ohlendieck, K. Opposite pathobiochemical fate of pyruvate kinase and adenylate kinase in aged rat skeletal muscle as revealed by proteomic DIGE analysis. *Proteomics* **2008**, *8* (2), 364–377.
- (47) Capitanio, D.; Vasso, M.; Fania, C.; Moriggi, M.; Viganò, A.; Procacci, P.; Magnaghi, V.; Gelfi, C. Comparative proteomic profile of rat sciatic nerve and gastrocnemius muscle tissues in ageing by 2-D DIGE. *Proteomics* **2009**, *9* (7), 2004–2020.
- (48) Donoghue, P.; Staunton, L.; Mullen, E.; Manning, G.; Ohlendieck, K. DIGE analysis of rat skeletal muscle proteins using nonionic detergent phase extraction of young adult versus aged gastrocnemius tissue. *J. Proteomics* **2010**, *73* (8), 1441–1453.
- (49) Seroussi, E.; Pan, H. Q.; Kedra, D.; Roe, B. A.; Dumanski, J. P. Characterization of the human NIPSNAP1 gene from 22q12: a member of a novel gene family. *Gene* **1998**, *212* (1), 13–20.
- (50) Wang, X. Y.; Smith, D. I.; Liu, W.; James, C. D. GBAS, a novel gene encoding a protein with tyrosine phosphorylation sites and a transmembrane domain, is co-amplified with EGFR. *Genomics* **1998**, *49* (3), 448–451.
- (51) Conley, K. E.; Jubrias, S. A.; Esselman, P. C. Oxidative capacity and ageing in human muscle. *J. Physiol.* **2000**, *526* (Pt 1), 203–210.
- (52) Chang, J.; Cornell, J. E.; Van Remmen, H.; Hakala, K.; Ward, W. F.; Richardson, A. Effect of aging and caloric restriction on the mitochondrial proteome. *J. Gerontol., Ser. A* **2007**, *62* (3), 223–234.
- (53) O'Connell, K.; Ohlendieck, K. Proteomic DIGE analysis of the mitochondria-enriched fraction from aged rat skeletal muscle. *Proteomics* **2009**, *9* (24), 5509–5524.
- (54) Zhang, T.; Birbrair, A.; Wang, Z. M.; Taylor, J.; Messi, M. L.; Delbono, O. Troponin T nuclear localization and its role in aging skeletal muscle. *Age (Dordrecht, Neth.)* **2013**, *35* (2), 353–370.
- (55) Mohan, S.; Radha, E. Age-related changes in rat muscle collagen. *Gerontology* **1980**, *26* (2), 61–67.

- (56) Sohal, R. S.; Orr, W. C. Relationship between antioxidants, prooxidants, and the aging process. *Ann. N.Y. Acad. Sci.* **1992**, *663*, 74–84.
- (57) Hall, D. M.; Xu, L.; Drake, V. J.; Oberley, L. W.; Oberley, T. D.; Moseley, P. L.; Kregel, K. C. Aging reduces adaptive capacity and stress protein expression in the liver after heat stress. *J. Appl. Physiol.* **2000**, *89* (2), 749–759.
- (58) Pahlavani, M. A.; Harris, M. D.; Moore, S. A.; Weindruch, R.; Richardson, A. The expression of heat shock protein 70 decreases with age in lymphocytes from rats and rhesus monkeys. *Exp. Cell Res.* **1995**, *218* (1), 310–318.
- (59) Dualan, R.; Brody, T.; Keeney, S.; Nichols, A. F.; Admon, A.; Linn, S. Chromosomal localization and cDNA cloning of the genes (DDB1 and DDB2) for the p127 and p48 subunits of a human damage-specific DNA binding protein. *Genomics* **1995**, *29* (1), 62–69.
- (60) Lombardi, A.; Silvestri, E.; Cioffi, F.; Senese, R.; Lanni, A.; Goglia, F.; de Lange, P.; Moreno, M. Defining the transcriptomic and proteomic profiles of rat ageing skeletal muscle by the use of a cDNA array, 2D- and Blue native-PAGE approach. *J. Proteomics* **2009**, *72* (4), 708–721.
- (61) Linton, P. J.; Dorshkind, K. Age-related changes in lymphocyte development and function. *Nat. Immunol.* **2004**, *5* (2), 133–139.
- (62) Paul, E.; Pozdnyakova, O. O.; Mitchell, E.; Carroll, M. C. Anti-DNA autoreactivity in C4-deficient mice. *Eur. J. Immunol.* **2002**, *32* (9), 2672–2679.
- (63) Gadjeva, M.; Verschoor, A.; Brockman, M. A.; Jezak, H.; Shen, L. M.; Knipe, D. M.; Carroll, M. C. Macrophage-derived complement component C4 can restore humoral immunity in C4-deficient mice. *J. Immunol.* **2002**, *169* (10), 5489–5495.
- (64) Ascenzi, P.; Fasano, M. Serum heme-albumin: an allosteric protein. *IUBMB Life* **2009**, *61* (12), 1118–1122.
- (65) Sierra, F.; Coeytaux, S.; Juillerat, M.; Ruffieux, C.; Gaudie, J.; Guigoz, Y. Serum T-kininogen levels increase two to four months before death. *J. Biol. Chem.* **1992**, *267* (15), 10665–10669.
- (66) Acuna-Castillo, C.; Leiva-Salcedo, E.; Gomez, C. R.; Perez, V.; Li, M.; Torres, C.; Walter, R.; Murasko, D. M.; Sierra, F. T-kininogen: a biomarker of aging in Fisher 344 rats with possible implications for the immune response. *J. Gerontol., Ser. A* **2006**, *61* (7), 641–649.
- (67) Walter, R.; Murasko, D. M.; Sierra, F. T-kininogen is a biomarker of senescence in rats. *Mech. Ageing Dev.* **1998**, *106* (1–2), 129–144.
- (68) Larsson, L. The age-related motor disability: underlying mechanisms in skeletal muscle at the motor unit, cellular and molecular level. *Acta Physiol. Scand.* **1998**, *163* (3), S27–S29.
- (69) Rostovtseva, T. K.; Bezrukov, S. M. VDAC regulation: role of cytosolic proteins and mitochondrial lipids. *J. Bioenerg. Biomembr.* **2008**, *40* (3), 163–170.
- (70) Carter, N. D.; Wistrand, P. J.; Isenberg, H.; Askmark, H.; Jeffery, S.; Hopkinson, D.; Edwards, Y. Induction of carbonic anhydrase III mRNA and protein by denervation of rat muscle. *Biochem. J.* **1988**, *256* (1), 147–152.
- (71) Larsson, L.; Salviati, G. Effects of age on calcium transport activity of sarcoplasmic reticulum in fast- and slow-twitch rat muscle fibres. *J. Physiol.* **1989**, *419*, 253–264.
- (72) Owens, E. L.; Lynch, C. J.; McCall, K. M.; Carter, N. D.; Vary, T. C. Altered expression of skeletal muscle proteins during sepsis. *Shock* **1994**, *2* (3), 171–178.
- (73) Steinghagen-Thiessen, E.; Hilz, H. The age-dependent decrease in creatine kinase and aldolase activities in human striated muscle is not caused by an accumulation of faulty proteins. *Mech. Ageing Dev.* **1976**, *5* (6), 447–457.
- (74) Nuss, J. E.; Amaning, J. K.; Bailey, E.; Doford, J. H.; Dimayuga, V. L.; Rabek, J. P.; Papaconstantinou, J. Oxidative modification and aggregation of creatine kinase from aged mouse skeletal muscle. *Ageing* **2009**, *1*, 6.
- (75) Marsh, D. R.; Criswell, D. S.; Hamilton, M. T.; Booth, F. W. Association of insulin-like growth factor mRNA expressions with muscle regeneration in young, adult, and old rats. *Am. J. Physiol.* **1997**, *273* (1 Pt 2), R353–R358.
- (76) Dirks, A. J.; Hofer, T.; Marzetti, E.; Pahor, M.; Leeuwenburgh, C. Mitochondrial DNA mutations, energy metabolism and apoptosis in aging muscle. *Ageing Res Rev* **2006**, *5* (2), 179–195.
- (77) Hagen, J. L.; Krause, D. J.; Baker, D. J.; Fu, M. H.; Tarnopolsky, M. A.; Hepple, R. T. Skeletal muscle aging in F344BN F1-hybrid rats: I. Mitochondrial dysfunction contributes to the age-associated reduction in VO₂max. *J. Gerontol., Ser. A* **2004**, *59* (11), 1099–1110.
- (78) Carter, E. E.; Thomas, M. M.; Muryinka, T.; Rowan, S. L.; Wright, K. J.; Huba, E.; Hepple, R. T. Slow twitch soleus muscle is not protected from sarcopenia in senescent rats. *Exp. Gerontol.* **2010**, *45* (9), 662–670.
- (79) Edstrom, E.; Ulfhake, B. Sarcopenia is not due to lack of regenerative drive in senescent skeletal muscle. *Ageing Cell* **2005**, *4* (2), 65–77.
- (80) Snow, L. M.; McLoon, L. K.; Thompson, L. V. Adult and developmental myosin heavy chain isoforms in soleus muscle of aging Fischer Brown Norway rat. *Anat. Rec., Part A* **2005**, *286* (1), 866–873.
- (81) Boffoli, D.; Scacco, S. C.; Vergari, R.; Solarino, G.; Santacroce, G.; Papa, S. Decline with age of the respiratory chain activity in human skeletal muscle. *Biochim. Biophys. Acta* **1994**, *1226* (1), 73–82.
- (82) Tobacman, L. S. Thin filament-mediated regulation of cardiac contraction. *Annu. Rev. Physiol.* **1996**, *58*, 447–481.
- (83) Szczesna, D.; Potter, J. D. The role of troponin in the Ca²⁺-regulation of skeletal muscle contraction. *Results Probl. Cell Differ.* **2002**, *36*, 171–190.
- (84) McArdle, A.; Vasilaki, A.; Jackson, M. Exercise and skeletal muscle ageing: cellular and molecular mechanisms. *Ageing Res. Rev.* **2002**, *1* (1), 79–93.
- (85) Sastre, J.; Pallardo, F. V.; Vina, J. The role of mitochondrial oxidative stress in aging. *Free Radical Biol. Med.* **2003**, *35* (1), 1–8.
- (86) Messina, S.; Vita, G. L.; Aguenouz, M.; Sframeli, M.; Romeo, S.; Rodolico, C.; Vita, G. Activation of NF-kappaB pathway in Duchenne muscular dystrophy: relation to age. *Acta Myol.* **2011**, *30* (1), 16–23.
- (87) Rossi, P.; Marzani, B.; Giardina, S.; Negro, M.; Marzatico, F. Human skeletal muscle aging and the oxidative system: cellular events. *Curr. Aging Sci.* **2008**, *1* (3), 182–191.
- (88) Emery, A. E. The muscular dystrophies. *Lancet* **2002**, *359* (9307), 687–695.
- (89) Oh-Ishi, S.; Kizaki, T.; Yamashita, H.; Nagata, N.; Suzuki, K.; Taniguchi, N.; Ohno, H. Alterations of superoxide dismutase isoenzyme activity, content, and mRNA expression with aging in rat skeletal muscle. *Mech. Ageing Dev.* **1995**, *84* (1), 65–76.
- (90) Sohal, R. S.; Wennberg-Kirch, E.; Jaiswal, K.; Kwong, L. K.; Forster, M. J. Effect of age and caloric restriction on bleomycin-chelatable and nonheme iron in different tissues of C57BL/6 mice. *Free Radical Biol. Med.* **1999**, *27* (3–4), 287–293.

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jjprot

Technical note

Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for Proteomics



F.V. Leprevost^a, D.B. Lima^a, J. Crestani^b, Y. Perez-Riverol^{c,d}, N. Zanchin^a,
V.C. Barbosa^e, P.C. Carvalho^{a,*}

^aLaboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

^bLaboratory for Regulation of Gene Expression in Microorganisms, Chemistry Institute, University of São Paulo, São Paulo, Brazil

^cDepartment of Proteomics, Center for Genetic Engineering and Biotechnology, Ave 31 e/158 y 190, Cubanacán, Playa, Ciudad de la Habana, Cuba

^dEMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

^eSystems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

ARTICLE INFO

Article history:

Received 20 April 2013

Accepted 13 June 2013

Available online 21 June 2013

Keywords:

Computational proteomics

Bioinformatics

Proteomics

Protein domains

Functional analysis

ABSTRACT

Mass-spectrometry-based shotgun proteomics has become a widespread technology for analyzing complex protein mixtures. Here we describe a new module integrated into PatternLab for Proteomics that allows the pinpointing of differentially expressed domains. This is accomplished by inferring functional domains through our cloud service, using HMMER3 and Pfam remotely, and then mapping the quantitation values into domains for downstream analysis. In all, spotting which functional domains are changing when comparing biological states serves as a complementary approach to facilitate the understanding of a system's biology. We exemplify the new module's use by reanalyzing a previously published MudPIT dataset of *Cryptococcus gattii* cultivated under iron-depleted and replete conditions. We show how the differential analysis of functional domains can facilitate the interpretation of proteomic data by providing further valuable insight.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

One of the goals of shotgun proteomics is to provide in-depth, holistic insights into cellular biology by first pinpointing differentially expressed proteins when comparing physiological states. Inferring exactly which proteins are in a mixture is an extremely challenging task, especially when analyzing data from higher-order organisms, in which case the number of peptides shared among proteins increases rapidly [1].

Typically, peptide spectrum matches (PSMs), the building blocks of computational shotgun proteomics, are mapped into protein groups that share identified peptides (Supplementary Fig. 1). In more complex scenarios, proteins share subsets of peptides, with different proteins giving rise to complex dependency graphs for treatment by protein-inference algorithms. The problem of deciding which proteins are truly in the mixture has been very well characterized by Nesvizhskii and collaborators [2,3], and computational solutions to tackle

* Corresponding author at: Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Rua Prof. Algacyr Munhoz Mader, 3775, ZIP: 81350-010, City: Curitiba, Paraná, Brazil. Tel.: +55 41 3316 3230; fax: +55 41 3316 3267.

E-mail address: paulo@pcarvalho.com (P.C. Carvalho).

it have been proposed [4,5]. It is now consensual in the proteomics community that a maximum-parsimony list of proteins is to be reported; i.e., one reports the smallest subset of proteins that explains all identified peptides. As a result, proteomic experiments cannot in general determine a mixture's correct protein contents. For example, a typical result for a *Homo sapiens* analysis might report somewhere near 2000 proteins according to the maximum-parsimony criterion but about 4000 when considering redundancies.

These limitations can, to some extent, obfuscate downstream functional analysis. In an attempt to help circumvent such difficulties, we introduce a data-analysis strategy stemming from the fact that proteins are composed of one or more regions that establish their biochemical functions. These "building blocks", known as functional domains, tend to be strongly conserved in nature. As such, inferring functional domains at a large scale and mapping peptide identifications into them, instead of into proteins, provides key benefits such as: a) simplifying (and eliminating redundancies in) the process of gaining functional insight into the biological system at hand; b) specifically addressing differentially expressed domains, and thus the key functional content of a given biological sample; c) providing direct access to the current functional state, thereby helping drive biochemical conclusions and culminating in an easier way to understand the relevant biochemical mechanisms.

PatternLab for Proteomics is a one-stop shop for proteomic data analysis, providing tools to handle data from mass spectra, conduct differential proteomics analyses, and more [6–8]. Within this environment, a cloud service has now been implemented that infers domains from the FASTA sequences of all proteins identified in the experiment at hand and maps peptide quantitation values into the corresponding functional domains. Recently, the growth of MS/MS data has motivated the proteomics community to seek cloud computing tools to enable small laboratories to analyze complex datasets [9,10]. Because inferring domains at a large scale is computationally intensive, resource demanding, and requires installing specialized software and databases, all this functionality is

already integrated into PatternLab following a cloud-client model. In essence, FASTA sequences of identified proteins are transmitted to our cloud servers, which perform domain inference by executing HMMER3 [11] on demand. The latter, briefly, uses a hidden Markov model (HMM) approach to scan profiles against the Protein Family (Pfam) database. Detailed instructions on how to use PatternLab are available [7,8]. The new option for performing a differential proteomic domain analysis (DPDA), or simply "differential dominomics", is integrated into the Regrouper module of the PatternLab pipeline and can be used by simply choosing to map values into domains instead of proteins. A brief overview of the pipeline is presented in Fig. 1.

We note that our cloud service, termed FioCloud, is hosted at the Fiocruz foundation (<http://fiocruz.br>), more specifically at Fiocruz Paraná. Fiocruz is one of the world's largest governmental agencies devoted to public health and research.

Our method significantly simplifies the process of understanding an organism's biology; therefore, some sensitivity loss may occur as a side effect. For example, for a group of proteins that share a common domain it is possible that while some are up-regulated others are down-regulated. Whenever this happens, dispersion is generated in the quantitation values and the proteins in question, consequently, are missed by our differential domain approach. Moreover, our method is blind to high-quality PSMs that do not map into any functional domain. On the other hand, we advocate that a differential domain prediction strategy can pave the way to a more effective analysis of organisms with poorly annotated genomes or when performing homology-driven proteomics [17], since protein domains tend to be more conserved than full-length protein sequences and are easier to predict [18]. That is, the method described herein is to be regarded as complementary to others, such as those based on homologous sequences. Note also that our method is blind to proteoforms, so these should be accounted for in standard differential proteomics analyses. Moreover, while a complementary strategy for inferring functional biology is by

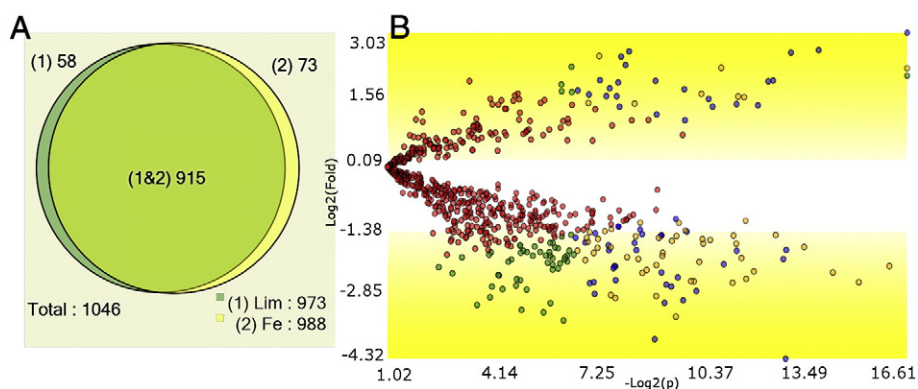


Fig. 1 – Workflow. Biological samples are analyzed by mass spectrometry. Mass spectra must be analyzed from a .sqt-compliant output [12], such as that generated by SEQUEST [13], ProLuCID [14], or the Spectrum Identification Machine (SIM) [15]. The PSMs are statistically filtered and organized using the Search Engine Processor (SEPro) [16]. The SEPro files are combined into PatternLab's index and sparse-matrix files for a single experiment using Regrouper, which offers the user the possibility of mapping quantitation values into protein functional domains. Differentially expressed domains are pinpointed using PatternLab's differential analyzer.



Fig. 2 – Differentially expressed domains. 58 and 73 domains were exclusively identified in the iron-depleted (Lim) and replete (Fe) conditions ($p < 0.05$), respectively, and 59 domains (blue dots) were found in both conditions but marked as differentially expressed ($q < 0.01$) by the TFold analysis.

mapping quantitation values into Gene Ontology (GO) terms (in fact, PatternLab includes specialized tools tailored to doing this [19]), usually an effective GO analysis can only be easily accomplished on well annotated organisms.

We exemplify the use of the new PatternLab functionality by reanalyzing the data of Crestani and collaborators, who have compared proteins from *Cryptococcus gattii* cultivated under iron-depleted and replete conditions [20]. The data were acquired using MudPIT [21] and searched with SEQUEST [13]. We then filtered for significant hits using the default parameters of SEPro [16] and used Regrouper for automatically inferring domains and mapping spectral counts into them by relying on HMMER3 and Pfam-A over the cloud, with Pfam accepting domains with HMMER E-Value $< 10E-6$ and i-EValue $< 10E-3$. Finally, PatternLab's statistical Venn Diagram [22] and TFold [23] modules were used for pinpointing domains exclusively identified in a single condition, as well as those differentially expressed (Fig. 2). Each domain out of a total of 1303 had at least one peptide mapped to it, and similarly none of 340 domains had any peptides mapped to it. Following domain identification, the information on which proteins were mapped to each domain is included in the domains' descriptions located in the index file. Downstream analysis in PatternLab can retrieve this information. We point out that the domain search results revealed interesting aspects that had remained unnoticed in the differential expression analysis. Examples are a FeoB_N domain, related to iron transportation and over-expressed in the iron-depleted condition, and three mitochondria-related domains (Mito_carr, FAD_binding_3, and Cyt-b5), over-expressed in the iron-replete condition. The complete results of all differentially expressed domains and all SEPro files discriminating our identified proteins and peptides are available at <http://proteomics.fiocruz.br/pcarvalho/dominomics>. A detailed protocol on how to use the new feature is available in Supplementary Part II.

2. Availability

All PatternLab modules are available for download at <http://proteomics.fiocruz.br>. All results of our *C. gattii* analysis are available at <http://proteomics.fiocruz.br/dpda>.

Financial support

The authors acknowledge CNPq, CAPES, FAPERJ, FAPESP, Fundação Araucária, and Fiocruz-PDTIS for the financial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.jprot.2013.06.013>.

REFERENCES

- [1] Perez-Riverol Y, Sanchez A, Ramos Y, Schmidt A, Muller M, Betancourt L, et al. In silico analysis of accurate proteomics, complemented by selective isolation of peptides. *J Proteomics* Sep 6 2011;74(10):2071–82.
- [2] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* Oct 2007;4(10):787–97.
- [3] Nesvizhskii AI, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov Today* Feb 15 2004;9(4):173–81.
- [4] Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res* Sep 2007;6(9):3549–57.
- [5] Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, et al. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* Aug 2009;8(8):3872–81.
- [6] Carvalho PC, Fischer JS, Chen EI, Yates III JR, Barbosa VC. PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics* 2008;9:316.
- [7] Carvalho PC, Yates Jr I, Barbosa VC. Analyzing shotgun proteomic data with PatternLab for proteomics. *Curr Protoc Bioinformatics* Jun 2010;13.13.1–5 [Chapter 13:Unit-13.13].
- [8] Carvalho PC, Fischer JS, Xu T, Yates III JR, Barbosa VC. PatternLab: from mass spectra to label-free differential shotgun proteomics. *Curr Protoc Bioinformatics* Dec 2012:13.19.1–13.19.18 [Chapter 13:Unit13.19].
- [9] Muth T, Peters J, Blackburn J, Rapp E, Martens L. ProteoCloud: A full-featured open source proteomics cloud computing pipeline. *J Proteomics* Aug 2013;88:104–8.

- [10] Trudgian DC, Mirzaei H. Cloud CFP: a shotgun proteomics data analysis pipeline using cloud and high performance computing. *J Proteome Res* Dec 7 2012;11(12):6282–90.
- [11] Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform* Oct 2009;23(1):205–11.
- [12] McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom* 2004;18(18):2162–8.
- [13] Eng JK, McCormack L, Yates A, Yates III JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–89.
- [14] Xu T, Venable JD, Park SK, Cociorva D, Lu B, Liao L, et al. ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol Cell Proteomics* 2006;5(S174).
- [15] Borges D, Perez-Riverol Y, Nogueira FC, Domont GB, Noda J, Leprevost FD, et al. Effectively addressing complex proteomic search spaces with peptide spectrum matching. *Bioinformatics* 2013;29(10):1343–4.
- [16] Carvalho PC, Fischer JS, Xu T, Cociorva D, Balbuena TS, Valente RH, et al. Search engine processor: filtering and organizing peptide spectrum matches. *Proteomics* Apr 2012;12(7):944–9.
- [17] Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, et al. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem* May 1 2001;73(9):1917–26.
- [18] Junqueira M, Carvalho PC. Tools and challenges for diversity-driven proteomics in Brazil. *Proteomics* Aug 2012;12(17):2601–6.
- [19] Carvalho PC, Fischer JS, Chen EI, Domont GB, Carvalho MG, Degraeve WM, et al. GO Explorer: a gene-ontology tool to aid in the interpretation of shotgun proteomics data. *Proteome Sci* 2009;7:6.
- [20] Crestani J, Carvalho PC, Han X, Seixas A, Broetto L, Fischer JS, et al. Proteomic profiling of the influence of iron availability on *Cryptococcus gattii*. *J Proteome Res* Jan 1 2012;11(1):189–205.
- [21] Washburn MP, Wolters D, Yates III JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* Mar 2001;19(3):242–7.
- [22] Carvalho PC, Fischer JS, Perales J, Yates JR, Barbosa VC, Bareinboim E. Analyzing marginal cases in differential shotgun proteomics. *Bioinformatics* Jan 15 2011;27(2):275–6.
- [23] Carvalho PC, Yates III JR, Barbosa VC. Improving the TFold test for differential shotgun proteomics. *Bioinformatics* Jun 15 2012;28(12):1652–4.