MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Programa de Pós-Graduação em Biologia Computacional e Sistemas

MODELAGEM CONCEITUAL DO SISTEMA DE BANCO DE DADOS PROTEINWORLDDB

MÁRCIA MÁRTYRES BEZERRA

Rio de Janeiro
Dezembro de 2012

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

MÁRCIA MÁRTYRES BEZERRA

Modelagem Conceitual do Sistema de Banco de Dados ProteinWorldDB

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Computacional e Sistemas

Orientador (es): Prof. Dr. Antonio Basílio de Miranda
Prof. Dr. Sérgio Lifschitz

RIO DE JANEIRO
Dezembro de 2012

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

B574 Bezerra, Márcia Mártyres

Modelagem conceitual do sistema de banco de dados ProteinWorldDB / Márcia Mártyres Bezerra. – Rio de Janeiro, 2012.

xi, 134 f. : il. ; 30 cm.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2012.

Bibliografia: f. 88-95

1. Banco de dados biológicos. 2. Modelagem conceitual de banco de dados. 3. Genômica comparativa. 4. Projeto comparação de genomas. 5. Sistema de banco de dados ProteinWorldDB. I. Título.

CDD 570.285

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

MÁRCIA MÁRTYRES BEZERRA

MODELAGEM CONCEITUAL DO SISTEMA DE BANCO DE DADOS PROTEINWORLDDB

ORIENTADOR (ES): Prof. Dr. Antonio Basílio de Miranda
Prof. Dr. Sérgio Lifschitz

Aprovada em: 20/dezembro/2012

EXAMINADORES:

Prof. Dr. Alberto Martín Rivera D Ávila (IOC/FIOCRUZ) **- Presidente**
Prof. Dr. Laurent Dardenne (LNCC/RJ)
Prof. Dr. Luiz Fernando Seibel (PUC/RJ)
Prof. Dr. André Nóbrega Pitaluga (IOC/FIOCRUZ)
Prof. Dr. Fabio Faria da Mota (IOC/FIOCRUZ)

Rio de Janeiro, 20 de dezembro de 2012

À

Guilherme, Daniel e Hélio

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

MODELAGEM CONCEITUAL DO SISTEMA DE BANCO DE DADOS PROTEINWORLDDB

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Márcia Mártyres Bezerra

Esta tese descreve o projeto conceitual do sistema de banco de dados ProteinWorldDB (PWDB). Um ponto importante da proposta do PWDB é permitir a construção de consultas e procedimentos no domínio da genômica comparativa sem a necessidade de comparação de sequências. Além disso, o PCG comparou milhões de sequências de proteína, incluindo o conjunto proteico total de centenas de genomas completos, utilizando programação dinâmica, e não um método heurístico, para os cálculos de similaridade. A estratégia do PCG, assim como a genômica, está fundamentada no conhecimento de que sequências biológicas por si só são pouco informativas; elas precisam ser analisadas a partir de um enfoque comparativo para a inferência de homologia. A comparação de sequências de diferentes organismos introduz uma perspectiva evolutiva ao processo, e o estudo comparativo de genomas completos pode ampliar a escala do conhecimento de um único processo biológico para o de sistemas biológicos complexos em células e organismos. Para responder eficientemente questões dessa natureza, o esquema conceitual apresentado associa bases de dados biológicos de referência aos índices de similaridade já pré-calculados e armazenados pelo PCG. Utilizando um formato gráfico de fácil compreensão para representar conceitos e relacionamentos (diagrama ER), o esquema foi proposto para facilitar o planejamento de consultas e procedimentos por pesquisadores da área de genômica (sem conhecimento de linguagens de bancos de dados), assim como guiar o desenvolvimento e a implementação física do PWDB por profissionais da área de computação. Alguns exemplos são apresentados com o objetivo de demonstrar a utilização do esquema conceitual para a especificação de consultas e procedimentos, mesmo antes da existência de um esquema lógico.

O esquema pode ser facilmente estendido. Módulos anexos podem ser inseridos/removidos para incluir outros projetos, baseados em comparação de sequências de proteína, que se beneficiem das informações fornecidas pelo módulo central do esquema e novas bases de dados, específicas de diferentes áreas (-ômicas, por exemplo), podem ser integradas ao esquema.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

CONCEPTUAL MODELING OF THE DATABASE SYSTEM PROTEINWORLDDB

ABSTRACT

PHD THESIS IN COMPUTATIONAL BIOLOGY AND SYSTEMS

Márcia Mártyres Bezerra

This thesis describes the conceptual design of the database system ProteinWorldDB (PWDB). An important point of the PWDB proposal is to allow the construction of queries and procedures in the field of comparative genomics without the need for sequence comparison. Moreover, the PCG compared millions of protein sequences, including the entire set of proteins from hundreds of complete genomes using dynamic programming, rather than a heuristic method, for calculating similarity

PCG's strategy, like that of genomic studies in general, is grounded in the knowledge that biological sequences alone are uninformative. They need to be analyzed from a comparative approach to infer homology. The comparison of sequences from different organisms introduces an evolutionary perspective to the process and the comparative study of complete genomes can expand our knowledge from a single biological process all the way to complex biological systems in cells and organisms. To efficiently answer questions of this nature, the conceptual schema links selected international reference biological databases to similarity indexes already precomputed and stored by the PCG. By using an easily understandable graphic format to represent concepts and relationships (ER diagram), the schema was proposed to help the design of queries and procedures by genomic researchers (who may not have knowledge of database languages) as well as to guide the development and physical implementation of the system by developers. Some examples are presented to demonstrate the use of the conceptual schema for specifying queries and procedures, even before the existence of a logical schema.

The schema can be easily extended. Additional modules can be inserted/removed to include other protein sequences comparisons projects that may benefit from the information provided by the schema´s central module. Likewise, new databases specific to different areas (-omics, for example) can be cross-referenced to the schema.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Esta tese discute a modelagem conceitual do sistema de banco de dados *ProteinWorldDB*[1] (PWDB), cujo conjunto de dados primário é o resultado de comparações "par-a-par" entre milhares de sequências de aminoácidos executadas pelo Projeto Comparação de Genomas[2] (PCG).

Como diferencial do PCG, pode-se citar:

- A utilização da capacidade ociosa de computadores pessoais de voluntários, através de uma infra-estrutura de computação distribuída, oferecida pelo *World Community Grid*[3] (WCG);
- O projeto possui um grande potencial multidisciplinar, permitindo a troca de experiência entre diferentes áreas de pesquisa;
- Foi o primeiro projeto da América Latina aceito para processamento pelo WCG;
- O programa de comparação de sequências utilizado foi o SSEARCH[4,5];
- O PCG gerou uma matriz de aproximadamente 1 Terabyte (TB), com 4,2 x 10^9 linhas. Cada linha exibe o resultado de similaridade entre um par de sequências de aminoácidos;
- Desta forma, índices de similaridade entre milhares de pares de sequências de aminoácidos já estão pré-calculados e armazenados nesta matriz, e podem ser recuperados sem a necessidade de uma nova comparação.

A estratégia do PCG está fundamentada no conhecimento de que sequências biológicas por si só são pouco informativas, elas precisam ser analisadas a partir de um enfoque comparativo, utilizando informações pré-existentes baseadas em relacionamentos e funções similares, para a inferência de homologia[6]. A comparação de sequências de diferentes organismos introduz uma perspectiva evolutiva ao processo[7].

[1] ANEXO I.
[2] ANEXO III.
[3] ANEXO III.
[4] O SSEARCH [http://fasta.bioch.virginia.edu/] é uma ferramenta, disponível gratuitamente, que busca o alinhamento local ótimo entre duas sequências, e utiliza o algoritmo *Smith-Waterman* [Smith and Waterman, 1981].
[5] Ao contrário de buscas baseadas em heurísticas, buscas ótimas garantem encontrar a melhor pontuação para o alinhamento, dado um determinado conjunto de parâmetros (http://www.ebi.ac.uk/services/proteins).
[6] Existência de relacionamento evolutivo. Ver DISCUSSÃO.
[7] Theodosius Dobzhansky's: "*Nothing in biology makes sense except in the light of evolution*" (The American Biology Teacher, March 1973 (35:125-129) (http://www.pbs.org/wgbh/evolution/library/10/2/l_102_01.html).

Um ponto relevante dessa proposta refere-se ao fato de que o foco tradicional da genética e biologia molecular esteve direcionado, ao longo dos anos, para o entendimento da função de um gene importante num processo biológico específico, enquanto que a área mais recente da genômica tem seu foco no conjunto completo de genes de um organismo (além de outras regiões estruturais e reguladoras). Sob essa nova perspectiva, o estudo comparativo de genomas completos pode ampliar o conhecimento de um único processo biológico para o de sistemas biológicos complexos em células e organismos.

A análise comparativa de sequências é parte da rotina da pesquisa biológica atual; o tempo dispendido no processo é considerável e o procedimento utilizado para as comparações é computacionalmente intensivo. Além disso, essas análises costumam ser apenas a primeira etapa de uma série de procedimentos mais elaborados, fornecendo os primeiros indícios para definir os próximos passos em direção a novas descobertas e inferências. Muitas vezes, comparações das mesmas sequências são repetidas inúmeras vezes.

O PCG comparou o conjunto proteico total de centenas de genomas[8] e, desta forma, permitiu que a etapa inicial de comparação de sequências fosse ultrapassada. Mais claramente, o PCG armazenou os índices de similaridade[9] obtidos a partir das comparações para serem reutilizados, abolindo a necessidade da repetição de uma mesma comparação.

O conjunto de requisitos do PCG[10] expande-se por um domínio bastante amplo da genômica, e a importância do projeto está diretamente ligada à forma como seus resultados serão tratados. Neste contexto, existe a necessidade de um sistema muito bem elaborado para armazenar e gerenciar os resultados eficientemente, de forma a atender estes requisitos.

Um sistema de banco de dados deve apresentar um projeto que vise à organização das informações e à utilização de técnicas para que o sistema apresente boa *performance* e facilidade de manutenção. A organização do banco de dados é uma das etapas mais importantes, e seu desenho deve ser representado, primeiramente, em um esquema conceitual, que é uma representação gráfica do modelo conceitual.

[8] DISCUSSÃO e ANEXO III.
[9] Validados estatisticamente pelo programa de comparação de sequências – ANEXO III.
[10] ANEXO III.

GENÔMICA, BIOINFORMÁTICA E BANCOS DE DADOS

Da Genômica à Biologia de Sistemas

Inicialmente, as bases de dados ganharam proeminência na biologia molecular como repositórios centrais de dados gerados pelos projetos de sequenciamento em larga escala. Com o crescimento de dados experimentais de diferentes áreas das ciências da vida devido aos avanços de tecnologias de alto rendimento, além das bases de dados primárias de sequências houve um grande aumento de bancos de dados de outras disciplinas biológicas para armazenar diferentes tipos de dados moleculares (FIGURA 1.1). Hoje, não mais a geração, mas sim a capacidade de processar, gerenciar e analisar esses dados, assim como interpretar as informações resultantes tornou-se um grande desafio para o avanço das ciências da vida [Katari *et al.*, 2010] – vistas como *discovery sciences* [Ideker *et al.*, 2001].

Esta grande produção de dados, diária, requer soluções de gerenciamento mais sofisticadas, e a disponibilidade da internet como uma moderna estrutura para trocas científicas tem gerado novas demandas com relação à acessibilidade dos dados. Além disso, o relativamente novo campo da Biologia de Sistemas tem aumentado ainda mais a demanda de requerimentos dos bancos de dados biológicos. A visão geral da biologia de sistemas é ir além da era dos estudos reducionistas de partes isoladas de interesse – por exemplo, proteínas e genes considerados individualmente – e atingir um conhecimento de estruturas mais complexas e sua dinâmica, como redes reguladoras, células, órgãos e, em última análise, o entendimento da biologia de todo o organismo como um sistema.

Uma grande quantidade dos dados produzidos atualmente ainda provém de projetos de sequenciamento de genomas, e irão continuar impulsionados pela acentuada queda no preço das novas tecnologias de sequenciamento e pela busca contínua de maior conhecimento sobre a vida, organismos, relações evolutivas e sistemas biológicos, adaptados e diferenciados pelo ambiente. No entanto, uma sequência por si só não é informativa; para o desenvolvimento de novas hipóteses é necessário analisar sua(s) função(ões) e relacionamento(s) com outras sequências. Atualmente, uma parte substancial da rotina diária da pesquisa biológica ainda é dispendida na análise de sequências; esse tipo de análise pode ser o passo inicial para a descoberta de novas conexões e regras biológicas importantes para um maior entendimento de sistemas complexos [Allen G, 2006].

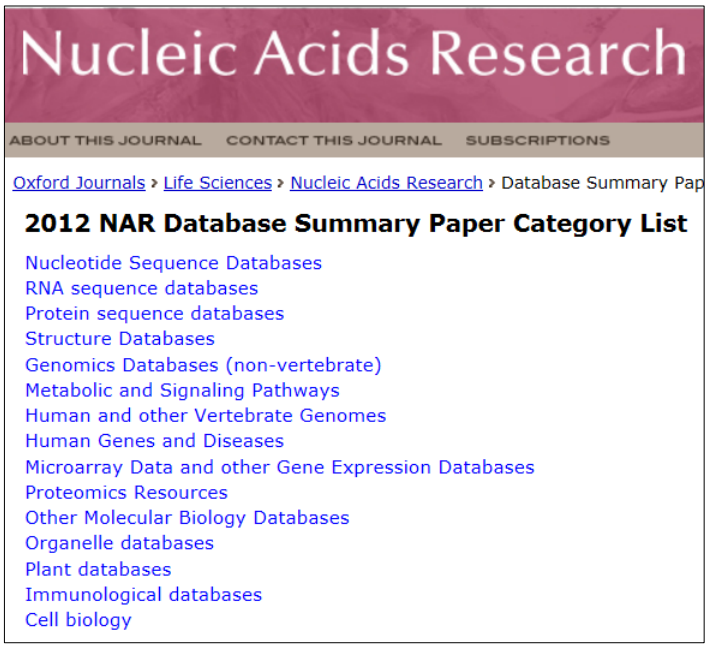


FIGURA 1.1. NAR *Database issue* – Vol. 40, January 2012: inclui 1.380 bancos de dados distribuídos em 14 categorias e 41 subcategorias. A edição de 2012 lista 92 novos bancos de dados *online* e 77 artigos de atualização de bancos de dados já descritos em edições anteriores.

Um problema que atinge todas as áreas das ciências biológicas é o da confiabilidade de dados; em inúmeras situações não existe garantia de que os dados armazenados nos bancos de dados sejam "biologicamente" corretos. Uma fonte comum de problemas é a propagação de erro em anotações de sequências através do uso de mecanismos automatizados de anotação [Philippi and Köhler, 2006]. Além disso, o erro humano é outra frequente fonte de enganos; apesar do procedimento manual de curar dados para assegurar a qualidade das bases de dados ser importante, existe um alto potencial para a introdução de erros. Anotações incorretas, se inseridas nos bancos de dados, multiplicam-se em subsequentes adições e podem se acumular numa proporção preocupante [Bell *et al.*, 2012].

Um ponto importante para uma interpretação mais holística das ciências biológicas, reconhecido na literatura científica, é que os blocos construtores básicos para a elaboração de modelos de sistemas biológicos são os dados experimentais existentes, que estão armazenados em milhares de bancos de dados distintos [Pennisi E, 2005; Roos DS, 2001; Augen J, 2001]. Neste contexto, temas centrais como uma correta modelagem e integração de dados, e a correção de anotações duvidosas são pré-requisitos fundamentais para qualquer estudo desde a genômica até a biologia de sistemas.

Além disso, vários tipos de informação costumam estar ausentes, incluindo a anotação funcional de genes e proteínas, relações genótipo-fenótipo e informações detalhadas de vias bioquímicas, dentre outros. Mesmo quando a informação baseada em similaridade de sequências é considerada, apenas 50% de todos os passos das reações de vias metabólicas podem ser conectadas a genes e proteínas que as catalisam [Karp *et al.*, 2005]. Consequentemente, a parametrização de modelos de sistemas biológicos pode ser considerada bastante difícil, interferindo na confiança e potencial preditivo dos resultados.

Muito dos atuais esforços em biologia de sistemas procura integrar os resultados das diferentes tecnologias científicas atuais. Uma das grandes dificuldades reside em converter dados em informação que forneça *insights* e represente conhecimento. O processo inicial requer limpeza e coerência de dados, tendo em mente que transformar informação em conhecimento requer interpretar o real significado dos dados [Brenner, 2003].

Desta forma, para lidar com a complexidade dos dados acumulados, extrair conhecimento do comportamento celular subjacente e eventualmente construir modelos preditivos, é necessário um amplo espectro de ferramentas computacionais. Mesmo assim, representações computacionais podem não ser suficientes devido a limitações computacionais, experimentais e metodológicas. [Mendes *et al.* 2004; Sokhansanj *et al.*, 2005] acreditam que o aumento da qualidade e coerência de dados, disponibilidade de bancos de dados integrados, e abordagens que possam gerenciar variabilidade experimental deverão ser cada vez mais necessários para uma maior confiabilidade de representações *in silico*.

Existe uma longa tradição de pesquisas sobre integração de bancos de dados na ciência da computação. Porém, apesar de já ser reconhecida há muitos anos como uma tecnologia chave nas ciências biológicas [Stein LD, 2003], esta ainda é uma área que requer constantes pesquisas devido à natureza da informação existente nas bases de dados biológicos: heterogeneidade, distribuição, tamanho, necessidade de frequentes atualizações e pobreza de semântica [Philippi S, 2004; Philippi and Köhler, 2006; Iskandar and Naomie, 2006].

Bioinformática

Partindo da análise de sequências biológicas, a tecnologia de coleta de dados genômicos foi sendo incrementada, assim como a geração de resultados oriundos das diversas ciências "–ômicas" (TABELA 1.1). Com isso, a necessidade de novos métodos para o gerenciamento e análise dessa massiva quantidade de dados foi ampliada, e o termo bioinformática evoluiu para incluir a análise matemática, estatística e computacional dos dados genômicos e demais ciências "–ômicas".

A bioinformática é um campo multidisciplinar, estende-se desde a modelagem de bancos de dados à engenharia de sistemas, inteligência artificial, matemática e estatística aplicada, com um foco mais direcionado para a ciência genômica. Hoje, a disciplina compartilha grande parte de seu domínio com a biologia computacional, sendo que esta última tem o foco mais direcionado ao desenvolvimento de modelos matemáticos para a simulação de sistemas biológicos. Neste contexto, é bom frisar que a utilidade de um modelo é muitas vezes influenciada pela qualidade dos dados experimentais e mecanismos subjacentes, fato que evidencia a importância do gerenciamento eficiente e a correta interpretação dos dados biológicos que vêm sendo produzidos.

TABELA 1.1 Exemplo de algumas ciências "–ômicas".

GENÔMICA	Envolve o estudo de genes e da sua função. A Genômica visa compreender a estrutura do genoma, incluindo mapeamento de genes, sequenciamento de DNA, e explorar os mecanismos moleculares e da interação de fatores genéticos e ambientais em organismos.
TRANSCRIPTÔMICA	É o estudo dos transcriptomas, o conjunto completo de transcritos de RNA produzido pelo genoma de uma só vez. É especificamente focada em como os padrões de transcrição são afetados por doenças, pelo desenvolvimento, ou fatores ambientais, tais como hormônios, drogas, etc.
PROTEÔMICA	É o estudo em larga escala de proteínas em sistemas biológicos. O proteoma é a totalidade dos componentes de proteínas, incluindo as modificações feitas a um conjunto específico de proteínas, produzidas por um organismo ou sistema. Este proteoma pode variar com o tempo e sofrem modificações sob condições experimentais diferentes ou *stress*, que uma célula ou organismo sofre.
METABOLÔMICA	É focada em perfis e na quantificação de pequenos compostos que ocorrem naturalmente e que coletivamente constituem o assim chamado metaboloma. Pequenas moléculas servem como assinaturas diretas de atividade bioquímica e, portanto, são mais fáceis de serem correlacionadas com fenótipos.

A utilização de softwares para o alinhamento e comparação de sequências biológicas[11] (FIGURA 1.2) para a identificação e mensuração da similaridade entre elas é um procedimento básico e fundamental em bioinformática. Trata-se de uma metodologia amplamente utilizada com o objetivo de adquirir informações sobre genes e proteínas desconhecidos, por exemplo, baseando-se em sequências já conhecidas e caracterizadas. O grau de similaridade[12], avaliado estatisticamente pelos programas de comparação de sequências[13], pode sugerir homologia (existência de relacionamento evolutivo) entre elas.

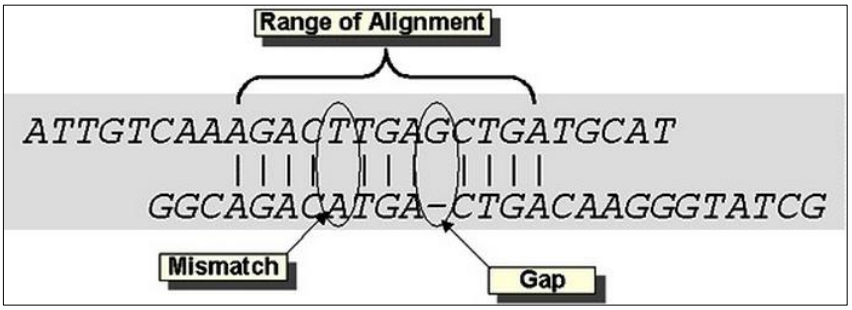


FIGURA 1.2. Alinhamento e comparação "par-a-par' de duas sequências de DNA (BLAST Glossary, http://www.ncbi.nlm.nih.gov/books/NBK62051/).

Os atuais programas para comparação de sequências utilizam diferentes métodos que fornecem resultados satisfatórios na maioria dos casos, mas também possuem seus prós e contras. Devem ser escolhidos de acordo com o propósito do experimento, do tamanho do conjunto de dados a ser comparado e da estrutura computacional disponível, sem abrir mão da precisão dos resultados. Em projetos de larga escala (por exemplo, comparação entre pares de todas as sequências de proteína de *n* genomas – "par-a-par" e "todas-contra-todas"[14]), o tempo de processamento e espaço computacional terá um grande peso.

[11] Para uma comparação "resíduo-a-resíduo", duas ou mais sequências são alinhadas. No alinhamento, as posições relativas das sequências são ajustadas para otimizar (normalmente maximizar) a pontuação do alinhamento – de acordo com alguma matriz de pontuação de referência. Em alguns casos, alguns espaços podem ser inseridos, com penalidades associadas, em uma ou mais sequências para otimizar esta pontuação. BLAST Glossary, http://www.ncbi.nlm.nih.gov/books/NBK62051/

[12] A similaridade entre duas sequências pode ser expressa como o percentual de resíduos idênticos e/ou percentual de substituições positivas entre elas (normalmente para aminoácidos).

[13] Por exemplo, BLAST (*Basic Local Alignment Search Tool*, [Altschul *et al.*, 1990, 1997]) e SSEARCH.

[14] A cada execução do programa de comparação, um par de sequências é alinhado e comparado; e todas as sequências são pareadas e comparadas (pelo menos) uma vez.

A etapa de comparação de sequências visando à identificação de homologia normalmente é apenas uma das primeiras fases de procedimentos computacionais automatizados, em que resultados gerados por um processo são utilizados como entrada para um processo posterior. Deste modo fica clara a importância da qualidade dos resultados gerados nas etapas iniciais. A possibilidade de não identificar (ou identificar incorretamente) um resultado inicial pode comprometer a identificação de algum outro padrão importante numa fase futura.

A escolha do algoritmo utilizado para as comparações de sequências certamente refletirá nos resultados das buscas de similaridade em bancos de dados e na identificação de homologias. Mas é preciso lembrar que o algoritmo utilizado não é o único parâmetro para medida de acurácia e precisão[15] dos resultados. A matriz de substituição[16] selecionada, assim como os valores escolhidos para a contabilidade de abertura e extensão de *gaps*[17], tem grande influência no resultado de qualquer método de comparação de sequências de nucleotídeos ou aminoácidos.

As vantagens e desvantagens de dois dos mais utilizados algoritmos de comparação de sequências, *BLAST* [Altschul *et al.*, 1990, 1997] e *Smith-Waterman* (SW) [Smith and Waterman, 1981], foram discutidas em [Shpaer *et al.*, 1996]. Estes autores mostraram diferenças entre os dois métodos com relação à acurácia e velocidade nas comparações com bases de dados. O algoritmo SW utiliza o método de programação dinâmica[18], e encontra o

[15]Exatidão ou Acurácia é o grau de concordância entre o resultado de uma medição e o valor verdadeiro do mensurando (grandeza específica submetida à medição). Precisão é um conceito qualitativo para indicar o grau de concordância entre os diversos resultados experimentais obtidos em condições de repetitividade. Assim, boa precisão significa erro estatístico pequeno, de forma que os resultados apresentam boa repetitividade. Note, entretanto, que mesmo com boa precisão a exatidão ou acurácia pode ser ruim caso exista erro sistemático grande (Introdução à teoria de erros, Instituto Tecnológico de Aeronáutica, http://www.fis.ita.br/labfis24/erros/errostextos/teor_erros1.htm, em 24/10/2013).

[16]Uma matriz de substituição contém valores proporcionais à probabilidade de que o aminoácido i seja transformado no aminoácido j, para todos os pares de aminoácidos. Tais matrizes são construídas através da montagem de uma grande e diversa amostra de alinhamentos entre pares de aminoácidos. Se a amostra é suficientemente grande para ser estatisticamente significativa, as matrizes resultantes devem refletir as verdadeiras probabilidades de mutações que ocorrem através de um período de evolução (NCBI – Glossário, http://www.ncbi.nlm.nih.gov/books/NBK21106/).

[17] Um espaço introduzido num alinhamento para compensar inserções e deleções em uma das sequências em relação à outra. Para prevenir o acúmulo de muitos espaços num alinhamento, impõe-se uma penalidade fixa na pontuação do alinhamento para a introdução do primeiro espaço, e a extensão do *gap* para englobar nucleotídeos ou aminoácidos adicionais é penalizado de acordo com o número de espaços introduzidos (NCBI – Glossário, http://www.ncbi.nlm.nih.gov/books/NBK21106/).

[18] Programação dinâmica funciona basicamente construindo-se uma tabela para armazenar resultados intermediários que são utilizados posteriormente para obter o resultado final. Para demonstrar a idéia geral, suponha o problema de ir do ponto A até o ponto E usando o caminho mais curto. No caso, o ponto C encontra-se entre A e E. Se o menor caminho de A até C é conhecido, então basta que se calcule como ir de C até E. Utilizando programação dinâmica pode-se armazenar numa tabela a solução A-C e usá-la posteriormente para decidir o melhor caminho de A até E. http://www.ncbi.nlm.nih.gov/books/NBK6831/, capítulo A05.

melhor[19] alinhamento local entre pares de sequências. Já o algoritmo BLAST identifica alinhamentos locais ótimos entre pares de sequências utilizando uma heurística[20] que procura inicialmente pequenos "pareamentos" (*words*), e só então estende os alinhamentos a partir dessas *words*. Segundo [Altschul *et al.*, 1990], este algoritmo é um desdobramento do algoritmo de SW, sendo um modelo otimizado em velocidade, ao contrário do cálculo mais acurado e exato do SW que, muitas vezes, torna-se inviável em projetos de larga escala devido ao consumo de tempo e espaço computacional.

[Uchiyama, 2007, Shpaer *et al.*, 1996; Pearson 1991, 1995] discutem a eficiência e aplicabilidade destes dois métodos. Os resultados mostram que o número de falso-positivos, assim como o de falso-negativos, é significativamente menor para o SW e que o risco de uma sequência facilmente detectada pelo SW não ser identificada pelo BLAST é considerável. De forma resumida, o algoritmo de SW, computacionalmente falando, garante encontrar o melhor alinhamento local entre as sequências, enquanto que o BLAST encontra alinhamentos ótimos mas não garante encontrar o melhor. De qualquer forma o BLAST fornece resultados bastante aceitáveis na maioria dos casos, sendo, de longe, o algoritmo mais utilizado na área.

A utilização do algoritmo de SW pode ser viável para comparações em larga escalase existir uma estrutura computacional com velocidade de processamento suficiente, e provavelmente distribuída[21]. Porém, pequenas instituições ou laboratórios podem não ter acesso a esse tipo de recurso; nestes casos, certamente a opção será por um algoritmo como o Blast.

Um fato importante a ser lembrado é que a existência de falso-positivos, com um valor de E-value[22] <= 0,01, é esperada quando milhões de comparações são executadas. De qualquer forma, a transferência de função e inferência de homologia não deve se basear unicamente nos valores de E-value. A fração de posições idênticas (ou posições positivas, no caso de

[19] Com melhor pontuação.
[20] Heurística é um método ou processo criado com o objetivo de encontrar soluções para um problema. É um procedimento simplificador (embora não simplista) que envolve a substituição de questões difíceis por outras de resolução mais fácil a fim de encontrar respostas viáveis, ainda que não ótimas. O procedimento pode ser tanto uma técnica de resolução de problemas, como uma operação de comportamento automática, intuitiva e inconsciente.
[21] Existem implementações do SW otimizadas em nível de hardware e em processamentos distribuídos em *grid* e *cloud*, por exemplo [ANEXO III, Liu *et al.*, 2013; Torbjørn R, 2011; Khajeh *et al.*, 2010; Rognes T, 2010; Liu *et al.*, 2009; Rudnicki *et al.*, 2009; Manavski andValle, 2008; Li *et al.*, 2007; Farrar M, 2007].
[22] O valor esperado representa o número de diferentes alinhamentos, com pontuação igual ou superior a um valor S, que é esperado ocorrer ao acaso numa busca de banco de dados. Quanto menor o E-value, mais significativa é a pontuação do alinhamento (NCBI – BLAST Glossary: http://www.ncbi.nlm.nih.gov/books/NBK62051/).

aminoácidos) entre um par de sequências assim como a extensão da área de sobreposição, dentre outras propriedades do alinhamento, têm um importante papel nas predições funcionais e evolutivas baseadas em similaridade de sequências [Rost B, 2002; Tian and Skolnick, 2003; Boekhorst and Snel, 2007].

Bancos de Dados Biológicos

Um banco de dados pode ser definido como qualquer coleção de dados relacionados e gerenciados por um sistema particular, chamado de SGBD – Sistema Gerenciador de Bancos de Dados. De uma forma mais restritiva, pode-se dizer que um banco de dados é uma coleção de dados persistente, logicamente coerente e inerentemente significativa, com relação a alguns aspectos do mundo real [Elmasri and Navathe, 2011]. A atividade de preparar um banco de dados pode ser dividida em:

- Coleta e organização d dados para que possam ser facilmente acessados;
- Disponibilização desses dados em um sistema multiusuário.

Os autores acima citam como relevantes algumas características de bancos de dados biológicos, como por exemplo:

- As definições dos dados biológicos devem ser passíveis de representação numa subestrutura de dados que garanta que informações importantes não sejam perdidas durante a modelagem dos dados;
- Devem ser flexíveis ao lidar com tipos e valores de dados. A imposição de restrições deve ser limitada, dentro do possível, uma vez que isso pode excluir exceções. A exclusão desses valores pode resultar em perda de informação relevante;
- Precisam dar suporte a consultas complexas. No entanto, usuários sem o conhecimento da estrutura de dados podem não conseguir construir, por conta própria, uma consulta complexa. Deste modo, o sistema deve fornecer ferramentas para que se construam tais consultas.

Com o amadurecimento da pesquisa genômica, além de dados de sequências, uma grande quantidade de outros dados biológicos tem sido gerada e armazenada em inúmeros bancos de dados. A informação biológica atual reside em algumas centenas de bancos de dados, públicos e privados, que provêem informações descritivas genômicas, proteômicas,

enzimáticas, de expressão gênica, variantes genéticas e ontologias, para citar algumas, suplementadas por múltiplas publicações científicas.

No entanto, um problema importante é conseguir relacionar uma mesma entidade biológica (no "mundo real") em diferentes bases de dados. Fontes de dados distintas geralmente usam identificadores particulares que em alguns casos podem ser conectados através de fontes de mapeamentos de identificadores disponíveis na *web*, mas muitos bancos de dados não provêem números de acesso únicos, e nos casos que possuem, muitas vezes os números não são estáveis nem em diferentes versões do mesmo banco de dados [Philippi and Köhler, 2006]. Como consequência, em muitos casos, é impossível reproduzir resultados de buscas acuradamente.

Bancos de dados têm sido utilizados para gerenciar e integrar grandes quantidades de dados complexos em outras disciplinas por décadas, e servem para dar suporte a métodos de análise provendo uma estrutura para integrar informação de uma variedade de fontes, permitindo buscas mais rápidas e poderosas. [Nelson *et al.*, 2003] fazem um comentário sobre o descuido com os princípios de modelagem de bancos de dados biológicos ser justificado devido à complexidade dos dados; os autores defendem que projetar um banco de dados ignorando todo o conhecimento já acumulado na área de computação é similar a projetar um experimento de biologia molecular ignorando os princípios fundamentais da replicação do DNA. E mesmo que se considere avanços na área de tecnologia de informação, ferramentas para aplicações de sistemas de bancos de dados não terão sucesso se forem implementadas a partir de uma base de dados ineficiente devido a um projeto de modelagem deficiente.

Alguns autores com formação em engenharia de software consideram que ferramentas de seu domínio vêm sendo aplicadas à genômica visando, primariamente, o desenvolvimento de algoritmos de análises poderosos e eficientes, mas sem uma atuação mais forte no desenvolvimento de sistemas de informação de qualidade [Mayordomo AM, 2011].

Integração de Dados

Devido ao número crescente de bancos de dados de biologia molecular e do seu conteúdo, a integração de bancos de dados nesse domínio é um tema importante de pesquisa.

Abordagens existentes têm em comum a compreensão de que são necessários esforços consideráveis para fornecer acesso integrado a fontes de dados heterogêneas [Philippi S, 2004].

Para [Pennisi E, 2005], um equívoco comum é a crença de que os principais problemas de integração de bancos de dados biológicos estão relacionados com a tecnologia utilizada. O autor acredita que, apesar do domínio de tais tecnologias poder ser um desafio, os maiores problemas estão, na realidade, relacionados a estrutura e conteúdo dos próprios bancos de dados que impedem o uso efetivo de tecnologias de integração. São problemas que não apenas possuem efeitos adversos relativos à tarefa de garantir a disponibilidade de dados para a comunidade científica em geral, mas tornam-se um obstáculo ainda maior para a biologia de sistemas [Philippi and Köhler, 2006].

Para a construção de modelos na área de biologia de sistemas o primeiro passo é a identificação de fontes de dados adequadas. [Philippi and Köhler, 2006] citam como um pré-requisito para identificar e usar dados, uma descrição, com meta-informação apropriada, ao menos do tipo de dado armazenado, o modo como foram produzidos, diretrizes sobre como foram curados, as estruturas usadas para armazenar os dados e informação sobre o gerenciamento de atualizações e versões. E concordam que, infelizmente, nem todo banco de dados biológicos fornece tal meta-informação.

Os repositórios de informação genômica são bastante heterogêneos, muitas vezes usam conceitos distintos ou versões do mesmo termo. Por exemplo, uma alteração numa sequência de DNA pode ser uma variação, mutação, polimorfismo, SNP. Nesses casos, apesar desses termos não representarem exatamente um mesmo conceito, podem ser usados com o mesmo significado em algumas situações, e isso causa confusão quando o dado precisa ser interpretado [Den Dunnen and Antonarakis, 2001]. Como alguns estudos indicam [Richesson and Turley, 2003], esse é um problema para o qual a utilização de metodologias de modelagem conceitual e ontologias são essenciais. Como auxílio a esse tipo de problema, ontologias e vocabulários controlados são frequentemente utilizados nas ciências biológicas como referências semânticas, que possuem definições comuns para entidades do mundo real (conceitos) e as relações entre elas.

Referências semânticas muitas vezes são utilizadas para codificar campos em bancos de dados – como por exemplo, "Taxonomy Ids" para espécies no NCBI[23] ou "números EC"[24] para funções enzimáticas – ao invés de serem criadas manualmente, sujeitas a erros de escrita e descrições livres. E podem ser usadas também para integração semântica de dados [Stevens *et al.*, 2000; Köhler *et al.*, 2003; Ashburner *et al.*, 2000; Philippi and Köhler, 2004]. Por exemplo, se duas fontes de dados armazenam dados sobre proteínas e na estrutura de um dos bancos os dados estão nomeados como "dados_ptn" e no outro como "dados_p", uma referência de ambos, semanticamente definida pelo conceito ontológico "proteína", pode ser explorada para conectar as entradas entre as duas fontes de dados, apesar das diferenças ao nível de esquemas.

Modelagem Conceitual

Na ciência da computação, modelagem conceitual é um campo heterogêneo, que engloba várias disciplinas relacionadas à construção de sistemas de software. Na área de bancos de dados o termo começou a ser utilizado referindo-se a representação de dados e suas inter-relações, as quais seriam gerenciadas por um sistema de informação, independentemente de qualquer característica de implementação deste sistema [Chen PP, 1976; Chen *et al.*, 1999].

Como os sistemas de software vêm se tornando mais complexos, e o domínio dos problemas tem se movido para além de conceitos familiares aos desenvolvedores, modelos conceituais vêm ganhando importância, atuando como o ponto inicial para o entendimento dos problemas de usuários, auxiliando na sua resolução. Deste modo, os modelos conceituais devem empregar uma linguagem que possa ser compreendida pelos usuários para o propósito de validação, e também transmitir informações do problema necessárias aos desenvolvedores para que estes possam construir o sistema posteriormente. O escopo do termo foi sendo ampliado gradativamente e adquirindo o significado de representação do domínio do problema, com o propósito de compreensão e comunicação entre desenvolvedores e usuários [Kaindls and Carroll, 1999; Loucopoulos and Karakostas, 1995; Wieringa R, 1995; Beringer D, 1994].

[23] NCBI - http://www.ncbi.nlm.nih.gov/taxonomy.
[24] Número atribuído a um tipo de enzima de acordo com um esquema de nomenclatura padronizado pelo *Enzyme Comission* do Comitê de Nomenclatura da União Internacional de Bioquímica e Biologia Molecular (IUBMB) - http://www.chem.qmul.ac.uk/iubmb/enzyme/.

Independente do rigor de significado nas diferentes áreas da computação, as características essenciais de qualquer modelo conceitual são representação e compreensão, isto é, permitir que desenvolvedores entendam o problema levantado por usuários, para atingir uma concordância com estes sobre o escopo da solução e, finalmente, usar a informação representada no esquema conceitual para construir um sistema de software que resolva o problema em questão (seja este banco de dados convencional, baseado em conhecimento, ou qualquer outro).

Resumidamente, de acordo com [Juristo and Moreno, 2000], modelos conceituais satisfazem as seguintes condições:

1. Independência entre a análise e as fases subsequentes, isto é, a possibilidade de análise e compreensão dos vários aspectos do problema antes de escolher o software e paradigmas de desenvolvimento.
2. Independência do sistema de computação, isto é, a possibilidade de usar resultados das análises como base para desenvolvimento em diferentes paradigmas.
3. Independência de evolução de tecnologia, isto é, a possibilidade do mesmo esquema ser válido mesmo quando os softwares são atualizados ou substituídos.

Uma vantagem de trabalhar no nível conceitual é que este é o mais estável. Não é afetado por alterações em interfaces de usuários ou técnicas de armazenamento ou acesso ao banco de dados. Se, por exemplo, um esquema conceitual é implementado em um SGBD relacional e, posteriormente, deseja-se fazer uma migração para um SGBD orientado a objeto, a não ser que o "Universo de Discurso" (UoD)[25] tenha se alterado, o esquema conceitual pode ser mantido sem nenhuma alteração. Será necessário apenas aplicar um diferente processo de mapeamento para os esquemas lógico e físico, e migrar os dados.

A modelagem conceitual e descrições ontológicas são amplamente utilizadas no campo de sistemas de informação auxiliando os desenvolvedores a realizarem seu trabalho em um alto nível de abstração, permitindo a compreensão e descrição do domínio do problema antes de atuar concretamente na sua solução [Chen *et al.*,1999].

[25] A porção do mundo real relevante para o banco de dados; às vezes referenciado também como mini-mundo (*miniworld*).

Na literatura, o termo modelagem conceitual muitas vezes é utilizado de forma pouco precisa, e não é raro encontrar modelos, ditos conceituais, que na realidade já estão condicionados a restrições computacionais próprias da abordagem de um desenvolvedor em particular[26]. Nestes casos, pode-se considerar que o enfoque do sistema de desenvolvimento de software estaria pré-condicionado desde o início de sua construção e, desta forma, o método de desenvolvimento estaria sendo escolhido antes que as necessidades do usuário fossem entendidas.

Pode-se citar algumas publicações que tratam especificamente de modelagem de bancos de dados biológicos: [Pastor O *et al.*, 2012; Mayordomo AM, 2011; Busch and Wedemann, 2009; Pastor O, 2008; Elmasri *et al.*, 2007; Xiaohua Zhou and Il-Yeol Song, 2005; Birney and Clamp, 2004; Chen and Carlis, 2003; Nelson *et al.*, 2003; Keet CM, 2003; Rojas-Mujica and Bornberg-Bauer, 2002; Bornberg-Bauer and Paton, 2002; Rubin *et al.*, 2002; Paton *et al.*, 2000; Juristo and Moreno, 2000; Navathe and Kogelnik, 1999; Chen *et al.*, 1999]. Dentre estes, a maioria discute conceitos e a importância do tema e sua ampla aplicação na área de sistemas de informação, e relata a pouca orientação na literatura sobre boas práticas em modelagem de bancos de dados biológicos e sua utilização. Alguns utilizam o termo modelagem conceitual quando na realidade estão se referindo claramente à modelagem lógica; e nenhum deles trata de um domínio ou representação esquemática que possa ser utilizado para o projeto desta tese.

Modelos de Dados

[Allen *et al.*, 2006][27] mostram que enquanto a modelagem de dados é uma área de pesquisa muito bem estabelecida na ciência da computação, ainda existem muitas e ricas oportunidades para pesquisa, e que novas áreas vêm surgindo continuamente, incluindo a área de genômica.

O modelo de dados relacional contribuiu para a separação da representação lógica dos dados (relações e tuplas) da implementação física (arquivos e mecanismos de acesso) (Codd EF, 1970). Desde então, vários modelos de dados semânticos têm sido propostos.

[26] Por exemplo, o modelo lógico relacional (Codd EF, 1970).
[27] O resumo da lista de discussões da AMCIS 2005, sobre avanços em modelagem de dados, patrocinado pelo "Special Interest Group on Systems Analysis and Design (SIGSAND)"

Exemplos incluem:

- Entity Relationship Model (ERM) (Chen PP, 1976),
- Extended ERM (Smith and Smith, 1977),
- Semantic Data Model (SDM) (Hammer and Mcleod, 1981),
- Unified Semantic Model (USM) (Ram S, 1995).

Um levantamento da literatura [Allen G *et al.*, 2006] sobre avaliação de métodos de modelagem revelou alguns atributos necessários, que incluem:

- Adequação ou riqueza do método de modelagem em representar a realidade subjacente;
- Legibilidade do esquema obtido com o método de modelagem;
- Quão fácil é a utilização do método de modelagem para a representação dos requerimentos.

Por exemplo, a legibilidade do método de modelagem indica essencialmente quão fácil é a leitura do esquema modelado e a reconstrução da realidade do domínio a partir desse esquema. É desejável em situações onde os esquemas são criados por um time de análise e precisam ser lidos e interpretados por outros analistas, desenvolvedores ou administradores de sistemas. No entanto, diferentes modelos podem enxergar a realidade de diferentes formas, assim, torna-se difícil isolar que aspecto de um modelo pode causar maior ou menor legibilidade.

Num modelo de dados, a descrição da base de dados e a base de dados em si são conceitos distintos. A descrição da base é um esquema, e uma das formas de visualização de um esquema é um diagrama do mesmo. Diferentes modelos de dados possuem diferentes convenções para o diagrama de um esquema[28].

[28] Veja por exemplo uma discussão dos modelos ER (Entity Relationship), OR (Object Relational) e UML (Unified Modeling Language) em [Halpin TA, 2004].

Sistema Gerenciador de Banco de Dados – SGBD

Sistemas gerenciadores de bancos de dados comerciais existem desde a década de sessenta. Dentre os mais antigos estão os modelos de rede e hierárquico (Bachman CW, 1969; Tsichritzis *et al.*, 1976).

SGDBs são aplicações de software especializadas, desenvolvidas para executar as tarefas fundamentais de bancos de dados como armazenar e organizar informações, garantir que os dados estejam livres de contradições internas, reforçar restrições específicas dos dados, e retornar resultados consistentes de consultas simultâneas de múltiplos usuários [Nelson *et al.*, 2003]. Os tipos de SGDBs mais comuns atualmente são o relacional, o orientado a objeto e o hierárquico.

Os SGDBs relacionais são certamente os mais populares. Pode-se citar alguns sistemas comerciais que utilizam o modelo de dados relacional, como Oracle[29], IBM's DB2[30], Microsoft's SQL Server[31] e Sybase[32], e alguns de código aberto como MySQL[33] e PostgreSQL[34].

SGBDs relacionais são amplamente utilizados e representam uma tecnologia adequada para gerenciar grandes quantidades de dados. Eles provêem facilidades para uma organização de dados estruturada e não redundante, além de uma linguagem de consulta declarativa – SQL – para gerenciá-los [Chen and Sidhu, 2007].

A primeira versão do PWDB foi implementada utilizando o SGBD relacional IBM-DB2 [Otto, **Bezerra** *et al.*, 2010] e o esquema relacional da segunda versão foi desenvolvido para o SGBD PostgreSQL [Lifschitz, **Bezerra** *et al.*, 2012].

[29] http://www.oracle.com
[30] http://www-3.ibm.com/ software/data/db2/
[31] http://www.microsoft.com/sql/ default.asp
[32] http://www.sybase.com/home
[33] http://www.mysql.com/
[34] http://www.postgresql.org/

2. OBJETIVOS

1. Fazer um estudo detalhado do arcabouço biológico necessário para suprir os requisitos do PCG;
2. Descrever e discutir detalhadamente a modelagem conceitual para a implementação do sistema de banco de dados PWDB, proposto para responder eficientemente esses requisitos através de consultas diretas ao banco e de procedimentos mais complexos;
3. Representar o modelo num esquema conceitual que permita:
   3.1. Que usuários das ciências biológicas (sem conhecimento em linguagens de consulta de banco de dados) possam esboçar consultas e procedimentos utilizando os objetos do esquema, mesmo antes da existência de um esquema lógico
   3.2. Direcionar o desenvolvimento e a implementação física do sistema.

Pontos importantes considerados:

- Análise detalhada de fontes de dados biológicos de referência para integrar o sistema;
- Seleção de bases de dados, a partir desta análise, que permitam a associação das sequências de proteína comparadas no PCG com dados de anotação e informações funcionais, dentro de um contexto genômico;
- Definição de protocolos para a construção de referências cruzadas entre as sequências de proteína e as diferentes bases de dados selecionadas.

Limites da modelagem:

A delimitação do domínio, as questões que poderão ser abordadas, e que respostas poderão ser obtidas com o esquema conceitual proposto dependerão das fontes de dados selecionadas, e das referências cruzadas que forem construídas com a matriz de similaridade resultante do PCG.

3. MATERIAL E MÉTODOS

3.1. PROJETO CONCEITUAL

O principal objetivo da modelagem conceitual de um banco de dados (BD) é a abstração, definição e conhecimento do domínio. Esta etapa busca representar, em uma linguagem de alto nível, os conceitos/objetos presentes no domínio do problema.

A definição e compreensão da semântica do domínio no projeto conceitual é bastante facilitada devido a característica do esquema conceitual de descrever o modelo de dados de maneira independente de representações computacionais.

Teoricamente, devido a razões como correção, clareza, riqueza de informação e portabilidade, o desenho do BD deve ser representado, primeiramente, em um esquema conceitual, que é uma representação gráfica do modelo conceitual, para só então ser efetuado o mapeamento para o esquema do BD. A qualidade de descrição de um esquema conceitual é fundamental, pois é a partir dele que se realiza o mapeamento para o modelo lógico que guiará a implementação do sistema (modelo físico).

Assim, um projeto de banco de dados deve evoluir de acordo com as seguintes etapas:

- **Projeto Conceitual**: representação dos requisitos de dados do domínio; não tem nenhuma dependência do Sistema Gerenciador de Banco de Dados (SGBD) e nem de requisitos computacionais;
- **Projeto Lógico**: representação do esquema conceitual em um modelo de BD. Nesse momento já existe uma dependência da classe do SGBD, mas não do SGBD específico. O esquema conceitual é mapeado num esquema lógico[35], descrito em termos de um modelo de dados genérico (por exemplo, relacional) escolhido de acordo com os propósitos da implementação;
- **Projeto Físico**: tem total dependência do SGBD; o esquema físico é construído adaptando o esquema lógico ao SGBD específico (por exemplo, PostgreSQL), e já são consideradas questões referentes à *performance* do sistema.

[35] Ver [Lifschitz, **Bezerra** *et al*., 2012]

A FIGURA 3.1 apresenta, de forma gráfica e concisa, as principais etapas do desenvolvimento de um sistema de banco de dados. Os projetos conceitual, lógico e físico são etapas independentes tratadas em diferentes momentos do projeto.

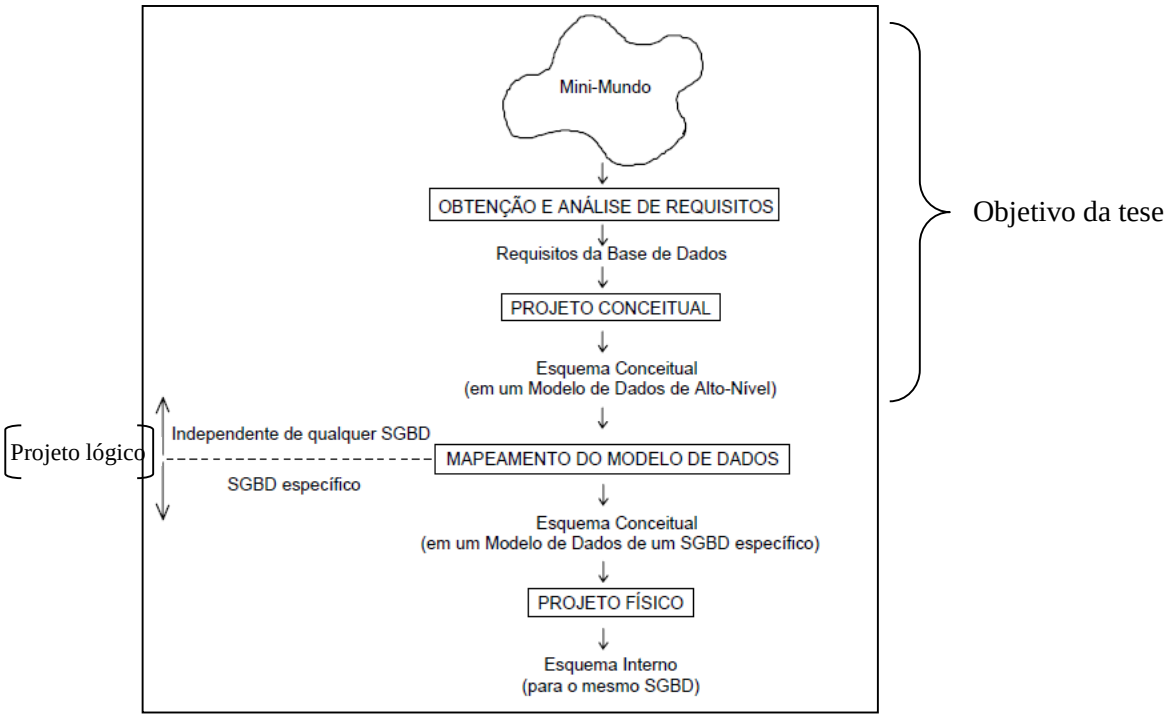


FIGURA 3.1. Diagrama simplificado com as principais fases do projeto de um BD (adaptado de [Elmasri and Navathe, 2011]).

Diagrama Entidade-Relacionamento (DER)

Neste trabalho, para a representação dos conceitos e objetos num esquema conceitual utilizou-se o DIAGRAMA ENTIDADE-RELACIONAMENTO (DER) [Harrington JL, 2009; Chen *et al.*, 1999; Elmasri R and Navathe, 2011].

Neste diagrama (FIGURA 3.2), os objetos são classificados basicamente em:

- ENTIDADE: abstração de um fato do mundo real para o qual se deseja manter os dados;
- RELACIONAMENTO: abstração de uma associação entre (ocorrências de) entidades;
- ATRIBUTO: abstração de uma propriedade de uma entidade ou de um relacionamento.

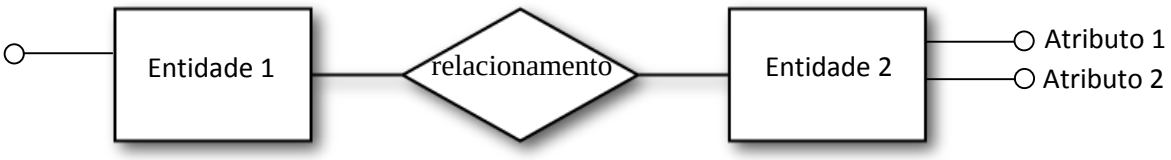


FIGURA 3.2: Elementos básicos de um esquema Conceitual representado por um Diagrama Entidade-Relacionamento: *Retângulos* representam ENTIDADES; *Losangos* representam RELACIONAMENTOS entre entidades e *Círculos* representam os ATRIBUTOS de entidades ou relacionamentos.

O objeto básico é a ENTIDADE. Uma entidade representa alguma coisa do mundo real que possui uma existência independente. Pode ser um objeto com uma existência física – uma pessoa por exemplo, ou pode ser um objeto com existência conceitual – uma empresa por exemplo. Uma entidade provê uma descrição das propriedades que são compartilhadas por uma coleção de instâncias de um domínio.

Os ATRIBUTOS de uma entidade têm propriedades particulares que a descreve. Uma entidade EMPREGADO pode ser definida por nome, trabalho, idade, endereço e salário, por exemplo. Os atributos de uma entidade indicam que valores podem ser armazenados para identificar ou descrever uma instância desse tipo.

Os RELACIONAMENTOS podem ter atributos. É o caso do valor de um atributo descrever uma relação sem ser atributo de nenhuma das duas entidades participantes. Um relacionamento também pode associar uma entidade a ela mesma.

Relacionamentos possuem certas "restrições" que limitam as possíveis combinações de entidades participantes em instâncias do relacionamento. São determinadas pelas situações do "mini-mundo" que os relacionamentos representam. Por exemplo, se existe uma regra que define que um empregado trabalha em apenas um departamento, essa restrição deve estar descrita no esquema.

Dois tipos frequentes de restrições de relacionamento são: *razão de cardinalidade* e *participação*. A razão de cardinalidade especifica a quantidade de instâncias do relacionamento que uma entidade pode participar. As mais comuns para relacionamentos binários são 1:1, 1:N e M:N.

Na FIGURA 3.3, por exemplo, o tipo de relacionamento binário trabalha-em, entre as entidades DEPARTAMENTO e EMPREGADO tem razão de cardinalidade 1:N, e significa que:

- Cada instância de DEPARTAMENTO pode estar relacionada a inúmeras instâncias de EMPREGADO (muitos empregados podem trabalhar em um departamento),
- Mas uma instância de EMPREGADO pode estar relacionada a apenas um DEPARTAMENTO (um empregado pode trabalhar em apenas um departamento).

FIGURA 3.3. Relacionamento trabalha-em

A restrição de participação define se a existência de uma entidade depende de um relacionamento com outra entidade; pode ter participação total ou parcial. Na FIGURA 3.4, por exemplo:

- Se existe uma regra de que todo departamento precisa ter um gerente, uma instância da entidade DEPARTAMENTO só existe se participar em uma instância do relacionamento gerencia – a participação é total.
- Nem todo empregado gerencia um departamento, assim, a participação de EMPREGADO no relacionamento gerencia é parcial, significando que nem todas as instâncias de EMPREGADO estarão relacionadas a instâncias de DEPARTAMENTO através do relacionamento gerencia.

FIGURA 3.4. Relacionamento gerencia

3.2. DADOS PRIMÁRIOS – SEQUÊNCIAS

***A**. Sequências Comparadas no WCG*

O conjunto básico de sequências comparadas no WCG foram sequências de aminoácidos, obtidas de duas fontes de dados: RefSeq v.21 e Uniprot-SwissProt v.5.5[36]. Além deste, também foi comparado um conjunto adicional de sequências tORF[37], específico do PCG.

Dentre as sequências comparadas no PCG existem[38]:

(a) Sequências de proteína que possuem anotação de genes, mRNAs[39] e CDSs[40] na sequência genômica;
(b) Sequências de proteína para as quais as únicas sequências de nucleotídeos de origem são mRNAs;
(c) Sequências de proteína provenientes de sequências genômicas, sem referência ao mRNA;
(d) Sequências de proteína obtidas diretamente do sequenciamento de moléculas de proteína, sem nenhuma sequência de nucleotídeos de origem;
(e) Sequências tORF que possuem, obrigatoriamente, a posição de sua sequência ncORF de origem definida numa sequência genômica completa de procarioto.

– Identificador de Sequências

Como o conjunto total de sequências comparadas no WCG inclui sequências de aminoácidos de duas bases de dados distintas de sequências de proteína – RefSeq e SwissProt, além de sequências tORF (inexistentes nestas bases), não existe um identificador natural que possa ser utilizado para definir unicamente cada sequência. Desta forma, foi criado o identificador "**fiocruzid**" para nomear as sequências independentemente de sua origem e tipo (FIGURA 3.5).

[36] ANEXO IV
[37] "*Non-coding*_ORF" (ncORF) e "*translated*_ORF" (tORF) são termos definidos para o esquema conceitual do PWDB para representar conceitos específicos do PCG. Não são proteínas cadastradas em bancos de dados de sequências. Maiores detalhes adiante nesse mesmo tópico 3.2 e em DISCUSSÃO.
[38] Maiores detalhes adiante nesse mesmo tópico 3.2 e em RESULTADOS
[39] RNA mensageiro
[40] *CoDing Sequence*

Algumas regras foram consideradas na construção deste *surrogate*[41,42]:

(a) Nas sequências provenientes da base de dados RefSeq, o **fiocruzid** equivale ao GI do NCBI[43], como por exemplo 51893456 na FIGURA 3.5.

(b) Nas sequências provenientes da base de dados SwissProt, o **fiocruzid** é um número sequencial começando por 150000 ou 900000 mais 8 casas decimais, como por exemplo 900000000000002 na FIGURA 3.5.

(c) Nas sequências tORF, o **fiocruzid** é um número sequencial simples de 8 casas decimais, como por exemplo 00000769 na FIGURA 3.5.

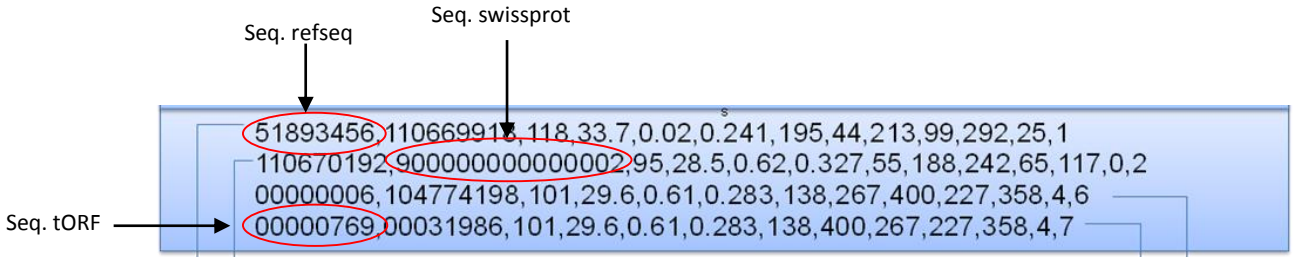


FIGURA 3.5. "Recorte" do *output* resultante da execução do SSEARCH. Cada linha apresenta o resultado da comparação de um par de sequências, identificadas pelos dois primeiros valores – por exemplo, na primeira linha, a sequência consulta (*query*) tem o identificador 51893456 e a sequência comparada (*subject*) tem o identificador 11066918. O restante das informações são valores de similaridade, cobertura e estatísticos do alinhamento[44].

– Tipos de Comparação[45]

No WCG foram executadas comparações "par-a-par" e "todas-contra-todas", da seguinte forma:

- Todas as proteínas RefSeq contra todas as proteínas RefSeq;
- Todas as proteínas RefSeq contra todas as proteínas SwissProt;
- Todas as proteínas SwissProt contra todas as proteínas SwissProt;
- Todas as proteínas SwissProt contra todas as sequências tORF;
- Todas as sequências tORF contra todas as sequências tORF[46].

[41] Uma chave artificial que permite identificar unicamente as entidades a representar.
[42] Os identificadores dos casos (b) e (c) abaixo, nas suas respectivas bases de dados de origem, podem ser recuperados dos dados de entrada do PCG.
[43] O número GI é um identificador único, interno do NCBI. Cada sequência de nucleotídeos e de proteína cadastrada no NCBI tem um número GI associado (fonte: glossário NCBI).
[44] O significado de cada "campo" de uma linha (separados por vírgulas) está especificado no ANEXO III.
[45] ANEXO III.
[46] Proteínas RefSeq NÃO foram comparadas com sequências tORF.

– Resultado das Comparações

O resultado do PCG é uma matriz, de aproximadamente 900 GB, com índices de similaridade entre pares de sequências de aminoácidos. Importante comentar que só foram armazenados os resultados de alinhamentos significativos[47], e que o resultado de uma comparação nada mais é do que uma linha com informações de similaridade, cobertura e estatísticas do alinhamento, fornecidas pelo programa SSEARH (FIGURA 3.5 acima).

Vale ressaltar o fato desta matriz disponibilizar 4.2×10^9 resultados de similaridade entre sequências, já pré-calculados, por um método não heurístico, que utiliza programação dinâmica. E que centenas de genomas completos (e incompletos) foram comparados "par-a-par" e todos-contra-todos".

***B**. Características das Sequências Consideradas no Projeto Conceitual do PWDB*

– Sequências ORF e CDS[48]

- Um "quadro aberto de leitura" (*Open Reading Frame* – ORF) é uma série de códons (trincas) de nucleotídeos que se estende até o primeiro códon de terminação, e pode ser ou não uma região codificadora de proteína.
- Uma sequência de nucleotídeos codificadora de proteína (*CoDing Sequence* – CDS) é uma ORF, com códon de início de tradução e códon de término de tradução, que codifica uma proteína.
- Assim, toda sequência CDS é uma ORF, mas nem toda sequência ORF é um CDS.

– Sequências ORF, ncORF e tORF

- "*Non-coding*_ORF" (ncORF) e "*translated*_ORF" (tORF) são nomenclaturas definidas para o projeto conceitual do PWDB.
- Sequências de nucleotídeos ncORFs representam ORFs numa sequência genômica que não foram identificadas como codificadores de proteína por métodos de predição de genes durante o processo de anotação do genoma.

[47] Atingiram os valores mínimos de *score* e estatísticos exigidos pelo programa, de acordo com os parâmetros selecionados para a comparação. Ver ANEXO III.

[48] Na nomenclatura do NCBI: CDS – "*CoDing Sequence*" – é região de nucleotídeos do mRNA que é traduzida em aminoácidos. Uma sequência codificadora de proteína (CDS) é uma sequência de nucleotídeos que começa com um códon de iniciação, termina com um códon *STOP* e determina a sequência de aminoácidos de uma proteína [Glossário NCBI].

- As sequências de aminoácidos tORFs foram obtidas pela "tradução conceitual"[49] de sequências ncORFs, com base no mesmo código genético utilizado para a tradução das sequências CDSs nas bases de dados de proteína.
- Como as sequências tORFs não foram preditas como proteínas, não existem nas bases de dados de proteínas.
- Apenas ORFs contidas integralmente em regiões descritas como não codificadoras foram consideradas como ncORFs; qualquer tipo de sobreposição de uma ORF com uma sequência anotada como codificadora, foi excluída.
- Sequências tORFs foram comparadas no WCG apenas com sequências de proteína da base de dados SwissProt[50].

– Sequências Genômicas[51]
  - *Status*[52] *Complete* – tipicamente significa que cada cromossomo está representado por apenas uma sequência com montagem de alta qualidade,
  - *Status Assembly* – tipicamente significa que existem montagens que ainda não estão no nível de cromossomo e/ou *draft*[53],
  - *Status In Progress* – indica que o projeto de sequenciamento está numa fase de pré montagem ou as sequências montadas/completas ainda não foram submetidas ao GenBank/EMBL/DDBJ[54].
  - Prefixo NC_[55]: foram obtidas por procedimento automatizado e possuem revisão de especialista para alguns registros. O sistema de coordenadas da sequência, posicionamento de genes e anotação são mais estáveis.
  - Prefixos NT_, NW_, NZ_[56]: indicam registros que não foram revisados individualmente; as atualizações do genoma são liberadas como um bloco. A montagem, anotação e posicionamento dos genes são provisórios. No PWDB, o tratamento dispensado a essas sequências deve ser diferenciado e mais cuidadoso, considerando que os relacionamentos entre as sequências de proteína, e suas sequências CDSs e genômicas podem ser incompletos ou não existir.

[49] Um códon de nucleotídeos codifica um aminoácido de acordo com um código genético.
[50] Sequências de proteína da base de dados RefSeq NÃO foram comparadas com sequências tORF
[51] Anexo IV – Base RefSeq.
[52] Propriedade que se refere ao estágio corrente do projeto de sequenciamento do genoma.
[53] Refere-se a uma sequência de DNA que ainda não está terminada mas tem, geralmente, alta qualidade (precisão > 90%).
[54] Ver ANEXO IV.
[55] Moléculas genômicas completas incluindo cromossomos, organelas e plasmídeos
[56] *Contig* ou *scaffold* e sequenciamento *whole genome shotgun* não finalizado [*unfinished* WGS].

***C.** Utilização das Sequências de Aminoácidos no PWDB* [57]

- As sequências de proteína provenientes de projetos genômicos completos[58] são indicadas para estudos comparativos de proteomas completos (“preditos”);
- As demais sequências de proteína[59] podem ser utilizadas para identificar e confirmar resultados de anotação e também como informação auxiliar em outros procedimentos. Podem ser úteis para (in)validar, por exemplo:
  - A existência de um gene não conhecido em genomas completos, cuja proteína traduzida de uma sequência similar exista nos bancos de dados,
  - Uma anotação experimental inexistente no conjunto de proteomas completos,
  - A existência de novas combinações de domínios em proteínas multi-modulares, não identificadas no conjunto de proteomas completos,
  - Dentre outros.
- O grupo experimental de sequências tORF foi criado para avaliar o potencial codificador de pequenas sequências ncORF[60].

3.3. Bases De Dados Biológicos[61]

Para a modelagem dos conceitos necessários para responder os requisitos do PCG foram construídas referências cruzadas entre bases de dados biológicos públicas e as sequências armazenadas na matriz de resultados do PCG.

Nesta etapa, algumas bases de dados foram detalhadamente estudadas, e algumas das características analisadas, relevantes para a escolha, foram:

- Não redundantes e curadas,
- Mantidas por grupos internacionais de pesquisa renomados[62] e com reconhecida competência técnica e potencial financeiro para sua manutenção e atualização,

[57] Ver RESULTADOS.
[58] Possuem sequências de proteína, CDS, gene e genômica.
[59] Não se pode garantir que todas as proteínas de um organismo estejam representadas nas bases de dados, e nem que possuem um genoma completo de referência sequenciado.
[60] O PCG comparou estas sequências numa tentativa de identificar sequências candidatas a serem codificadoras – não detectadas por métodos computacionais de identificação de genes – utilizando o grau de similaridade com sequências de proteína já conhecidas e armazenadas em bancos de dados
[61] ANEXO IV.
[62] Muitas vezes geograficamente dispersos.

- Bem estabelecidas no meio acadêmico e de pesquisa,
- Disponíveis sem custos para a comunidade científica em geral.

As bases de dados inicialmente selecionadas para integrarem o PWDB foram:

- Bases de dados de referência de sequências de nucleotídeos (RefSeq e Gene),
- Bases de dados de referência de sequências e informações proteicas (RefSeq e UniProt),
- Base de dados de domínios proteicos (Pfam),
- Base de dados de ontologias de produtos gênicos (Gene Ontology – GO),
- Base de dados de vias biológicas, classes enzimáticas e outras informações de sistemas biológicos (KEGG).

Regras para a associação destas fontes externas com os dados de similaridade da matriz de resultados do PCG, e a definição de referências entre conceitos semanticamente equivalentes nas diferentes bases de dados foram definidas durante o processo de modelagem conceitual. O resultado é uma consequência imediata do trabalho desta tese e gerou o protocolo para a carga dos dados do PWDB. No entanto, o detalhamento desta etapa foge ao escopo dessa discussão[63].

[63] Ocorre numa fase posterior à modelagem conceitual. Para maiores detalhes ver [Tristão and Lifschitz, 2009].

4. RESULTADOS

4.1. MODELAGEM CONCEITUAL

Em bioinformática, o termo *hit* costuma ser utilizado para descrever alinhamentos recuperados para uma sequência consulta (*query*), numa busca por sequências similares em bancos de dados. Uma única busca pode retornar uma lista de *hits* para uma mesma *query*[64]. Em outros casos, o termo refere-se à sequência pareada (sequência *subject*) que apresenta similaridade na comparação com a sequência *query*. Da mesma forma, poderá existir uma lista de *hits* para uma *query* numa mesma busca num banco de dados.

No caso do PCG, as comparações foram "par-a-par" e o algoritmo utilizado foi o SSEARCH. Desta forma, tem-se um único resultado para cada par de sequências comparado, com as informações do alinhamento ótimo. Além disso, para que o resultado de uma comparação seja armazenado na matriz do PCG, é necessário que haja uma similaridade estatisticamente significativa[65] entre o par.

Nesta tese, o conceito "***hit***" foi definido como o resultado da comparação de um par de sequências de aminoácidos armazenado na matriz do PCG, e equivale a uma linha da matriz (FIGURA 3.5[66]).

É importante notar que:

- A única informação das sequências de proteína armazenadas na matriz do PCG são seus identificadores (fiocruzid/PWDId[67]), que é a única opção para o estabelecimento de qualquer relacionamento com bases de dados externas.
- Os identificadores que não correspondem ao GI da base de dados RefSeq podem ser recuperados das suas respectivas bases de dados de origem[68].

Na modelagem do PWDB, o passo inicial foi relacionar os identificadores das sequências de cada *hit* com fontes de dados externas de anotação, para a recuperação de informações preditas

[64]Para a escolha do(s) melhor(es) *hit*(s) diferentes medidas do alinhamento e composição e estrutura das sequências devem ser considerados.
[65] De acordo com as exigências do PCG e do programa de comparação SSEARH.
[66] Em MATERIAL E MÉTODOS
[67] No esquema em [Lifschitz, **Bezerra** *et al.*, 2012] e nesta tese, o identificador fiocruzid aparece como PWDId.
[68] Ver 3.2. Identificador de Sequências

para essas proteínas – de funções, ontologias, domínios proteicos, dentre outras –, e também para comparar anotações das diferentes fontes de dados. O objetivo deste procedimento é tentar transferir a anotação predita/confirmada de uma sequência (baseada no conjunto de bases de dados escolhido) para as sequências comparadas no PCG que apresentem índices de similaridade estatisticamente significativos, assim como outras características relevantes dos alinhamentos.

Além do requisito de anotação, foi definida uma estratégia adicional de forma a responder questões genômicas mais amplas. Nesse caso, a questão chave foi considerar as sequências de nucleotídeos da base de dados RefSeq, que deram origem às sequências de aminoácidos comparadas no PCG; e posicionar estas sequências em seus genomas. Para isso, utilizou-se parte do conhecimento clássico do dogma central da biologia:

GENE → *transcrição* → **mRNA** → *tradução* → **PROTEÍNA**

Essa abordagem se adapta bem aos propósitos do PWDB, pois no PCG apenas sequências proteicas foram consideradas[69], as quais, "conceitualmente" (no mundo real), possuem sua origem em regiões genômicas (e gênicas) codificadoras de proteínas.

Desta forma, as sequências de proteína foram associadas às suas sequências de nucleotídeos de origem através da sequência codificadora de proteína (CDS), baseando-se no conceito "*CDS feature*" descrito no NCBI *Data Model* [Ostell *et al.*, 2001] (FIGURA 4.1).

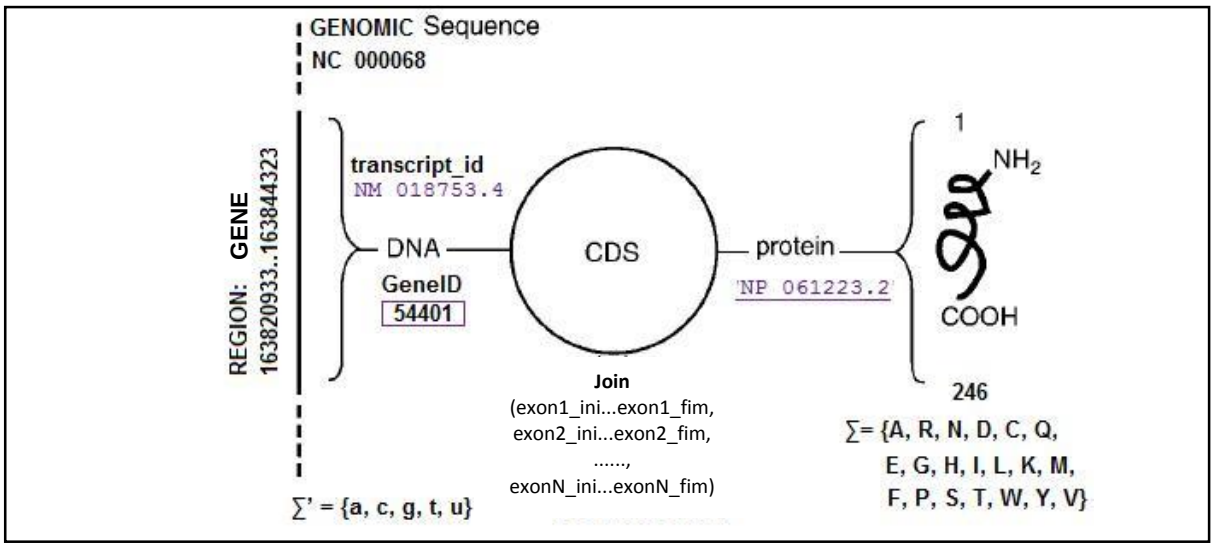


FIGURA 4.1. "CDS *feature*" – NCBI *data model* (adaptado de [Ostell *et al.*, 2001] com dados da FIGURA 4.2).

[69] Excluindo as sequências torf que não são proteínas.

De acordo com esse raciocínio, pode-se "traçar um caminho"[70], desde uma sequência de proteína até sua sequência genômica, utilizando a sequência **CDS** como o ponto central para a transformação:

SEQUÊNCIA DE NUCLEOTÍDEOS → **CDS** → SEQUÊNCIA DE AMINOÁCIDOS

Este "caminho" PROTEÍNA → SEQUÊNCIA GENÔMICA (FIGURA 4.2) pode ser descrito, resumidamente, como segue:

- Um GENE codificador de proteína é uma subsequência de uma SEQUÊNCIA GENÔMICA[71] [DNA, $\sum = \{a, c, g, t\}$]
- A transcrição de um GENE codificador de proteína gera uma sequência TRANSCRITA primária [RNA, $\sum' = \{a, c, g, u\}$]
- A sequência TRANSCRITA primária é processada[72] gerando um mRNA MADURO [RNA, $\sum' = \{a, c, g, u\}$]
- A sequência mRNA MADURO possui a subsequência codificadora de proteína (CDS) e regiões reguladoras[73],
- A sequência CDS é traduzida[74] numa sequência de PROTEÍNA [$\sum'' = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$].

Na FIGURA 4.2 pode-se ver informações da proteína **NP_061223**:

A.1 em sua Sequência Genômica (Cromossomo 2, Accession: NC_00068),

A.2 em sua Sequência Gênica (NC_00068 REGION 163.820.933 .. 163.844.323),

A.3 em sua Sequência mRNA (transcript_id: NM_018753),

A.4 em seu CDS (protein_id: NP_061223),

B.1 em sua Sequência mRNA (DBSOURCE RefSeq: accession NM_018753),

B.2 em seu CDS (*coded_by* = NM_018753:173..913)

[70] Este "caminho" seria uma forma de leitura do diagrama apresentado mais adiante e seus conceitos. No momento não se trata de manipulacao de modelo conceitual, pois isso ocorre após a transformação para um modelo lógico (relacional, por exemplo). Mas pode-se ver que definições e suas representações já estão bem claros desde o modelo conceitual, ponto fundamental da proposta desta tese.
[71] Na maioria dos casos são sequências de DNA, mas existem genomas de RNA [$\sum' = \{a, c, g, u\}$].
[72] O processo é diferente em eucariotos e procariotos.
[73] A concatenação de subsequências codificadoras da sequência transcrita primária (exons), mais as regiões reguladoras UTR.
[74] De acordo com um Código Genético, uma trinca de mRNA codifica um aminoácido. Por exemplo, a trinca de nucleotídeos AUG codifica o aminoácido Metionina (M).

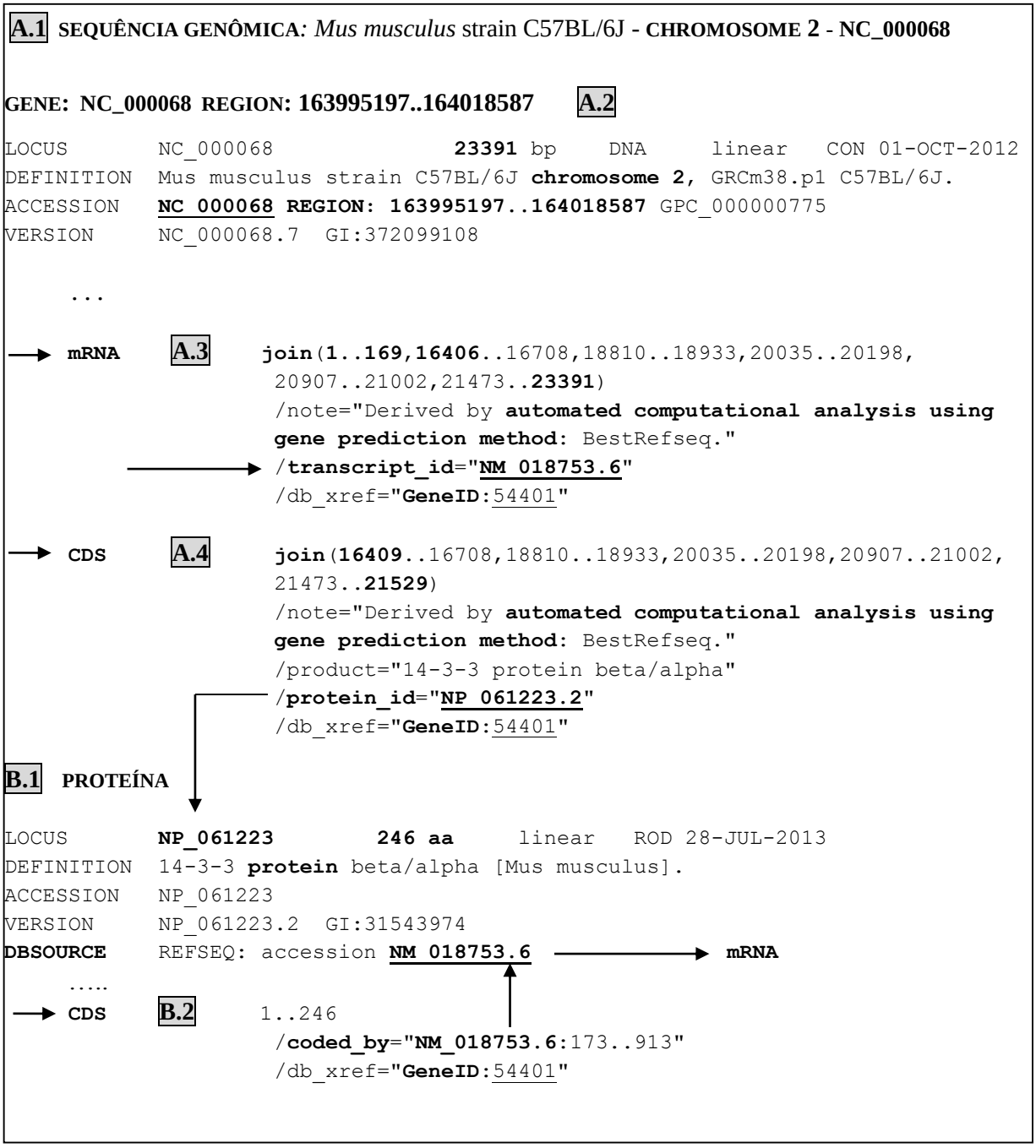


FIGURA 4.2[75]. Exemplo usado na FIGURA 4.1.

Pode-se ver mapeado na sequência do **gene** em A.3: as posições dos fragmentos que formam o mRNA[76] (com o identificador do transcrito – *transcript_id*) e em A.4[77]: as posições dos exons do mRNA que formam a região codificadora da proteína (CDS, com o identificador da sequência de aminoácidos codificada – *protein_id*).

Na sequência da **proteína**, pode-se ver em B.1: o identificador do transcrito que deu origem à proteína, e em B.2: a posição ocupada pela região codificadora da proteína no transcrito.

[75] Informações obtidas na base de dados RefSeq.
[76] Exons e regiões reguladoras
[77] Note a diferença da posição inicial e final em A.3 e A.4.

4.2. ESQUEMA CONCEITUAL

Nessa tese, foi utilizado um DIAGRAMA ENTIDADE-RELACIONAMENTO para a representação do modelo conceitual; e apenas com o propósito de simplificar sua descrição, o esquema da FIGURA 4.8 será apresentado, inicialmente, como quatro módulos independentes que se conectam; nomeados como: CENTRAL, TAXONOMIA, HIT e ANOTAÇÃO.

A. MÓDULOS

1) **CENTRAL** (ou Dogma Central) (FIGURA 4.3): envolve as entidades que representam as sequências de aminoácidos e nucleotídeos.

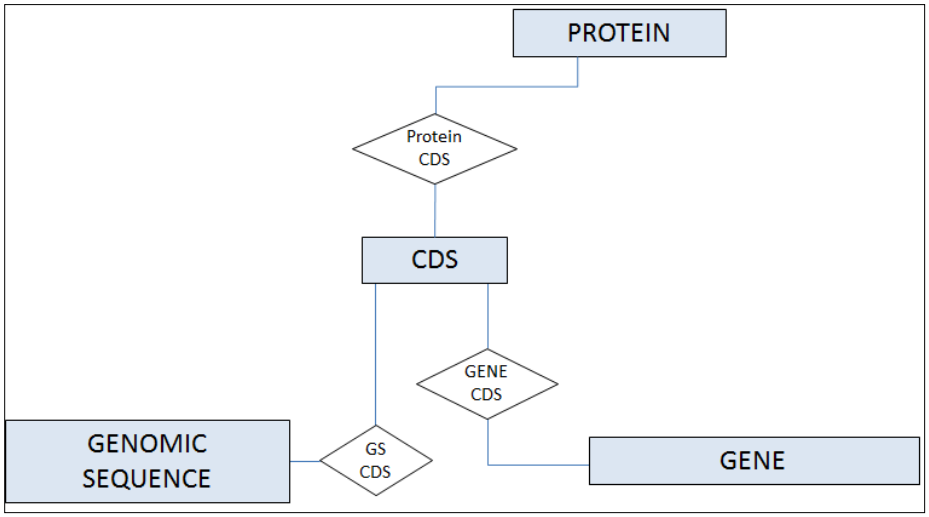


FIGURA 4.3. Módulo CENTRAL

A ligação entre as entidades PROTEIN, GENE e GENOMIC SEQUENCE é intermediada pela entidade CDS através do posicionamento das subsequências codificadoras do gene (exons) no sistema de coordenadas da sequência genômica que o contém.

Para facilitar a leitura do esquema é importante lembrar que os elementos:

- GENOMIC SEQUENCE, GENE, CDS, ORF e ncORF referem-se a sequências de nucleotídeos,
- PROTEIN, ORF_T e tORF referem-se a sequências de aminoácidos.

A base de dados de referência para a associação das sequências de proteína com suas sequências de nucleotídeos de origem é o NCBI Refseq. Adicionalmente, informações complementares podem ser obtidas nas bases de dados UniProt-SwissProt e NCBI Entrez-Gene.

Além do contexto inicial, uma sequência de proteína (FIGURA 4.4):

- Pertence a um táxon (módulo TAXONOMY),
- A comparação de sequências gera medidas de similaridade entre elas (módulo HIT)
- A descrição da sequência é obtida de uma anotação (módulo ANOTATION).

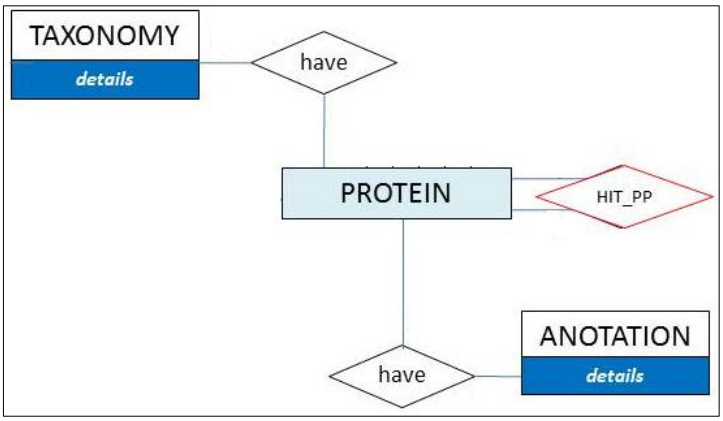


FIGURA 4.4. Módulos TAXONOMIA, HIT E ANOTAÇÃO, detalhados a seguir.

2) **HIT** (FIGURA 4.5): considera os resultados de similaridade entre sequências de aminoácidos, calculados e armazenados pelo PCG.

Os *hits* são relacionamentos entre sequências de aminoácidos:

- `HIT_PP` – apenas proteínas,
- `HIT_OP` – proteínas e tOrfs
- `HIT_OO` – apenas tOrfs

FIGURA 4.5. Módulo HIT.

3) **TAXONOMIA** (FIGURA 4.6): utiliza a taxonomia do NCBI e relaciona cada sequência de aminoácidos ao seu táxon[78].

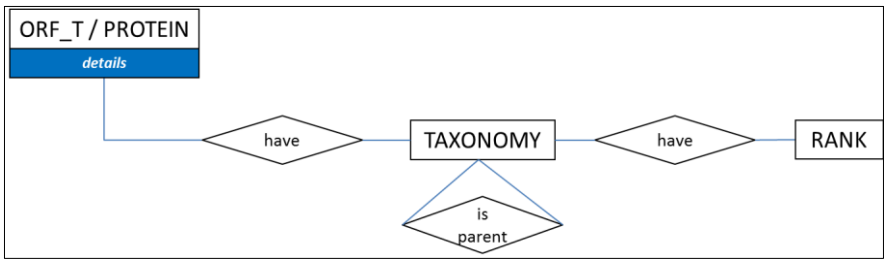


FIGURA 4.6. Módulo TAXONOMIA.

4) **ANOTAÇÃO** (FIGURA 4.7): permite a construção de referências cruzadas entre bases de dados de anotação de proteínas e o módulo CENTRAL, através da entidade PROTEIN.

No caso específico do PWDB, as primeiras bases de dados selecionadas[79] foram:

- UniProt
- Pfam,
- Gene Ontology (GO),
- KEGG.

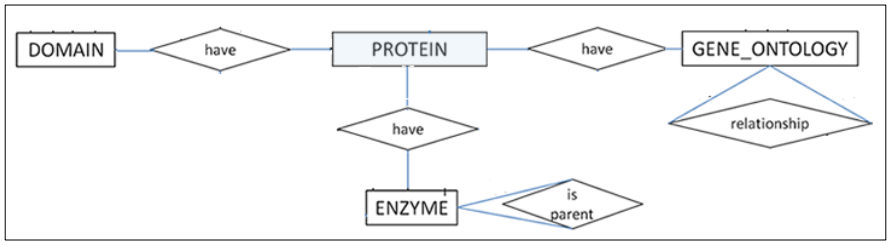


FIGURA 4.7. Módulo ANOTAÇÃO[80].

B. ENTIDADES E RELACIONAMENTOS

A FIGURA 4.8 apresenta o esquema conceitual completo – com entidades, atributos e relacionamentos – discutidos e detalhados em seguida.

[78] A especificação da entidade RANK, assim como dos relacionamentos *have* e *is parent* estão descritos em [Tristão and Lifschitz, 2009].
[79] ANEXO IV.
[80] A especificação dos atributos e relacionamentos das entidades DOMAIN, ENZIME e GENE_ONTOLOGY estão descritos em [Tristão and Lifschitz, 2009].

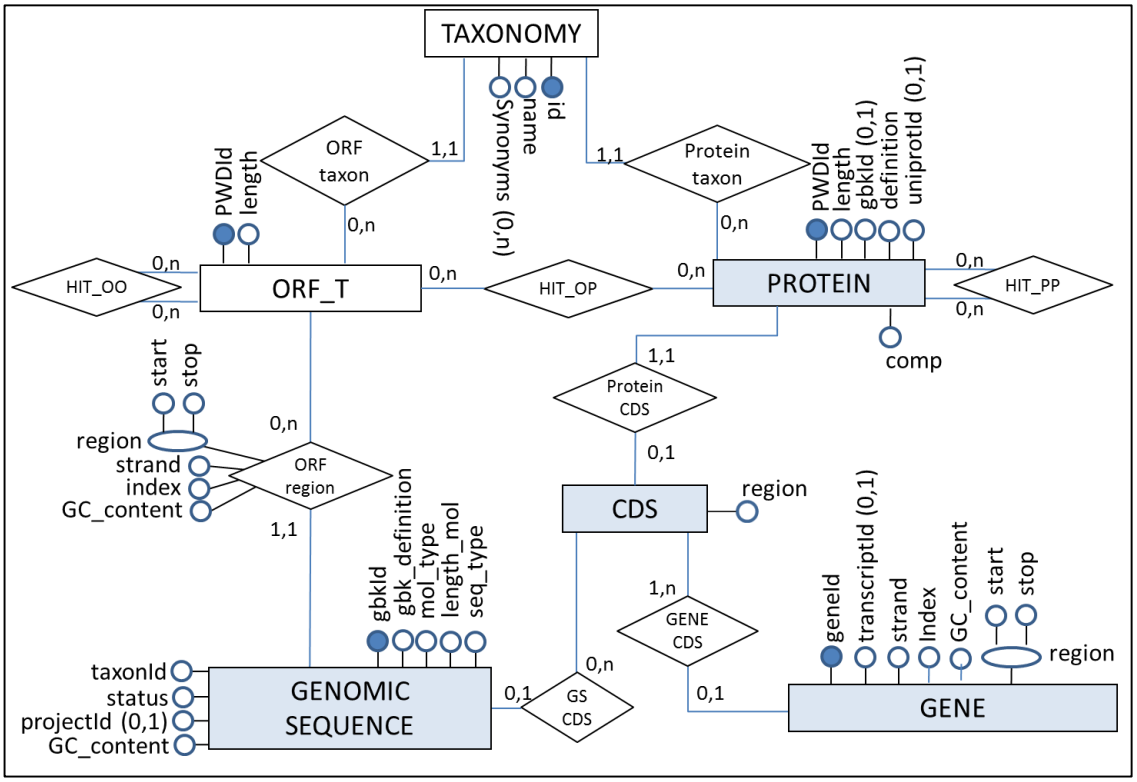


FIGURA 4.8. Esquema Conceitual representado por um diagrama Entidade-Relacionamento

B.1. ENTIDADES

PROTEIN

A entidade representa sequências de aminoácidos anotadas como proteína nas bases de dados de referência RefSeq e SwissProt.

Atributos[81]:

- PWDId/fiocruzid[82]: identificador único das sequências comparadas no PCG, e das sequências RefSeq
- definition: definição da proteína na base de dados RefSeq[83],
- length: comprimento da sequência de proteína,
- gbkId: identificador *accession-version* na base de dados Refseq
- uniprotId: identificador da sequência na base de dados UniProt.
- comp: sinalizador que define se a sequência foi comparada no PCG.

[81] Atributos não obrigatórios: gbkId, UniprotId
[82] Rever MATERIAL E MÉTODOS 3.2.
[83] Ou SwissProt, no caso de sequências não representadas na base de dados RefSeq.

Se a sequência foi comparada, comp=1,

Se não foi comparada, comp=0.

No caso específico do PWDB, as sequências de proteína foram originadas das bases:

- Refseq v.21: foram comparadas no PCG [comp=1], possuem o identificador gbk_id, e poderão ter ou não o identificador uniprotId (0,1).
- Swissprot v.51.5: foram comparadas no PCG [comp=1], possuem o identificador uniprotId, e poderão ter ou não o identificador gbk_id (0,1);
- Refseq v.33: não foram comparadas no PCG [comp=0], possuem o identificador gbk_id, e poderão ter ou não o identificador uniprotId (0,1). Apesar de não terem sido comparadas no PCG, possuem o identificados PWDId que é o GI do NCBI[84].

ORF_T

Apesar das sequências tORFs serem "computacionalmente" sequências de aminoácidos (como as sequências de proteína), elas são conceitualmente distintas e, portanto, estão representadas por uma entidade própria: ORF_T. Essas sequências foram "traduzidas" a partir de sequências ncORFs e não existem em bases de dados de proteína[85].

Atributos:

- PWDid/fiocruzid[86]: identificador único da sequência tORF no PCG,
- length: comprimento da sequência tORF.

CDS

A entidade representa a região codificadora de uma proteína[87].

É uma entidade cuja propriedade básica é permitir a associação entre as entidades PROTEIN, GENE e GENOMIC SEQUENCE, através do posicionamento das subsequências codificadoras do gene (*exons*) no sistema de coordenadas da sequência genômica que o contém.

Atributos:

- region: região definida pelas posições dos *exons* na sequência genômica[88]. Cada *exon* na sequência genômica correspondente a uma subsequência em CDS.

[84] Rever MATERIAL E MÉTODOS seção 3.2
[85] Não foram identificadas como sequências codificadoras de proteína por métodos de predição de genes
[86] Rever MATERIAL E MÉTODOS seção 3.2.
[87] Formada pela concatenação dos exons. A sequência completa é codificadora.
[88] FIGURA 4.2, A.4: join (16409..16708, 18810..18933, ... , 21473..21529).

GENE

A entidade GENE representa uma região gênica de uma sequência genômica.

Atributos[89]:

- geneId: identificador único da base de dados NCBI Gene,
- transcriptId: identificador da sequência transcrito.
- strand: sentido de leitura da sequência gênica em relação à sequência genômica referência[90].
- index: posição que o gene ocupa em relação aos outros genes da mesma sequência genômica:

  Número total de genes da Sequência Genômica = N,
  Posição $gene_1$ < posição $gene_2$ < ... < posição $gene_N$,
  index $gene_1$ = 1º, index $gene_2$ = 2º, ..., index $gene_N$ = $N^{ésimo}$.

- GC_content: conteúdo GC[91] da região gênica.
- region: região ocupada pela sequência gênica na sequência genômica[92].

GENOMIC SEQUENCE

A entidade representa uma sequência genômica.

Atributos[93,94]:

- gbkId: identificador da sequência na base de dados Refseq,
- gbk_definition: definição da sequência na base de dados Refseq,
- mol_type: tipo de molécula (DNA/RNA),
- seq_type: tipo de sequência (cromossomo, organela, etc),
- length_mol: comprimento da sequência genômica,
- taxonId: identificador do organismo de origem no NCBI Taxonomy DB
- status: definição do estado corrente do projeto de sequenciamento.
- projectId: identificador do projeto genômico.
- GC_content: conteúdo GC da sequência genômica.

[89] Atributos não obrigatórios: transcriptId,
[90] Positivo, se o gene estiver na sequência referência (posição códon start < posição códon stop); e negativo, se estiver na sequência complementar (posição códon start > posição códon stop na sequência referência).
[91] Ver DISCUSSÃO.
[92] Inclui as regiões codificadoras, não codificadoras e reguladoras do gene.
[93] Maiores detalhes em ANEXO IV – parte 2
[94] Atributos não obrigatórios: status, projectId

B.2. RELACIONAMENTOS

> HIT_OO, HIT_OP, HIT_PP:

Os *hits* foram modelados como relacionamentos entre sequências de aminoácidos comparadas no PCG[95]:

- HIT_PP: relacionamento apenas entre sequências de proteína,
- HIT_OP: relacionamento entre sequências de proteínas SwissProt e tORFs.
- HIT_OO: relacionamento apenas entre sequências tORFs.

HIT_PP

PROTEIN [0,n] PROTEIN

- Instâncias que não foram comparadas no PCG[96] não têm *hits* [0],
- Instâncias que não apresentam similaridade significativa com outra proteína comparada no PCG não têm *hits* [0].
- Instâncias da entidade PROTEIN podem ter *hits* com mais de uma proteína comparada no PCG [n].

HIT_OP

PROTEIN [0,n] ORF_T

- Instâncias da entidade PROTEIN que não foram comparadas com instâncias da entidade ORF_T[97] não têm *hits* [0].
- Instâncias da entidade PROTEIN[98] comparadas com instâncias da entidade ORF_T que não apresentam similaridade significativa não têm hits [0].
- Todas as sequências SwisProt foram comparadas com sequências tORFs.
- Instâncias da entidade PROTEIN podem ter *hits* com mais de uma tORF [n].

ORF_T [0,n] PROTEIN

- Instâncias da entidade ORF_T que não apresentam similaridade significativa com instâncias da entidade PROTEIN[99] não têm *hits* [0].

[95] Origem nas bases de dados Swissprot v.51.5 e Refseq v.21. Proteínas da versão v.33 da base RefSeq (que não existiam ainda na versão 21) não foram comparadas; isto é, não possuem hits. Ver 4.2.
[96] Proteínas da versão 33 da base de dados RefSeq (que não existiam ainda na versão 21, não foram comparadas; isto é, não possuem *hits*).
[97] Sequências RefSeq não foram comparadas com sequências torf.
[98] Sequências SwissProt.
[99] Sequências SwissProt.

– Todas as sequências tORF foram comparadas no PCG.
– Instâncias da entidade ORF_T podem ter *hits* com mais de uma proteína [n].

HIT_OO

ORF_T [0,n] ORF_T

– Instâncias da entidade ORF_T que não apresentam similaridade significativa com outra tORF não têm *hit* [0].
– Todas as sequências tORFs foram comparadas no PCG.
– Instâncias da entidade ORF_T podem ter *hits* com mais de uma tORF [n].

> PROTEIN CDS, GENE CDS, GS CDS

– A entidade CDS é associada à entidade PROTEIN através do relacionamento PROTEIN CDS;
– A entidade CDS é associada à entidade GENE através do relacionamento GENE CDS;
– A entidade CDS é associada à entidade GENOMIC SEQUENCE através do relacionamento GS CDS.

PROTEIN CDS

PROTEIN [0,1] CDS

– Instâncias da entidade PROTEIN cuja sequência CDS é desconhecida[100] não participam do relacionamento[0].
– Uma instância da entidade PROTEIN se relaciona com apenas 1 instância em CDS [1].

CDS [1,1] PROTEIN

– Toda instância de CDS se relaciona com a entidade PROTEIN e qualquer instância em CDS se relaciona com apenas uma instância em PROTEIN.

GENE CDS

GENE [1,n] CDS

– Toda instância de GENE se relaciona com CDS[101], e uma sequência gênica em GENE pode possuir 1 ou mais instâncias em CDS[102].

[100] Material e Métodos seção 3.2.
[101] Devido as instâncias da entidade GENE terem sua origem na base de dados NCBI-Gene.
[102] *Splicing* alternativo, por exemplo

CDS [0,1] GENE

- Instâncias de CDS cuja região esteja definida apenas para a sequência genômica (sem referência à região gênica) não participam do relacionamento [0].
- Genes não representados na base de dados NCBI-Gene não participam do relacionamento [0].
- Uma instância em CDS tem sua origem em apenas 1 gene em GENE [1].

GS CDS

GS [0,n] CDS

- Instâncias de GENOMIC_SEQUENCE que não tenham a predição de sequências CDS com anotação de sua posição[103] não participam do relacionamento [0].
- Uma instância em GENOMIC_SEQUENCE pode possuir várias regiões codificadoras em CDS [n].

CDS [0,1] GS

- CDSs que não possuem informação de sequência genômica[104] não participam do relacionamento [0].
- Uma instância de CDS tem sua origem em apenas 1 instância de GENOMIC_SEQUENCE [1]

> PROTEIN TAXON, ORF TAXON

PROTEIN TAXON

PROTEIN [1,1] TAXONOMY

- Uma sequência de proteína em PROTEIN tem sua origem, obrigatoriamente, em 1 e somente 1 táxon[105].

TAXONOMY [0,n] PROTEIN

- Podem existir táxons em TAXONOMY que não tenham sequências de proteína depositadas nos bancos de dados RefSeq e SwissProt [0].
- Um organismo em TAXONOMY pode possuir n sequências em PROTEIN [n].

[103] Ver Material e Métodos, seção 3.2. B. Sequências genômicas ou CDS com apenas mRNA como origem.
[104] CDSs que possuem apenas sequências mRNA de origem, por exemplo.
[105]Todas as sequências depositadas em bases de dados biológicos possuem um organismo de origem. No caso das bases RefSeq e SwissProt, a taxonomia utilizada é a do NCBI.

ORF TAXON

ORF_T [1,1] TAXONOMY

- Uma sequência tORF em ORF_T tem sua origem, obrigatoriamente, em 1 e somente 1 táxon.

TAXONOMY [0,n] ORF_T

- Podem existir táxons em TAXONOMY que não tenham sequências depositadas na base de dados RefSeq v.21[106] [0].
- Um organismo em TAXONOMY pode possuir n sequências em ORF_T [n].

> **ORF REGION**

ORF_T [1,1] GENOMIC_SEQUENCE

- Uma sequência tORF em ORF_T tem sua origem, obrigatoriamente, em 1 e somente 1 sequência genômica em GENOMIC_SEQUENCE[107].

GENOMIC_SEQUENCE [0,n] ORF_T

- Instâncias de GENOMIC_SEQUENCE que não tenham sido utilizadas para a predição de sequências ncORF não participam do relacionamento [0].
- Uma instância em GENOMIC SEQUENCE pode estar associada a várias instâncias em ORF_T [n].

[106] tORFS foram geradas a partir de ncORFs de genomas completos de procariotos da base RefSeq v.21
[107]Toda sequência torf foi obtida de uma única sequência ncorf, com origem numa única sequência genômica.

5. DISCUSSÃO

5.1. MODELAGEM

Modelos tentam representar parte de alguma realidade (um "mini mundo"). No entanto, em ciências, na grande maioria das situações não se conhece integralmente tal realidade e/ou não se consegue representá-la adequadamente.

Nas ciências biológicas, existe uma "realidade" que tenta explicar o conhecimento atual na área – baseada em experimentos *in vivo* realizados em situações particulares e controladas, e em hipóteses que estendem os resultados obtidos para situações mais abrangentes (que devem ser revalidadas experimentalmente). Grande parte desta "realidade" experimental encontra-se armazenada em bancos de dados. Estes dados, e a meta informação associada (anotação)[108], podem ser utilizados em procedimentos computacionais[109] para dar suporte à inferências e hipóteses com o poder de incrementar o conhecimento, realimentando o processo. Desta forma, tem-se também a "realidade" representada nos bancos de dados biológicos, que pode não conter todo o conhecimento experimental atual, mas pode conter hipóteses e inferências geradas por métodos computacionais com um nível de confiabilidade alto, porém ainda não testadas experimentalmente.

Com relação ao "mini mundo" considerado na modelagem conceitual do PWDB:

(A) Na ciência genômica, *parte* do conhecimento atual entende que um organismo:

- Possui um genoma (com uma ou mais entidades genômicas),
- As unidades genômicas possuem regiões gênicas codificadoras de proteína (e outras regiões codificadoras, não codificadoras, reguladoras e estruturais),
- Genes codificadores de proteína são transcritos em moléculas de mRNA,
- As moléculas de mRNA, após processamento e eliminação de introns[110], mantêm uma região linear e contínua codificadora de proteína[111],
- A região codificadora é traduzida, de acordo com um código genético, numa molécula de proteína.

[108] Algumas sugestões de leitura: [Stein L, 2001; Reeves *et al.*, 2009; Poptsova MS and Gogarten JP, 2010; Klimke *et al.*, 2011].
[109] Nestes procedimentos, a precisão e confiabilidade das anotações são críticas para a obtenção de resultados corretos.
[110] Há diferenças entre procariotos e eucariotos.
[111] Este processamento pode variar e gerar diferentes regiões codificadoras.

(B) Nos bancos de dados, as moléculas biológicas estão representadas por sequências[112] que podem ser manipuladas computacionalmente.

– Um organismo pode ter um genoma:
  - Totalmente sequenciado, ou
  - Parcialmente sequenciado, ou
  - Que não foi sequenciado.

– Regiões gênicas e seus mRNAs, assim como suas regiões codificadoras podem ser:
  - Conhecidas experimentalmente, ou
  - Identificadas apenas por procedimentos computacionais, e anotadas manualmente, ou
  - Identificadas apenas por procedimentos computacionais, e anotadas de forma automatizada sem interferência humana, ou
  - Inexistentes em bancos de dados.

– Proteínas[113] podem ser:
  - Geradas por tradução computacional – a partir de regiões codificadoras de seus genomas, genes ou mRNAs – e confirmadas experimentalmente.
  - Geradas por tradução computacional – a partir de regiões codificadoras de genomas, genes ou mRNAs – sem confirmação experimental.
  - Identificadas experimentalmente e depositadas em bancos de dados de sequências de proteína, apesar de suas sequências genômica, gênica ou mRNA não existirem em bancos de dados.

– Bancos de dados de sequências:
  - A cobertura do número de organismos e genomas aumenta imensamente a cada versão disponibilizada.
  - O escopo e funcionalidade varia de acordo com o conjunto de dados que armazenam, redundância, procedimentos automáticos e/ou manuais para curar os dados de sequências e suas anotações, dentre outras características.

[112] As sequências de DNA e RNA são representadas pelo alfabeto: $\sum = \{a, c, g, t, u\}$. As sequências de proteína são representadas pelo alfabeto $\sum'' = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$.
[113] A função da maior parte das sequências de proteína armazenadas em bancos de dados biológicos não foi determinada experimentalmente, mas extrapolada a partir de genes homólogos [Altenhoff AM *et al.*, 2012].

- São atualizados com grande rapidez e a quantidade de novos dados introduzida é enorme (a TABELA 5.1 apresenta um exemplo para a base RefSeq). Os procedimentos computacionais para curar estes dados evoluem com grande velocidade e o processo manual, indispensável, tenta acompanhar este processo.

Release date **Sep 09, 2013**, Number of Accessions Included: 41.958.567, Directory: complete

Number of **taxids**: 29.414

Number of Accessions and total length per molecule type:

Genomic:	4291237	310993467663
RNA:	4528216	8557926514
Protein:	33139114	11248966865

RefSeq Status Counts:

Status	RNA	Protein
Reviewed	150541	257778
Validated	49469	119667
Provisional	2120208	10480550
Predicted	23290	22379
Inferred	7539	7884
Model	2177169	2045119
Unknown	0	20205737

Release date **Jan 07, 2007**, Number of Accessions Included: 4.742.335, Directory: complete

Number of **taxids**: 4.079

Number of Accessions and total length per molecule type:

Genomic:	688455	72372319088
RNA:	819522	1492671478
Protein:	3234358	1144795927

RefSeq Status Counts:

Status	RNA	Protein
Reviewed	56655	89264
Validated	9616	121124
Provisional	402948	1742224
Predicted	14762	696301
Inferred	29	50
Model	323872	296967
Unknown	11640	4011

TABELA 5.1. Estatísticas de sequências da base de dados RefSeq em 2013 e 2007 (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-statistics/).

– Anotação: [Reeves *et al.*, 2009] fazem uma revisão dos vários aspectos de anotação: tipos, metodologias e disponibilidade. Além disso ressaltam os recentes esforços direcionados a anotações integradas, e discutem:

- Anotação de genomas e proteomas: sua organização, interpretação e integração,
- Ferramentas de anotação e bancos de dados: fornecem meios para a disseminação dos dados e compreensão de sua importância biológica.

Uma grande contribuição dos bancos de dados biológicos atuais é permitir a comparação de genomas de diferentes organismos. A comparação de sequências biológicas, buscando identificar similaridade entre elas, é uma ferramenta poderosa para auxiliar no desenvolvimento de hipóteses com o objetivo de caracterizar genes e outras sequências homólogas, e inferir relações estruturais, funcionais e evolutivas entre elas e seus organismos. Somente através de comparações é possível identificar características compartilhadas e/ou únicas de diferentes processos biológicos, organismos e grupos taxonômicos. A disponibilidade de sequências de organismos relacionados tem enriquecido bastante as análises genômicas comparativas.

Inúmeros algoritmos desenvolvidos para a resolução de problemas genômicos se propõem a resolver "todos os casos"[114]; muitas vezes utilizando métodos de aprendizado, *clustering* e *pipelines* automatizados sem interferência humana[115]. Desta forma, podem apresentar resultados com baixa confiabilidade nos casos de exceção, que podem se propagar em análises posteriores.

Poder utilizar recursos oriundos de diferentes fontes que integram diferentes tipos de informação biológica e métodos computacionais é uma boa opção para pesquisas pontuais e específicas, e é um procedimento que pode revelar novos dados, além de ressaltar possíveis incoerências ou erros. O esquema desenvolvido para o PWDB prevê este tipo de integração, com a vantagem de agregar o resultado de similaridade, já pré-calculado, resultante da comparação de milhares de sequências de proteína.

No projeto conceitual desenvolvido nesta tese foram considerados os conceitos listados em (A), e restrições necessárias foram introduzidas de acordo com as limitações listadas em (B). Dentre algumas situações, pode-se citar:

– Uma proteína sempre é codificada por um CDS, de acordo com (A), no entanto,

[114] Fato comum em computação.
[115] Dependem do conjunto de dados inicial e refletem o conhecimento da época de seu desenvolvimento. Sofrem interferência da representatividade de sequências e organismos nas bases de dados, e qualidade e confiabilidade de anotações.

existem restrições no relacionamento PROTEIN CDS, limitadas por (B);

- Um CDS sempre tem sua origem em um gene, de acordo com (A), no entanto, existem restrições no relacionamento GENE CDS, limitadas por (B);
- A existência de três atributos identificadores para a entidade PROTEIN é necessária devido às diferentes bases de dados de origem das sequências e, também, para a construção de referências cruzadas com outras bases de dados;
- O atributo *comp* da entidade PROTEIN só é necessário pois existem dois conjuntos de sequências de proteína distintos a serem considerados no momento da instanciação: o conjunto de proteínas comparadas no WCG, e o conjunto de proteínas de versões posteriores das bases de dados RefSeq e/ou Uniprot.
- No caso de projetos de sequenciamento de genomas completos, os atributos *projectId* e *status* são necessários para fornecer informações sobre as diferentes sequências genômicas e a fase do sequenciamento[116].

O conhecimento em (A) já está bastante sedimentado, mas não deixa de ser apenas um "recorte" da biologia molecular e genômica. Particularidades e exceções existem, e outras novas podem ser descobertas, sem que se altere o conceito primário – representado pelo MÓDULO CENTRAL do esquema conceitual da FIGURA 4.8 (TABELAS 5.2 e 5.3).

CONCEITO	REPRESENTAÇÃO
Classificação dos organismos	**TAXONOMY**
Sequência genômica	**GENOMIC SEQUENCE** $\sum = \{a, c, g, t\}$
Gene	**GENE** $\sum = \{a, c, g, t\}$
Sequência **C**o**D**ificadora de Proteína	**CDS** $\sum' = \{a, c, g, u\}$
Proteína	**PROTEIN** $\sum'' = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$
ncORF traduzida	**ORF_t** $\sum'' = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$
Genoma	Conjunto de todas GENOMIC SEQUENCE de um mesmo táxon
Transcrito (mRNA)	Representado como atributo em GENE
ncORF	Representado como atributo em ORF-region $\sum = \{a, c, g, t\}$

TABELA 5.2. Resumo de entidades e demais conceitos do esquema conceitual da FIGURA 4.8.

[116] ANEXO IV.1.

CONCEITO	RELACIONAMENTO
Associa uma sequência de proteína ao seu táxon	PROTEIN TAXON
Associa uma sequência torf ao seu táxon	ORF TAXON
Relaciona uma sequência de proteína a sua sequência cds	PROTEIN CDS
Posiciona uma sequência cds em sua sequência genômica	GS CDS
Relaciona uma sequência cds ao seu gene	GENE CDS
Posiciona uma ncorf em sua sequência genômica	ORF REGION
Relaciona duas sequências de proteína, comparadas par-a-par, cujo índice de similaridade atingiu o *cut-off* exigido	HIT_PP
Relaciona uma sequência de proteína[117] com uma sequência torf, comparadas par-a-par, cujo índice de similaridade atingiu o *cut-off* exigido	HIT_OP
Relaciona duas sequências torf, comparadas par-a-par, cujo índice de similaridade atingiu o *cut-off* exigido	HIT_OO

TABELA 5.3. Resumo de relacionamentos do esquema conceitual da FIGURA 4.8.

Os atributos que definem as entidades representadas no esquema conceitual foram concebidos para responder questões genômicas bastante amplas, e podem ser enriquecidos com informações complementares oriundas de bases de dados associadas ao MÓDULO CENTRAL.

5. 2. QUESTÕES FUNDAMENTAIS EM GENÔMICA

Vários conceitos baseados em estrutura, organização, funcionalidade e adaptação de genomas são utilizados em estudos evolutivos, desde a era pré-genômica, quando poucos métodos bioquímicos clássicos eram os únicos disponíveis, até os dias atuais, com novos e variados métodos experimentais de produção em larga escala e estrutura computacional robusta para interpretação e armazenamento dos resultados produzidos.

Uma discussão detalhada destes conceitos[118] está aquém da proposta desta tese, mas alguns deles serão mencionados, a seguir, de forma a tornar mais claro os requisitos do PCG e determinadas propriedades escolhidas para a definição dos elementos do esquema conceitual.

Questões fundamentadas nestes conceitos podem ser respondidas completa ou parcialmente através de consultas diretas a um sistema de banco de dados implementado a partir

[117] No PCG, sequências SwissProt
[118] Sugestões para uma revisão geral: [Buschman F, 2001] e [Gregory TR, 2005]. Dentre artigos científicos atuais pode-se sugerir [Gillis J and Pavlidis P, 2011] e outros que serão citados ao longo da Discussão.

do esquema da FIGURA 4.8, ou através de protocolos que utilizem resultados intermediários (fornecidos pelo sistema) como *input* para novas consultas/procedimentos, ou ainda, como *input* para ferramentas externas ao sistema.

Além disso, pode-se gerar, por exemplo, listas com prováveis homólogos – para proteínas de interesse – enriquecidas com informações variadas como anotações provenientes de mais de uma base de dados, posições relativas ocupadas nos genomas homólogos, distâncias evolutivas[119] entre eles, o conteúdo GC das regiões gênicas, identificadores das proteínas, transcritos, genes e sequências genômicas etc., facilitando e ampliando novas consultas com as proteínas selecionadas.

Homologia (Ortologia e Paralogia) e Analogia

A duplicação de DNA é uma das principais forças que direcionam a evolução dos organismos pois cria o material genético "cru" para a seleção agir e modelar. A redundância, que permite a diversificação, ocorre com relação a um único gene, regiões genômicas ou todo o genoma; e a compreensão deste processo requer a aplicação de abordagens comparativas.

O conceito de **homologia** por muitos anos foi utilizado apenas por biólogos evolutivos, e, atualmente, tem sido muitas vezes mal compreendido e indevidamente utilizado quando explorado fora do seu contexto evolutivo. É um conceito com enorme potencial para estudar a evolução de espécies, genomas e genes [Descorps-Declère S, 2008]. A duplicação de genes é uma importante fonte de inovação funcional, e a definição de famílias de genes e compreensão das complexas relações entre seus membros é fundamental em estudos evolutivos, fisiológicos e de adaptação a ambientes diversificados e hostis, por exemplo.

Genes homólogos (FIGURA 5.1) – pares de genes descendentes de um ancestral comum – são basicamente classificados como **ortólogos** (se divergiram após um evento de especiação) ou **parálogos** (se divergiram após um evento de duplicação, antes ou depois de um evento de especiação) [Fitch WM, 1970; 2000]. Estes são conceitos-chave em genômica evolutiva, e uma distinção clara entre eles é relevante em uma ampla gama de contextos como a evolução de genomas, função de produtos gênicos, redes celulares e anotação funcional [Koonin EV, 2005].

[119] Pode ser medida por diferentes métodos

Analogia, por outro lado, é definida como a relação entre dois caracteres quaisquer, descendentes de caracteres ancestrais não relacionados entre si, cuja similaridade se origina por convergência (Fitch WM, 1970; 2000). Existe um considerável número de enzimas análogas[120], com atividades e especificidades similares sem uma origem evolutiva compartilhada, ou similaridade de sequência, por exemplo.

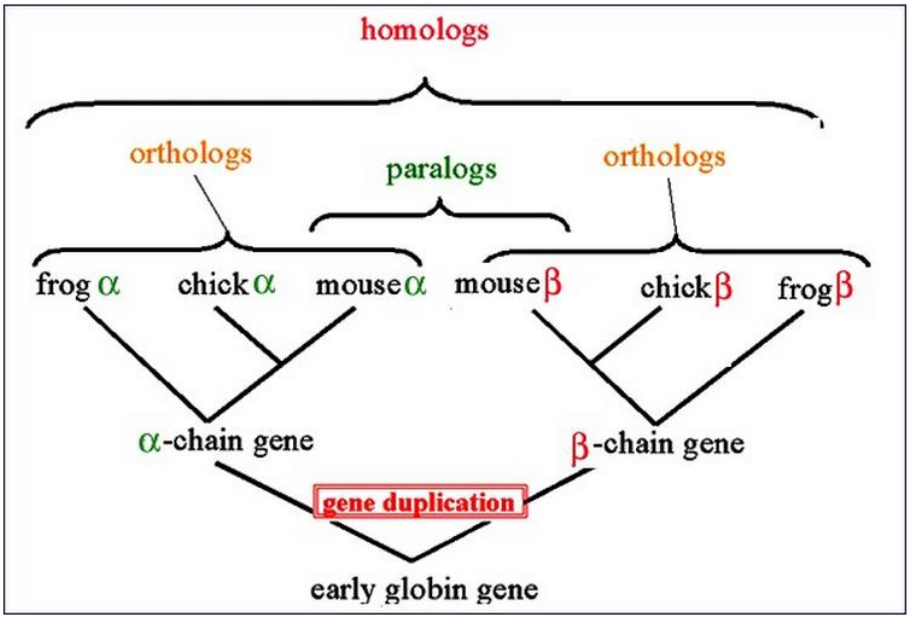


FIGURA 5.1. Exemplos de homologia, ortologia e paralogia
(BLAST Glossary, http://www.ncbi.nlm.nih.gov/books/NBK62051/)

[Fulton DL *et al.*, 2006] citam como um equívoco comum a utilização do termo ortólogo com o significado de "genes funcionalmente equivalentes em diferentes espécies". Mesmo que muitos concordem que ortólogos tendem a apresentar funções similares, isto não é um requerimento para ortologia. Para os autores, várias situações em que pesquisadores tentam identificar ortólogos, num estudo genético ou genômico, o que eles realmente desejam é identificar o subgrupo de ortólogos que são especificamente equivalentes funcionalmente.

Existe uma discussão sobre o assunto – a conjectura de ortólogos [Altenhoff AM, 2012], que postula que genes ortólogos são funcionalmente mais semelhantes do que genes parálogos. Uma parte considerável de pesquisadores apoia a hipótese de que após uma duplicação a função da proteína muda rapidamente originando parálogos com diferentes funções, enquanto que ortólogos tendem a reter a função ancestral.

[120] Ver [Guimarães ACR, 2010].

A identificação automática de genes ortólogos, assim como de famílias de parálogos, tem sido um componente essencial de várias aplicações de bioinformática, desde anotação de novos genomas à priorização de alvos para experimentos, e é comumente realizada a partir de sequências de proteína. É um ponto crítico para a construção de uma classificação evolutiva robusta de genes, estudar a complexidade das relações entre sequência e função e, consequentemente, uma anotação funcional confiável [Fulton DL *et al.*, 2006].

Erros na predição de verdadeiros ortólogos podem ter um impacto bastante negativo em análises posteriores (incluindo genômica funcional e análises proteômicas), e tem levado a um interesse cada vez maior em métodos de predição de ortólogos de alta qualidade. Análises de genomas completos indicam que muitas famílias de genes (essencialmente parálogos) foram formadas antes da divergência da maioria das espécies atualmente comparadas. Consequentemente, ortólogos são tipicamente mais similares entre si do que entre os outros genes do genoma, e é por isso que a similaridade de sequências é geralmente utilizada para inferior ortologia de genes entre duas ou mais espécies. Porém, se um gene não está presente no conjunto de dados de um organismo comparado (devido a sequência incompleta do genoma ou perda de genes, por exemplo), métodos computacionais podem predizer incorretamente um parálogo como um ortólogo (Fulton DL *et al.*, 2006).

Estudos de genômica comparativa utilizam, muitas vezes, genomas incompletos, especialmente em grandes projetos de sequenciamento de eucariotos. Além disso, a perda gênica é uma importante força direcionando a evolução de bactérias [Reference Genome Group of the Gene Ontology Consortium, 2009]. Portanto, é importante estar atento ao fato de que muitos dos atuais bancos de dados de ortólogos devem conter falso-positivos devido a limitações dos métodos de inferência.

[Chen and Zhang, 2012] analisaram estudos funcionais de homólogos com sequências de proteína idênticas e identificaram a existência de viés experimental, erros de anotação, e inferências funcionais baseadas em evidência rotulada como experimental (em GO), quando na realidade o gene que possuía a evidência experimental não era o gene anotado, mas seu homólogo.

Eventos de transferência lateral de genes (*Lateral Gene Transfer* – LGT) e casos de proteínas com múltiplos domínios são fatores bastante complicadores para algoritmos que

dependem da correta distinção entre ortólogos e parálogos. [Dalquen *et al.*, 2013] compararam a precisão de métodos de inferência de ortologia e concluíram que LGT diminui drasticamente a precisão de todos os métodos analisados; e relataram com preocupação que nenhum dos programas investigados incorporava um método específico para a detecção de LGT.

Sintenia[121] e Colinearidade[122]

Vários métodos para detecção de homólogos utilizam o conceito de sintenia e o contexto genômico em que o gene está inserido. Comparações entre genomas eucarióticos relacionados revelaram vários graus nos quais genes homólogos permanecem nos cromossomos correspondentes e exibem conservação de ordem (colinearidade) durante a evolução [Koonin EV, 2005].

O multi-alinhamento de regiões cromossômicas colineares (referidas como blocos colineares) pode revelar antigos eventos de duplicação de genomas inteiros (*Whole Genomic Duplication* – WGD), duplicação regionais e rearranjos, e relações cromossômicas complexas [Yupeng *et al.*, 2012]. Padrões de sintenia e colinearidade podem fornecer *insights* sobre a história evolutiva de genomas e permitir análises subsequentes potencialmente úteis. Em muitas análises consideram-se genes "âncora", que estão localizados em posições colineares nestes blocos, e genes "não-âncora", que são mantidos para obtenção de ganhos/perdas genéticas ou transposição. Os genes de ancoragem são mais susceptíveis de serem homólogos do que os genes "não-âncora" (Jun *et al.*, 2009, Casneuf *et al.*, 2006).

[Kassahn *et al.*, 2009] comentam, com relação à evolução dos cordados, que grupos de genes co-duplicados podem ser remanescentes de antigos eventos de duplicação em pequena escala (envolvendo segmentos cromossômicos ou *clusters* de genes) que ocorreram em diferentes momentos evolutivos. E que combinações recentes de grupos co-duplicados distintos em diferentes regiões cromossômicas podem ser, provavelmente, o resultado de rearranjos de segmentos genômicos, incluindo grupos sintênicos de genes.

[Henricson *et al.*, 2011] demonstraram que os introns podem manter a sua posição por longos períodos evolutivos. Para os autores, pode ser possível utilizar a conservação das

[121] Regiões de genomas distintos com considerável similaridade de sequência e probabilidade de descendência de um ancestral comum.
[122] Colinearidade, uma forma mais específica de sintenia, considera a conservação da ordem dos genes.

posições dos introns como um fator de discriminação na detecção de ortologia, e que as posições dos introns em genes ortólogos tendem a ser mais conservadas do que em genes não ortólogos.

[El-Mabrouk and Sankoff, 2012] apresentam uma revisão de estudos de rearranjos de genomas[123], e de como lidar com genes duplicados. Apresentam as questões:

- Como genomas atuais evoluíram a partir de um genoma ancestral comum?
- Quais os cenários evolutivos mais realísticos para explicar a ordem de genes observada?
- Qual o conteúdo e estrutura dos genomas ancestrais?

Os autores tratam da identificação de ortólogos, parálogos e blocos sintênicos; e consideram três níveis de organização de genes:

- Famílias de genes (evolução através de duplicação, perda e especiação),
- *Clusters* de genes (evolução através de duplicações em série),
- Genômica (todos os tipos de eventos de rearranjo, incluindo WGD).

Genes Órfãos / Genes Únicos (Taxonomicamente Restritos)

Genes órfãos tem sua origem principalmente em eventos de duplicação, e uma das possibilidades é que eles representem genes específicos de espécie. Segundo [Mazza *et al.*, 2009], trabalhos recentes têm demonstrado que genes órfãos em *Drosophila* e primatas evoluem três a quatro vezes mais rápido do que a média de genes. Os autores citam que a função destes genes é muitas vezes mal caracterizada, e que eles possuem propriedades distintas, como alta especificidade tecidual, rápida evolução e pequeno tamanho. Comentam ainda que dependendo da sensitividade e especificidade dos métodos utilizados para identificar genes ortólogos, a fração de genes sem ortólogos entre espécies é variável; e que a qualidade da montagem do genoma pode interferir nos resultados. Em alguns casos a divergência de sequências entre espécies pode ser tão grande que uma possível ortologia entre genes pode se tornar indetectável.

A identificação de genes restritos em diferentes níveis taxonômicos tem valor prático e científico. Proteínas específicas de linhagens (*strain*), espécie e gênero podem fornecer *insights* sobre critérios que definem um organismo e suas relações com organismos próximos. Informação sobre a presença ou ausência de genes é uma poderosa ferramenta para adquirir conhecimento sobre metabolismo, patogenicidade, fisiologia e comportamento em diferentes

[123] Genômica comparativa, baseada na representação dos genomas como sequências ordenadas de assinaturas de genes [El-Mabrouk and Sankoff, 2012].

organismos [Eisen and Fraser, 2003; Tatusov *et al.* 1997; Siew and Fischer, 2003]. Podem auxiliar na discriminação de linhagens patogênicas e não-patogênicas e fornecer uma lista reduzida de alvos diagnósticos para serem validados em laboratório.

Os resultados de [Mazumder *et al.*, 2005] (foram considerados apenas genomas de procariotos) mostraram que a maioria das proteínas únicas não apresentavam uma anotação funcional (estavam anotadas como hipotéticas), e as restantes, em sua maioria, apresentavam alguma relação com patogênese ou virulência, ou eram derivadas de fagos. Alguns fatos levantados no estudo:

- O número de genes únicos codificados em organismos específicos pode depender da definição de "único", dos parâmetros do método e do banco de dados utilizado.
- Sequências podem ser incorretamente identificadas como únicas devido à ausência de genomas próximos para comparação, e também devido a existência de múltiplos genomas da mesma espécie ou gênero.
- A identificação de genes específicos de espécie ou gênero poderá ocorrer com maior confiança quando existir uma maior representatividade de genomas nos bancos de dados para comparação.

Natureza Modular das Proteínas, Domínios Proteicos, Fusão/Fissão [124]

A estrutura modular das proteínas pode ser uma limitação para vários métodos de comparação de sequências e, portanto, a detecção e posicionamento de módulos, definindo uma estrutura de domínios, pode ajudar a remontar a história evolutiva da proteína, identificando eventos de duplicação, reorganização e fusão. Inúmeras proteínas são compostas por uma combinação de domínios discretos, associados a funções específicas que surgem em diferentes momentos evolutivos. O surgimento de novos domínios está relacionado com a diversificação e adaptação funcional da proteína. Como novos domínios surgem e como evoluem ainda é uma intensa área de pesquisas.

Segundo [Toll-Riera and Mar Albà, 2013], as proteínas tendem a ganhar domínios ao longo do tempo. Muitas são compostas por domínios de diferentes idades e as regiões correspondentes aos domínios adquiridos mais recentemente tendem a evoluir mais

[124] Jachiet *et al.*, 2013; Dimitriadis *et al.*, 2011; Adam *et al.*, 2010; Reid *et al.*, 2010; Durrens *et al.*, 2008; Kummerfeld SK and Teichmann SA, 2005; Long M, 2000; Rentzsch R and Orengo CA, 2013; Forslund K, 2011; Kummerfeld SK, Teichmann AS, 2009; Bagowski CP,2010.

rapidamente. A identificação de domínios com origem evolutiva recente é crucial para o entendimento das adaptações específicas de espécie e específicas de linhagem, mas são domínios ainda pobremente caracterizados. Os autores compararam propriedades evolutivas de domínios de proteínas humanas de diferentes idades: específicas de mamíferos, específicas de vertebrados e mais antigas. E encontraram que quando os domínios de diferentes idades se combinam na mesma proteína, o domínio mais recente tende a evoluir bem mais rápido do que os domínios mais antigos, reforçando a idéia de que o tempo decorrido desde a origem de uma sequência determina em grande parte sua taxa evolutiva atual.

Comparações de proteínas com múltiplos domínios podem ter ótimos resultados estatísticos de similaridade com proteínas com um único domínio (confirmando que possuem um domínio comum), mas uma análise detalhada da cobertura do alinhamento vai mostrar que a similaridade ocorre apenas na região do domínio comum.

Fusão gênica é um processo evolutivo no qual genes inicialmente separados se fusionam numa única ORF, a qual é expressa como uma cadeia de proteína multi-domínio [Reid *et al.*, 2010][125]. Através da detecção destes eventos tenta-se inferir que as proteínas não fusionadas são funcionalmente relacionadas (i.e. participam de um processo biológico comum: seus produtos gênicos interagem como parte de um complexo proteico com mais de uma sub-unidade ou numa via metabólica comum). Procedimentos computacionais que identificam eventos de fusão com o objetivo de predição de associações funcionais costumam usar sequências de proteína completas ou famílias de domínios (fusão gênica e fusão de domínios), e os dois tipos de enfoque são compatíveis e podem ser combinados. No caso de fusão de domínios, a anotação de famílias de domínios (e.g. Pfam[126]) é usada para identificar domínios distintos em diferentes proteínas de um genoma, que ocorrem fusionados em uma única proteína de outro genoma.

[Kassahn *et al.*, 2009] compararam a expressão e arquitetura de domínios de WGD em *Danio rerio* com seus ortólogos de cópia única em *Mus musculus* e encontraram vários exemplos que suportam um modelo de neo funcionalização:

- Duplicatas-WGD adquiriram novos domínios proteicos mais frequentemente do que genes de cópia única,

[125] O artigo cita alguns métodos para detecção de eventos de fusão (gene/domínio).
[126] ANEXO IV.3

- Alterações pós-WGD no nível de regulação gênica foram mais comuns do que alterações no nível de proteína,
- Concluíram que a consequência mais significante de WGD para a evolução de vertebrados foi permitir um controle da regulação do desenvolvimento mais especializado, via aquisição de novos domínios de expressão espaço-temporal.

No estudo, foram extraídos tripletos de proteínas peixe-humano quando um único *locus* humano possuía uma ou mais "âncoras" genômicas dentro do genoma do peixe. Estas proteínas do peixe compartilhavam significante similaridade de sequência com a proteína humana e estavam localizadas em regiões genômicas que compartilhavam pelo menos três homólogos próximos peixe-humano, sugerindo que as duas proteínas do peixe tinham sido originadas de um cromossomo ancestral compartilhado através de duplicação, provavelmente como resultado de WGD. Este tipo de procedimento utilizado pelos autores pode ajudar em predições de sintenia para duplicatas de WGD mantidas no genoma.

Ainda de acordo com [Kassahn *et al.*, 2009], a natureza das alterações é semelhante tanto após eventos de especiação quanto de duplicação, mas mudanças na arquitetura de domínios acontecem com uma frequência maior após duplicações de genes. Os autores acreditam que eventos de inserção/deleção de domínios parecem gerar uma maior perturbação na estrutura de domínios existente numa proteína quando ocorrem em posições internas, e que este fato talvez favoreça, em grande parte, a tendência destes eventos ocorrerem com maior frequência em torno das regiões terminais de proteínas

Usando-se, por exemplo, os recursos do banco de dados Pfam pode-se traçar a evolução de famílias de domínios específicos. [Buljan and Bateman, 2009] mostram que a trajetória da superfamília de imunoglobulinas pode ser traçada por mais de 500 milhões de anos, desde a sua expansão para uma das maiores famílias do genoma humano. Esta família de proteínas teve sua origem em receptores de membrana de animais primitivos como esponjas (Porifera); e sua estrutura e sequência pode ser recuperada[127] analisando-se a manutenção e variação de diferentes domínios encontrados em anticorpos do sistema imunológico, e em proteínas neurais e musculares de humanos.

[127] Desde a divergência ocorrida durante a evolução dos vertebrados.

Fatores de transcrição são um bom exemplo da natureza modular e embaralhamento (*shuffling*) de proteínas, ampliando o controle da regulação gênica e desenvolvimento de organismos [Riechmann *et al.*, 2000].

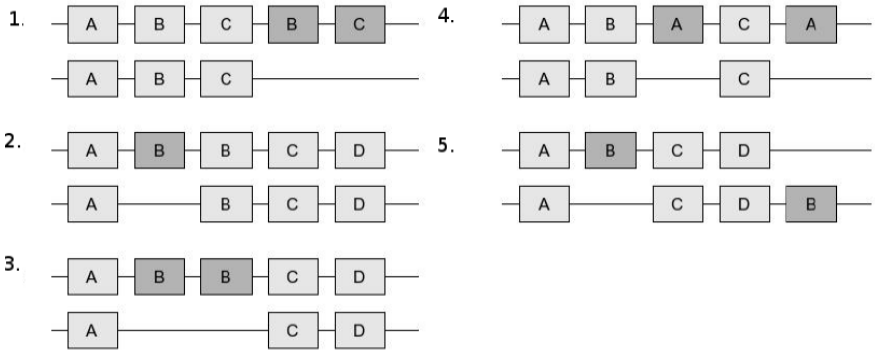
De acordo com [Sjölander *et al.*, 2011], métodos de identificação de ortólogos geralmente não consideram alterações na arquitetura de domínios[128], que tendem a modificar a função de proteínas. Os autores apresentam uma revisão de questões envolvidas na predição de ortólogos quando o conjunto de dados utilizado inclui sequências com estruturas heterogêneas de arquitetura de domínios, com atenção particular aos métodos desenhados para aplicações em larga escala. A FIGURA 5.2 mostra como as combinações de domínios foram anotadas no trabalho.

[Sjölander *et al.*, 2011] concluem que:

- Ortólogos exibem uma maior conservação da arquitetura de domínios do que parálogos.
- Interpretaram este fato como a indicação de uma pressão seletiva mais forte sobre ortólogos para manter a arquitetura de domínios requerida para as proteínas executarem uma função específica.
- Ortólogos, assim como os parálogos mais próximos, geralmente possuem arquiteturas de domínios bastante similares, mesmo em grandes distâncias evolutivas.
- A alteração mais comum observada em arquiteturas de domínios, tanto em pares de ortólogos quanto de parálogos, envolve inserção/deleção de novos domínios, enquanto que embaralhamento e duplicação/deleção de segmentos são menos frequentes.
- É possível inferir que ortologia está acompanhada de uma forte pressão para manter a arquitetura de domínios, e que a ordem específica de domínios parece ser importante para a função executada por um grupo de proteínas ortólogas.

[128] Arquitetura de domínios, refere-se ao arranjo sequencial de domínios ao longo de uma sequência de proteína. Também referido como arranjo de domínios ou ordem de domínios, significa especificamente o arranjo sequencial de domínios conhecidos Pfam-A ao longo de uma proteína, no sentido N- para C-terminal [Sjölander *et al.*, 2011].

Todos os alinhamentos que não foram classificados como idênticos ou totalmente diferentes foram anotados como segue:



Anotações de eventos de troca de domínios. Domínios alinhados estão sombreados em cinza claro e domínios não alinhados em cinza escuro.

Todos os casos de domínios não alinhados receberam uma das cinco anotações:

1. *Duplicação/deleção de segmento*. Um segmento de dois ou mais domínios foi duplicado nas proximidades ou perdido,
2. *Diferença de repetição*. A primeira proteína possui um ou mais domínios B do que a outra proteína. Como o domínio não alinhado está localizado próximo a um domínio alinhado do mesmo tipo, o domínio não alinhado é anotado como uma diferença de repetição,
3. *Inserção/deleção de um novo domínio*. A primeira proteína tem dois domínios B não alinhados. Ambos serão anotados como inserção/deleção de um novo domínio, pois a outra proteína não possui este tipo de domínio.
4. *Inserção/deleção de um domínio existente*. A primeira proteína possui dois domínios A não alinhados. Como a outra proteína tem um domínio deste tipo, os domínios não alinhados na primeira proteína serão anotados como inserção/deleção de um domínio existente,
5. *Embaralhamento de domínios* (*shuffling*). Ambas proteínas possuem domínios B não alinhados. Nenhum deles ocorre próximo a um domínio B alinhado e portanto não podem ser anotados como repetição. Estes domínios serão anotados como um embaralhamento.

FIGURA 5.2. Adaptação da Figura 1 de [Sjölander K *et al.*, 2011].

5.3. VALIDAÇÃO DO MODELO

Os objetos representados num esquema conceitual são definidos com base num conjunto semântico, o qual pode ser entendido intuitivamente por desenhistas e usuários do sistema.

O diagrama ER de um esquema conceitual é uma ferramenta que permite a descrição de consultas e procedimentos sem envolvimento com o esquema lógico. Este fato permite que usuários leigos em linguagens de bancos de dados possam contribuir desde as fases iniciais do projeto do sistema, auxiliando os desenvolvedores na compreensão do problema e implementação dos procedimentos.

A seguir são apresentados alguns exemplos de consultas simples e procedimentos mais complexos com o objetivo de demonstar mais claramente como inúmeras questões podem ser detalhadas pelos próprios usuários a partir do diagrama da FIGURA 4.8.

CONSULTAS SIMPLES:

1. *Contar proteínas no banco de dados que tenham sido comparadas?*
   - Contar ocorrências na entidade PROTEIN usadas em comparações (atributo *comp* = 1).
2. *Proteínas do PCG que possuem uma sequência genômica origem?*
   - Checar ocorrências em PROTEIN, usadas nas comparações (*comp* = 1), que tenham um elemento associado na entidade GENOMIC SEQUENCE.
3. *Quantos genomas de um dado grupo taxonômico (e.g. Vertebrata) estão representados no banco de dados?*
   - Inicialmente é necessário recuperar o *taxonId* para o grupo na entidade TAXONOMY e, então, contar todos os nodos com *rank* espécie (ou o menor *rank*, se houver), pertencentes a esse grupo, que possuam sequências genômicas associadas na entidade GENOMIC SEQUENCE.
4. *Quantas proteínas, com sequências genômicas, pertencem a um dado grupo taxonômico?*
   - É similar ao item anterior, mas é necessário considerar todas as sequências genômicas (na entidade GENOMIC SEQUENCE) de cada *taxonId* das espécies (ou o menor rank) pertencentes ao grupo taxonômico desejado em TAXONOMY; e contar todas as ocorrências na entidade PROTEIN, que estejam associadas a um elemento da entidade GENOMIC SEQUENCE pertencente a esse grupo taxonômico.
5. *Retornar todos os hits (com proteínas) para uma dada protein X, com e-value abaixo de um cut-off*:
   - Fazer uma busca nas ocorrências do relacionamento HIT_PP, restrita a *e-values* < *cut-off* escolhido, e listar aquelas em que *query_gi* ou *subject_gi* seja igual a X.

CONSULTAS COMPLEXAS*:*

Seguem dois exemplos de procedimentos para a identificação de genes taxonomicamente restritos e parálogos. Nestes exemplos, a análise considerará apenas proteínas oriundas de genomas completos (atributo *status* = complete na entidade GENOMIC SEQUENCE), para garantir comparações entre proteomas completos.

A. Identificação de proteínas restritas a uma espécie ou a um grupo taxonômico

Resumidamente deseja-se identificar proteínas, do proteoma completo ("predito") de uma espécie, que não tenham similaridade com nenhuma proteína de outro proteoma; ou identificar proteínas dos proteomas completos das espécies de um grupo taxonômico T que tenham similaridade com proteínas de todas as espécies do grupo T, e não tenham similaridade com nenhuma proteína fora do grupo taxonômico T.

A.1 Genes restritos a uma espécie (ou o menor *rank*, que pode estar no nível taxonômico de espécie, *strain* ou sub-*strain*).

1. Obter o *taxonId* do organismo de interesse na entidade TAXONOMY (todas as sequências genômicas desse *taxonId* devem ter *status* = complete),
2. Para esse *taxonId*, considerar todas as sequências genômicas da entidade GENOMIC SEQUENCE que tenham elementos associados e comparados em PROTEIN (*comp* = 1),
3. Seja SeqGen o conjunto de sequências genômicas definidas em 2.
4. Para o conjunto SeqGen, identificar, dentre todos os seus elementos associados na entidade PROTEIN, aqueles que não tenham *hit* (e-value < *cut-off*) com sequências de proteína não associadas com o conjunto SeqGen.

A.2 Genes restritos a um grupo taxonômico (gênero, por exemplo)

1. Obter os *taxonId* das espécies (ou nó folha inferior) do genêro X na entidade TAXONOMY (todas as sequências genômicas, para cada *taxonId*, devem ter *status* = complete),
2. Seja T o conjunto de *taxon_id* das espécies pertencentes ao gênero X.
3. Para cada *taxonId* pertencente a T,
   Seja SeqGen o conjunto de todas as sequências genômicas da entidade GENOMIC SEQUENCE desse *taxonId*, que tenham elementos associados e comparados em PROTEIN (*comp* = 1).
4. Seja SeqGlobal o conjunto de SeqGen.
5. Para o conjunto SeqGlobal,
   Identificar, dentre todos os seus elementos associados na entidade PROTEIN, aqueles que

satisfaçam as duas condições abaixo:

- NÃO TENHAM *hit* (e-value < *cut-off*) com sequências de proteína não associadas em SeqGlobal,
- TENHAM *hit* (e-value < *cut-off*) com pelo menos uma sequência de proteína associada em cada SeqGen.

Como exemplo, pode-se citar o gênero *Escherichia* que possui a espécie *E. coli* (dentre várias outras). Essa espécie possui a linhagem (*strain)* O157:H7 (dentre várias outras), e *E. coli* O157:H7 possui várias sub-*strains*. Através dos procedimentos descritos acima, pode-se obter subconjuntos iniciais, por exemplo, para:

- Genes únicos de *Escherichia* (ocorrem em todas as espécies/linhagens do gênero e não ocorrem fora do genêro);
- Genes restritos à espécie *E. coli* (não ocorrem em outras espécies de *Escherichia*);
- Genes restritos à uma linhagem de *E. coli* (O157:H7, ou qualquer outra),
- Genes compartilhados e únicos de linhagens patogênicas e não patogênicas.

É indicado que haja uma segunda etapa com procedimentos que comparem o resultado obtido em A.1 ou A.2 com todas as outras proteínas da entidade PROTEIN (que não tenham sido utilizadas na primeira fase), para confirmar se não existem proteínas similares em organismos que não foram considerados por não possuírem genomas completos.

B. Identificação de Genes Parálogos

Resumidamente, deseja-se identificar proteínas do proteoma completo (“predito”) de um organismo que tenham similaridade com outras proteínas do mesmo proteoma.

1. Obter o *taxonId* do organismo de interesse na entidade TAXONOMY (todas as sequências genômicas desse *taxonId* devem ter *status* = complete),
2. Para esse *taxonId*, considerar todas as sequências genômicas da entidade GENOMIC SEQUENCE, que tenham elementos associados e comparados em PROTEIN (*comp* = 1),
3. Seja SeqGen o conjunto de sequências genômicas definidas em 2.
4. Para o conjunto SeqGen, identificar, dentre todos os seus elementos associados na entidade PROTEIN, grupos de proteínas que tenham *hits* entre si (e-value < *cut-off*).

Nos dois exemplos acima, os possíveis resultados representam uma primeira etapa na identificação de genes restritos e parálogos. Diferentes valores de *cut-off* devem ser testados, e avaliações adicionais.

C. Outros Procedimentos

A seguir, são apresentados outros exemplos de procedimentos que podem ser especificados utilizando o esquema conceitual da FIGURA 4.8, como nos casos A. e B. descritos acima. Vale frisar que muito do poder de consulta do esquema provém de suas conexões com as bases de dados associadas, e que cada consulta/procedimento pode ser apenas a primeira fase de protocolos com múltiplas etapas. Resultados intermediários ou finais podem ser combinados de forma a atingir uma resposta mais específica.

- Famílias de Parálogos: podem ser comparadas entre diferentes proteomas identificando, por exemplo, perdas, duplicações e diversificação nas famílias. Consultas adicionais:
  - Recuperar os domínios das proteínas na base de dados Pfam,
  - Utilizar a posição dos domínios Pfam para um entendimento da estrutura de domínios nos membros da família. Isto pode permitir o agrupamento de parálogos mais similares dentro da família (resultantes de ciclos independentes de duplicação). Esses grupos podem facilitar a comparação da história evolutiva da família em diferentes organismos (expansão/redução e diversificação).
  - O atributo *region* da entidade CDS pode ser utilizado para mapear a família de genes em seu genoma.
  - O alinhamento em HIT_PP pode se estender apenas ao longo de um domínio que define a família (comum a todos os parálogos do proteoma), e não ao longo do comprimento total das sequências comparadas.
  - Adicionar anotação GO da base Gene Ontology, e *definition* da entidade PROTEIN, para comparar anotações funcionais; e números EC da base KEGG no caso de enzimas.
  - O resultado pode fornecer indícios para uma busca mais detalhada de proteínas não identificadas no proteoma de um dos organismos (com uma função específica, por exemplo), mas que esteja presente no proteoma de outro organismo como integrante de uma família gênica compartilhada entre os dois.
  - Dentre outros.

- Ortólogos: pode-se utilizar uma sequência representativa[129] para sub-grupos distintos[130] dentro de uma família de genes parálogos (partindo do caso acima). A análise de similaridade, auxiliada por informações de parálogos, pode facilitar a identificação de ortólogos entre diferentes proteomas completos (todas as sequências genômicas na entidade GENOMIC SEQUENCE de cada proteoma devem ter *status* = complete). Adicionalmente, pode-se considerar nas análises:
  - A cobertura dos alinhamentos (relacionamento HIT_PP), que deve se estender pelo comprimento total das sequências (ao contrário de parálogos, como citado anteriormente);
  - Adicionar anotação GO da base Gene Ontology, e *definition* da entidade PROTEIN, para comparação das anotações funcionais; e números EC da base KEGG no caso de enzimas.
  - Confirmar os domínios Pfam e a estrutura de domínios das proteínas.
  - O atributo *region* da entidade CDS permite:
    - Analisar a conservação de número e posição de introns,
    - Acessar a sequência genômica e região promotora dos genes para comparação,
  - Dentre outros.

- Colinearidade: para os casos descritos acima, e várias outras consultas, outros fatores importantes podem ser analisados:
  - A vizinhança dos genes sob avaliação. Pode-se utilizar o atributo *index* de GENE.
  - Confirmar a existência de *clusters* de genes e a ordem dentro do *cluster*, e comparar essa estrutura com outros proteomas.
  - Avaliar o contexto genômico (ver abaixo).
  - Dentre outros.

- Contexto genômico
  - Várias análises genômicas avaliam o conteúdo GC de genes e genomas,
  - Genes, assim como regiões genômicas, com uma composição anormal do conteúdo GC, com relação ao valor médio de genes e do genoma é um dos parâmetros utilizado em procedimentos para identificação de transferência

[129] Ou gerar uma sequência consenso. Existem ferramentas para isto.
[130] Parálogos mais similares

lateral de genes (LGT) e ilhas genômicas (patogenicidade, resistência a antibióticos, dentre outras), por exemplo.

- Alguns tipos de elementos genômicos estão associados com regiões que apresentam conteúdo GC diferenciado.
- A entidade GENE possui a informação do conteúdo GC de genes (atributo *GC_content*);
- A entidade GENOMIC SEQUENCE possui a informação do conteúdo GC de sequências genômicas (atributo *GC_content*).
- Para análise de transferência lateral de genes, pode-se utilizar como parâmetros:
  - O conteúdo GC de genes individualmente e de *clusters* de genes, quando for o caso.
  - Colinearidade de genes em *clusters* (os atributos *index* e *strand* da entidade GENE podem auxiliar).
  - A existência de genes ou *clusters* de genes similares no genoma analisado e sua ausência em grupos taxonômicos próximos, porém existentes em organismos mais distantes.
  - Inversamente, a existência de genes ou *clusters* de genes similares no genoma analisado e em grupos taxonômicos vizinhos.
- O atributo *region* da entidade CDS serve como ponto de referência para o acesso a regiões genômicas específicas, de forma que as questões descritas acima e várias outras (como por exemplo, a comparação de regiões promotoras de grupos específicos de genes) possam ser avaliadas.
- Dentre outros.

- Vias Bioquímicas
  - Supondo uma via bioquímica de interesse, pode-se utilizar como referência a mesma via já bem estudada num organismo modelo, por exemplo. E tentar reconstruir essa via em outros organismos de interesse.
  - Analisar proteínas similares para cada etapa de uma via em diferentes organismos, considerando as anotações funcionais oferecidas pelo sistema. Restringir o conjunto de dados utilizando os atributos dos elementos do esquema, associados às anotações. No caso de enzimas, considerar os números EC.
  - Esse processo pode fornecer *insights* sobre a existência e conservação de vias bioquímicas e realçar a ausência de proteínas/enzimas em diferentes organismos.

- O banco de dados KEGG é referência para este tipo de estudo.
- O estudo comparativo de vias bioquímicas fornece informações do estilo de vida dos organismos, fenótipos, adaptação a ambientes adversos, evolução, dentre outras.
- Existem também as enzimas análogas que não apresentam similaridade de sequência[131] mas possuem função similar[132], e podem atuar num mesmo processo bioquímico em diferentes organismos. Para estes casos, pode-se utilizar informações funcionais das bases de dados associadas e números EC, avaliando os elementos em PROTEIN, associados a GENOMIC SEQUENCE com *status* = complete, que foram comparados no PCG (atributo *comp* = 1 em PROTEIN) mas não possuem elementos em HIT_PP.
- Dentre outros.

Alguns atributos dos elementos do esquema conceitual desenvolvido podem ser relevantes para complementar tais consultas; como por exemplo:

Entidade PROTEIN

gbkId: permite a construção de referências cruzadas entre a base de dados RefSeq (seção proteína) com outras bases de dados.

uniprotId: permite a construção de referências cruzadas entre a base de dados UniProt com outras bases de dados.

comp: sempre será utilizado para limitar (ou excluir) o conjunto de dados do PCG (sequências de aminoácido que foram comparadas).

Entidade CDS

region: sua função principal é mapear proteínas, CDSs, transcritos e genes na sequência genômica. Pode ser utilizado como posição de referência para o acesso e análise de regiões genômicas específicas.

Entidade GENE

geneId: permite a construção de referências cruzadas entre a base de dados Gene com outras bases de dados.

[131] Não terão ocorrências em HIT_PP.
[132] Mesmo número EC e anotações funcionais similares.

transcriptId: pode ser útil em consultas direcionadas à expressão gênica, transcriptomas, perfis de expressão diferencial em células ou tecidos, por exemplo, e permitir a construção de referências cruzadas com bases de dados.

Entidade GENOMIC SEQUENCE

gbkId: permite a construção de referências cruzadas entre a base de dados RefSeq (seção genoma) com outras bases de dados.

mol_type: pode limitar uma consulta a um tipo específico de molécula.

seq_type: pode limitar uma consulta à tipo(s) específico(s) de unidade(s) genômica(s).

length_mol: útil no cálculo de estatísticas, com relação a unidades elementares das sequências genômicas (e genoma).

status: importante para limitar consultas a proteomas "preditos" completos (complete).

A combinação de propostas, como as descritas acima, assim como a adição de informações das bases de dados associadas (Uniprot, RefSeq, Gene, Pfam, EC, GO, KEGG), permite uma ampla utilização do esquema apresentado. Adicionalmente, todas as consultas podem ser limitadas a uma espécie (ou *rank* inferior) ou a um grupo taxonômico, além de módulos Pfam, números EC e anotação GO, permitindo um maior controle do usuário. Importante lembrar que as consultas podem ser limitadas, ou não, ao conjunto de proteomas "preditos" completos (*status* = complete), e que as sequências de proteína que não pertencem a este grupo podem ser importantes para (in)validar resultados.

Uma informação de grande utilidade para a tomada de decisão em qualquer procedimento é utilizar os códigos de confiabilidade de anotação das bases de dados associadas (PROTEIN EXISTENCE no UniProt, REVISION no RefSeq, EXISTENCE CODE no GO, por exemplo).

Como comentário final deve-se frisar que um ponto fundamental, e um diferencial da proposta desenvolvida nesta tese, é que a especificação e programação de consultas não precisa considerar a etapa de comparação de sequências, que já foi executada pelo PCG, permitindo um ganho de processamento considerável no momento da execução dos procedimentos[133].

[133] Após a implementação física do sistema

5.4. CONSULTAS IMPLEMENTADAS NO PWDB V.1[134]

Resumidamente, usuários podem fazer "*downloads*", comparar e analisar resultados de similaridade filtrados por genomas, funções de proteínas (utilizando números EC, domínios Pfam, GO, palavras-chave), *clusters*, dentre outras facilidades disponibilizadas através do menu da FIGURA 5.3.

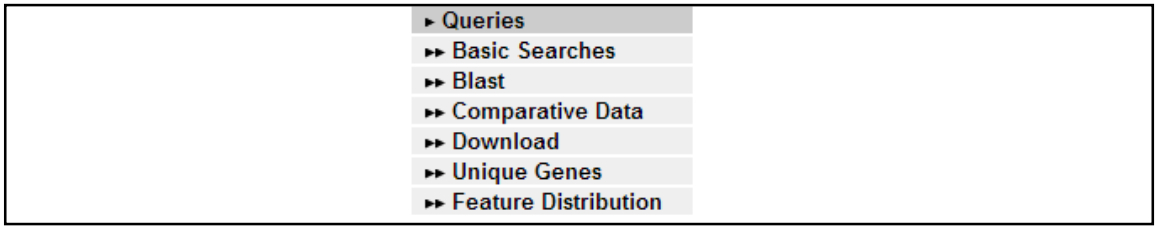


FIGURA 5.3. Menu de consultas da interface do PWDB v.1.

EXEMPLOS:

⇨ ***Consulta 1***: Opção: "**Basic Searches**" (FIGURA 5.3)
Domínio **Pfam** PF00226,
Genoma: *Escherichia coli 536*
FIGURA 5.4

⇨ ***Resultado 1***: FIGURA 5.5

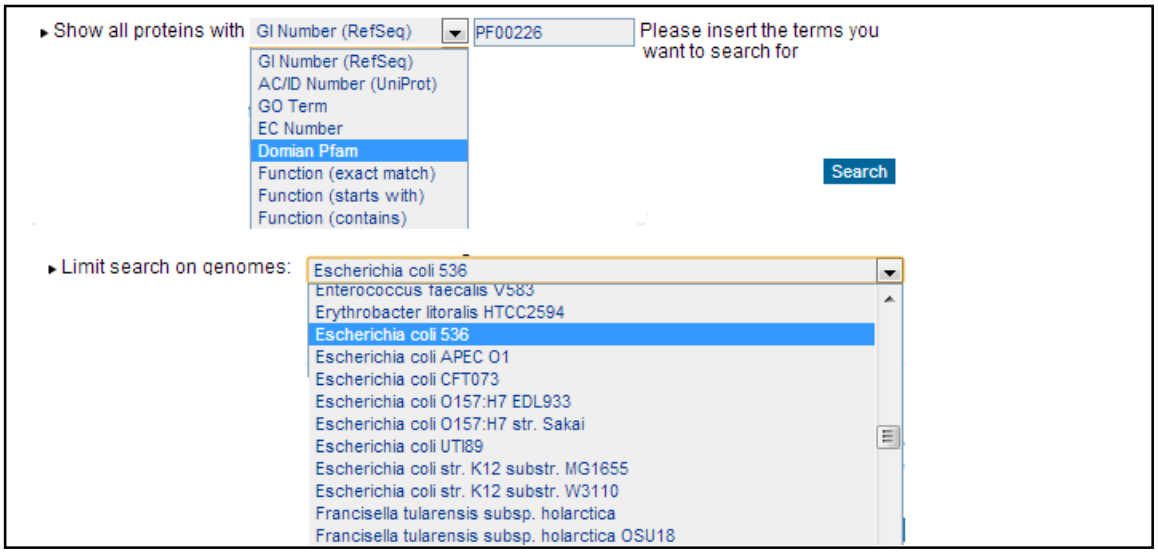


FIGURA 5.4. Parâmetros da *Consulta 1*.

[134] Maiores informações sobre as consultas implementadas podem ser vistas em [Otto, **Bezerra** *et al.*, 2010] e no *site* do PWDB v.1.

As proteínas presentes no banco de dados PWDB v.1 que preenchem os requisitos da Consulta 1 estão listadas na FIGURA 5.5 - ***Resultado** 1*. Cada proteína apresenta:

- Os identificadores das bases RefSeq (linha superior da 1ª coluna) e UniProt (linha inferior da 1ª coluna).
- A anotação das bases RefSeq (linha superior da 2ª coluna) e UniProt (linha inferior da 2ª coluna).
- As duas anotações básicas (que descrevem a proteína) podem ser comparadas.

	Primary / Secondary ID	**Annotation (Function)**	**GO-terms**	**Pfam**
	110640269 YP_667997 Q0TLT3 Q0TLT3_ECOL5	DnaJ-like protein DjlA SubName: Full=DnaJ-like protein DjlA;	GO:0031072	**PF00226**
	110642692 YP_670422 Q0TEV8 HSCB_ECOL5	chaperone protein HscB RecName: Full=Co-chaperone protein hscB; AltName: Full=Hsc20;	GO:0031072 GO:0006457 GO:0051087	**PF00226** PF07743
*	110641882 YP_669612 Q0TH68 Q0TH68_ECOL5	*putative* TPR repeat protein SubName: Full=Putative TPR repeat protein;	GO:0031072	**PF00226** PF08238
*	110641883 YP_669613 Q0TH67 Q0TH67_ECOL5	*putative* TPR repeat protein SubName: Full=Putative TPR repeat protein;	GO:0031072	**PF00226** PF08238
*	110640228 YP_667956 Q0TLX4 Q0TLX4_ECOL5	chaperone protein DnaJ RecName: Full=Chaperone protein dnaJ;	GO:0031072 GO:0006457 GO:0005737 GO:0006260 GO:0008270 GO:0051082 GO:0005524 GO:0009408	**PF00226** PF01556 PF00684
	110641184 YP_668914 Q0TJ66 CBPA_ECOL5	curved DNA-binding protein RecName: Full=Curved DNA-binding protein;	GO:0031072 GO:0006457 GO:0005737 GO:0003681 GO:0009295 GO:0051082	**PF00226** PF01556

FIGURA 5.5[135]. *Resultado 1* – seis proteínas selecionadas de acordo com os parâmetros da *Consulta1*.

Consulta 2: Opção: "**Comparative Data**" (FIGURA 5.3)

Domain: PF00226 limitado ao **genoma**: Escherichia coli 536
Similaridade com o Genoma: **todos**
FIGURA 5.6

⇨ ***Resultado 2***: FIGURA 5.7.

[135] Os três registros marcados com o símbolo * serão comentados na FIGURA 5.7

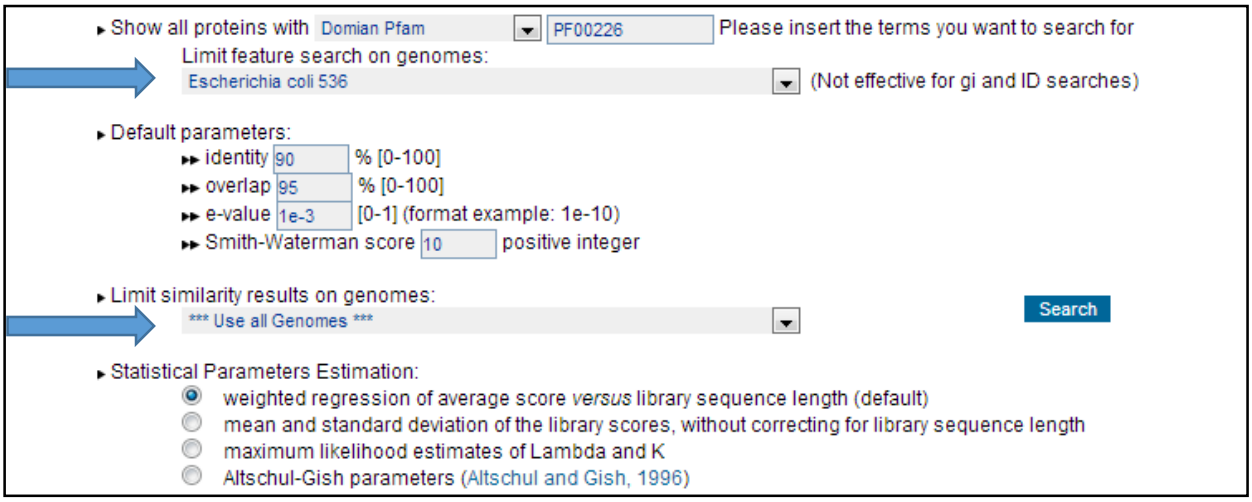
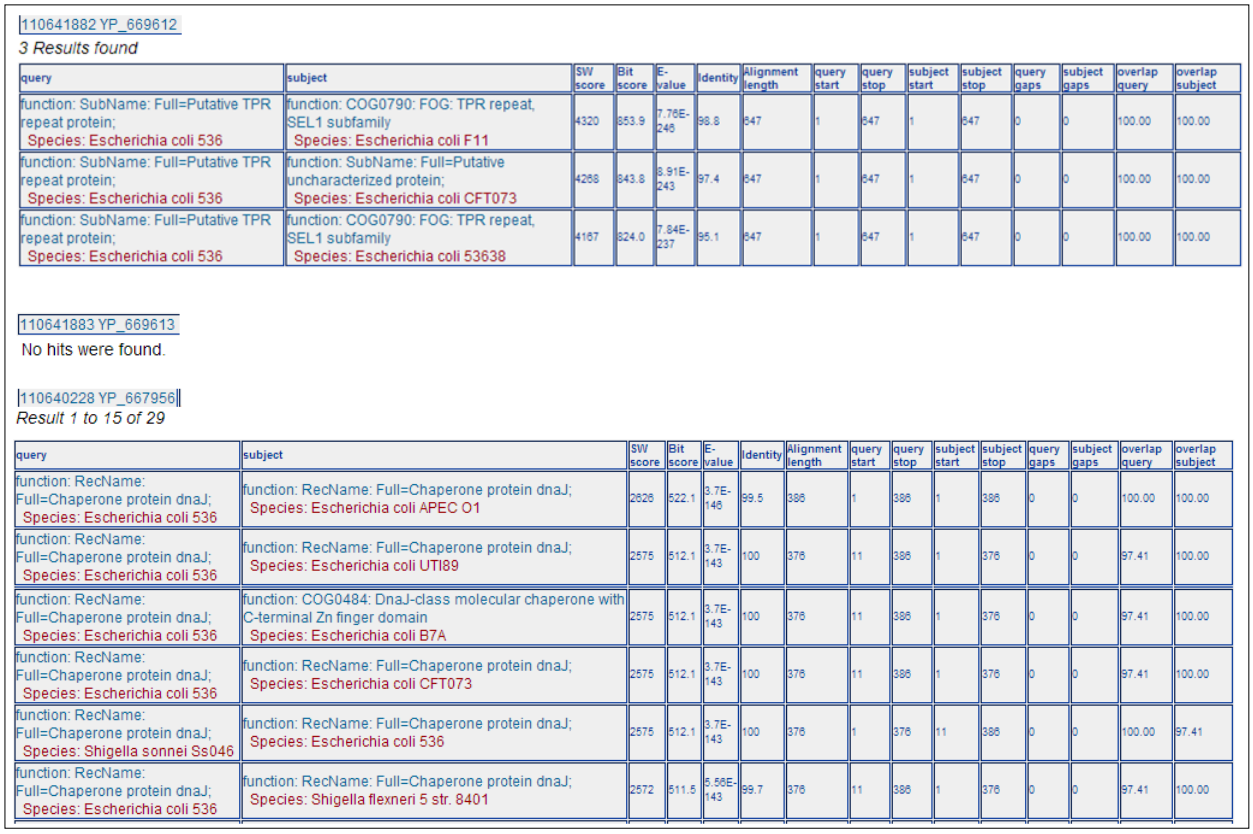


FIGURA 5.6. Parâmetros da *Consulta 2*.



110641882 YP_669612

3 Results found

query	subject	SW score	Bit score	E-value	Identity	Alignment length	query start	query stop	subject start	subject stop	query gaps	subject gaps	overlap query	overlap subject
function: SubName: Full=Putative TPR repeat protein; Species: Escherichia coli 536	function: COG0790: FOG: TPR repeat, SEL1 subfamily Species: Escherichia coli F11	4320	853.9	7.76E-246	98.8	647	1	647	1	647	0	0	100.00	100.00
function: SubName: Full=Putative TPR repeat protein; Species: Escherichia coli 536	function: SubName: Full=Putative uncharacterized protein; Species: Escherichia coli CFT073	4268	843.8	8.91E-243	97.4	647	1	647	1	647	0	0	100.00	100.00
function: SubName: Full=Putative TPR repeat protein; Species: Escherichia coli 536	function: COG0790: FOG: TPR repeat, SEL1 subfamily Species: Escherichia coli 53638	4167	824.0	7.84E-237	95.1	647	1	647	1	647	0	0	100.00	100.00

110641883 YP_669613

No hits were found.

110640228 YP_667956

Result 1 to 15 of 29

query	subject	SW score	Bit score	E-value	Identity	Alignment length	query start	query stop	subject start	subject stop	query gaps	subject gaps	overlap query	overlap subject
function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli 536	function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli APEC O1	2626	522.1	3.7E-146	99.5	386	1	386	1	386	0	0	100.00	100.00
function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli 536	function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli UTI89	2575	512.1	3.7E-143	100	376	11	386	1	376	0	0	97.41	100.00
function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli 536	function: COG0484: DnaJ-class molecular chaperone with C-terminal Zn finger domain Species: Escherichia coli B7A	2575	512.1	3.7E-143	100	376	11	386	1	376	0	0	97.41	100.00
function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli 536	function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli CFT073	2575	512.1	3.7E-143	100	376	11	386	1	376	0	0	97.41	100.00
function: RecName: Full=Chaperone protein dnaJ; Species: Shigella sonnei Ss046	function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli 536	2575	512.1	3.7E-143	100	376	1	376	11	386	0	0	100.00	97.41
function: RecName: Full=Chaperone protein dnaJ; Species: Escherichia coli 536	function: RecName: Full=Chaperone protein dnaJ; Species: Shigella flexneri 5 str. 8401	2572	511.5	5.56E-143	99.7	376	11	386	1	376	0	0	97.41	100.00

FIGURA 5.7. *Resultado 2* – A *Consulta 2* tem os mesmos parâmetros **pfam** e **genoma** da *Consulta 1* (cujo resultado são as seis proteínas da FIGURA 5.5), além do limite de busca de genomas = todos. Na FIGURA 5.7 estão listados os resultados de apenas três destas sequências – marcadas na FIGURA 5.5 com o símbolo *

Iniciando a pesquisa na opção "**Comparative Data**" pode-se:

- Escolher parâmetros de e-value, identidade, sobreposição e pontuação SW,
- Escolher parâmetros estatísticos,

- Limitar a busca a uma característica (domínio Pfam, número EC, GO) de um genoma específico:
  - No exemplo, domínio PF00226, com a opção "*Limite feature search on genomes*" = *Escherichia coli 536* – a consulta será feita para cada uma das seis sequências listadas no Resultado 1 da FIGURA 5.5.
- Limitar os *hits* a um genoma específico.

***Consulta** 3*: Opção: "**Unique Genes**" (FIGURA 5.3)
Cluster: Identidade 80%; cobertura do alinhamento: 90%
3.1 Genoma: *Saccharomyces cerevisiae* (FIGURA 5.8)
3.2 **Genoma**: *E. coli* 536

⇨ ***Resultado** 3.1*: FIGURA 5.9

⇨ ***Resultado** 3.2*: FIGURA 5.10

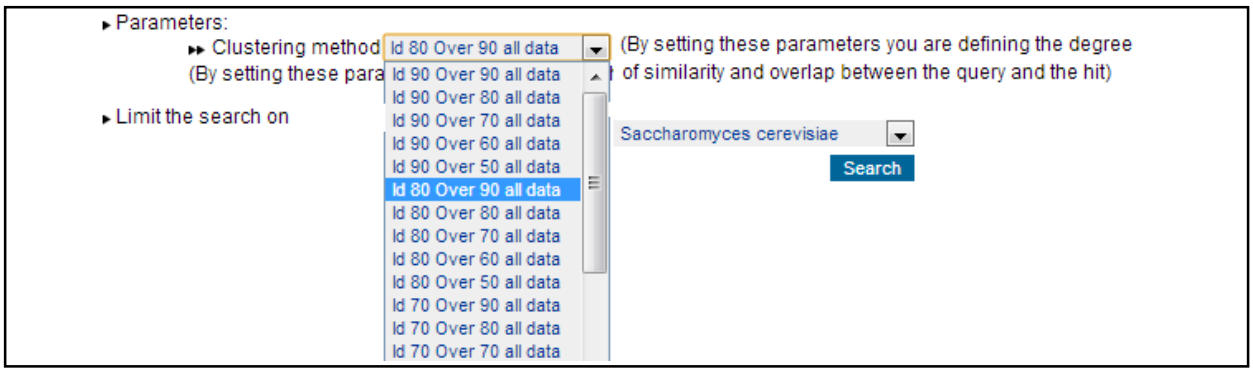


FIGURA 5.8. Parâmetros da *Consulta 3*.

O termo "*Unique Genes*" deve ser usado com muita cautela. Nesta opção de consulta:

- O resultado está baseado num método de agrupamento (*cluster*) específico[136].
- Proteínas multi-modulares são fatores complicadores para estes métodos.
- O resultado depende também da representatividade dos grupos taxonômicos no banco de dados (com proteomas completos) e suas distâncias evolutivas[137].
- Se os parâmetros forem relaxados, o resultado será alterado.

[136] [Otto *et al.*, 2008]. Existem vários métodos de agrupamento.
[137] Por exemplo, o grande número de variantes da espécie *E. coli* e outras espécies evolutivamente próximas presentes no banco de dados PWDB.

1[ST] ID	Annotation (Function)	GO-terms	EC's	Pfam
O93938	Full=Aquaporin-2;	GO:0016021-GO:0005215-GO:0005783-GO:0005515-GO:0005886-GO:0055085	-	PF00230
P12962	Full=Cap-associated protein CAF20;	GO:0017148-GO:0005515-GO:0005845-GO:0003743-GO:0005737-GO:0006412-GO:0010606	-	-
P36028	Full=ABC transporter NFT1;	GO:0016021-GO:0055085-GO:0005524-GO:0042626	-	PF00005 PF00664
P36039	Full=ER membrane protein complex subunit 3;	GO:0005783-GO:0005515-GO:0016021	-	PF05863
P36135	Full=Protein UTH1;	GO:0001308-GO:0005576-GO:0007047-GO:0007049-GO:0019898-GO:0000917-GO:0005741-GO:0009277-GO:0000422-GO:0006950-GO:0007005	-	PF03856
P36143	Full=Glycogenin-1;	GO:0005515-GO:0005978-GO:0008466	EC2.4.1.186	PF01501
P38626	Full=NADH-cytochrome b5 reductase 1;	GO:0005792-GO:0016021-GO:0055114-GO:0004128-GO:0005515-GO:0005741-GO:0005783	EC1.6.2.2	PF00175 PF00970
P38858	Full=6-phosphogluconolactonase 3; Short=6PGL;	GO:0005737-GO:0017057-GO:0005515-GO:0009051-GO:0005634	EC3.1.1.31	-
P39546	Full=DUP240 protein YAR023C;	GO:0016021	-	PF00674
P40033	Full=37S ribosomal protein RSM18, mitochondrial; Flags: Precursor;	GO:0005763-GO:0032543-GO:0003735	-	PF01084
P42844	Full=Mitochondrial protein import protein ZIM17;	GO:0005743-GO:0005759-GO:0030150-GO:0050821-GO:0005515-GO:0006457-GO:0006986-GO:0008270	-	PF05180
P46993	Full=Protein ASG7;	GO:0000747-GO:0016021-GO:0005886	-	-
P46997	Full=J protein JJJ2;	GO:0005737-GO:0051082-GO:0005634-GO:0006457-GO:0031072	-	PF00226
P48559	Full=GTP-binding protein YPT11;	GO:0016020-GO:0048309-GO:0048313-GO:0003924-GO:0005783-GO:0000001-GO:0000131-GO:0005515-GO:0005525-GO:0005934-GO:0005935-GO:0051645	-	PF00071
P50088	Full=Stationary phase gene 1 protein;	GO:0016021-GO:0005739	-	-
P50273	Full=Mitochondrial translation factor ATP22; Flags: Precursor;	GO:0005743-GO:0005515-GO:0006417-GO:0006461	-	-
Q00947	Full=Transcription factor STP1; Flags: Precursor;	GO:0005886-GO:0005515-GO:0006388-GO:0003700-GO:0003704-GO:0005634-GO:0008270-GO:0045944	-	PF00096
Q02721	Full=Meiotic recombination protein REC102;	GO:0042138-GO:0000794-GO:0005515-GO:0007131	-	-
Q04597	Full=Uncharacterized protein YDR114C;	-	-	-
Q06053	Full=tRNA-dihydrouridine synthase 3;	GO:0005634-GO:0050660-GO:0003676-GO:0008270-GO:0005737-GO:0006400-GO:0017150-GO:0055114	-	PF00642 PF01207
Q06406	Full=U6 snRNA-associated Sm-like protein LSm6;	GO:0003723-GO:0005515-GO:0006364-GO:0005688-GO:0006402-GO:0008033-GO:0000398-GO:0000932-GO:0005730-GO:0005732-GO:0046540	-	PF01423
Q07535	Full=Uncharacterized protein YDL118W;	GO:0005515	-	-
Q08337	Full=Putative uncharacterized transporter YOL162W/YOL163W;	GO:0006810-GO:0016021	-	-
Q12173	Full=Transposon Ty3-G Gag polyprotein; Contains: Full=Spacer peptide p3; Contains: Full=Nucleocapsid protein p9; Short=NCp9; Short=NC;	GO:0005737-GO:0008270-GO:0003676	-	PF00098
Q12478	Full=Putative uncharacterized protein YLR156W/YLR157W-D/YLR159W/YLR161W;	GO:0005515	-	-
Q2V2Q1	Full=Uncharacterized protein YCL058W-A;	-	-	-
Q6B308	Full=Putative glucosamine--fructose-6-phosphate aminotransferase [isomerizing]; Short=GFAT;	GO:0004360-GO:0005529-GO:0005975-GO:0006541	EC2.6.1.16	PF00310 PF01380
6323731	hypothetical protein; Ymr085wp (refSeq 21)	-	-	-

FIGURA 5.9. *Resultado 3.1:* 28 proteínas do genoma ***S. cerevisiae*** não possuem identidade de pelo menos 80% com cobertura do alinhamento de 90% com as outras proteínas do PWDB v.1.

The following unique genes were found with cluster parameters *Id 80 Over 90 all data*.

Results 1 to 10 of 11 Escherichia coli 536

Primary / Secondary ID	Annotation (Function)	GO-terms	EC's	Pfam
110640243 YP_667971 (show hits)	hypothetical protein ECP_0030	-	-	-
110640277 YP_668005 (show hits)	hypothetical protein ECP_0066	-	-	-
110640326 YP_668054 (show hits)	uropathogenic specific protein	-	-	-
110640355 YP_668083 (show hits)	hypothetical protein ECP_0144	-	-	-
110640453 YP_668181 (show hits)	H repeat-associated protein of Rhs element	-	-	-
110640488 YP_668216 (show hits)	hypothetical protein ECP_0282	-	-	-
110640515 YP_668243 (show hits)	hypothetical membrane protein	-	-	-
110640522 YP_668250 (show hits)	hypothetical protein ECP_0317	-	-	-
110640529 YP_668257 (show hits)	putative LysR-family transcriptional regulator	-	-	-
110640530 YP_668258 (show hits)	lysyl-tRNA synthetase, heat inducible	-	-	-

FIGURA 5.10. *Resultado 3.2*: 11 proteínas do genoma de ***E. coli* 536** não possuem identidade de pelo menos 80% com cobertura do alinhamento de 90% com as outras proteínas do PWDB v.1.

Uma das facilidades bastante interessante oferecida já na primeira versão do PWDB é a possibilidade de *download* do resultado da comparação entre dois proteomas completos (FIGURA 5.11).

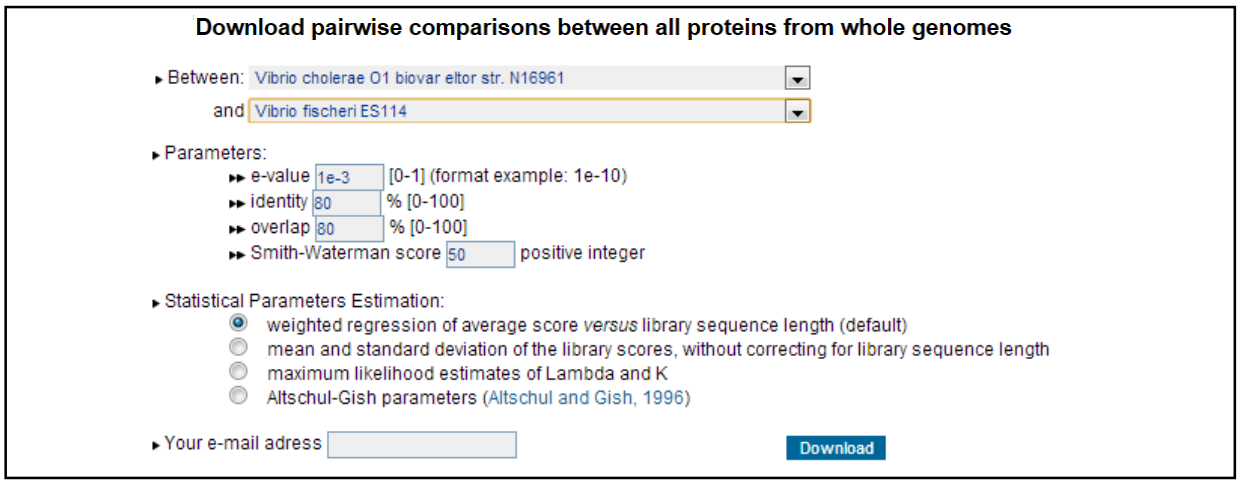


FIGURA 5.11. Menu para a escolha de parâmetros para *download* do resultado da comparação entre dois proteomas (“preditos”) completos.

Uma nova implementação do PWDB baseada no esquema da FIGURA 4.8, tem o potencial de oferecer novas possibilidades para pesquisas mais avançadas. Todas as consultas já disponíveis na primeira versão são passíveis de serem respondidas de forma ampliada, e o controle do usuário na elaboração de consultas pode ser mais preciso e pontual, fornecendo condições para a construção de procedimentos mais complexos. Além disso, pode-se pensar num sistema que recupere resultados de consultas intermediárias e os reutilize como *input* em etapas posteriores.

Com o esquema conceitual proposto nessa tese, todas as consultas desenhadas para o PWDB v.1, por exemplo, podem ser limitadas ao táxon folha da hierarquia (espécie ou níveis inferiores) ou nodos superiores como gênero, ordem e assim por diante. Podem também ser limitadas aos genomas que possuem todas as suas sequências genômicas representadas no banco de dados, isto é, o proteoma completo (“predito”) do organismo estaria representado na entidade PROTEIN. A *Consulta 1* da FIGURA 5.4, por exemplo, poderia ser limitada ao genoma de uma das variantes de *E. coli*[138] (*rank* inferior à espécie; por exemplo: *E. coli* O157:H7 str. Sakai, taxonId = 386585), ou considerar todas as variantes, optando por um grupo de hierarquia superior como espécie (*E. coli*, *taxonId* = 562), gênero (*Escherichia*, *taxonId* = 561), classe (Gammaproteobacteria, *taxonId* = 1236) e etc.

[138] Única opção através da interface do PWDB v.1

5.5. EXTENSÃO DO MODELO

Um ponto a ser ressaltado no esquema da FIGURA 4.8, refere-se a entidade ORF_T e o relacionamento GS ORF_T. Este conjunto de elementos deve ser visto como um anexo ao módulo CENTRAL. As sequências tORF foram concebidas para um estudo específico do PCG.

Considerando a ideia de módulos anexos, pode-se afirmar que o esquema conceitual proposto ultrapassou o objetivo original de modelagem direcionada, especificamente, aos dados e requerimentos do PCG. Módulos anexos podem ser inseridos/removidos para incluir outros projetos baseados em comparações de sequências de proteína, que se beneficiem das informações fornecidas pelo módulo CENTRAL do modelo.

Da mesma forma, outras bases de dados específicas – por exemplo, bases de dados de: expressão gênica, transcriptomas, estruturas 3D de proteínas, proteomas (experimentais), interações de proteínas, dentre outras – podem ser integradas ao esquema a partir de referências cruzadas não só com a entidade PROTEIN, mas também com as entidades GENE e GENOMIC SEQUENCE, através dos atributos que referenciam os identificadores únicos das bases de dados RefSeq, UniProt, Gene e Taxonomy. Novas bases adicionadas ao esquema podem fornecer informações extra para responder requisitos e/ou necessidades futuras e incrementar o processo de anotação, além de oferecer novos parâmetros para consultas e procedimentos.

Conjuntos com novas sequências de proteína (versões mais recentes das bases RefSeq e UniProt) podem ser comparadas e seus *hits* serem adicionadas aos do conjunto original do PCG, assim como outros conjuntos de dados de proteína podem ser comparados, e seus resultados substituírem os *hits* do PCG.

As informações centrais do esquema que têm como referência as bases de dados RefSeq, Gene e UniProt, podem ser instanciadas e atualizadas a qualquer momento, utilizando *scripts* já desenvolvidos e disponíveis para a carga de dados [Tristão and Lifschitz, 2009].

6. CONCLUSÕES

O propósito original do estudo desta tese – modelagem conceitual de bancos de dados biológicos –, surgiu com a proposta para o desenvolvimento de um sistema de banco de dados para armazenar e gerenciar o resultado do Projeto Comparação de Genomas (PCG).

O interesse em desenvolver um sistema para o PCG deve-se a importância de seu resultado:

- Mais de 4 milhões de sequências de aminoácidos foram comparadas "par-a-par",
- Centenas de genomas completos foram utilizados,
- O programa de comparação utilizado foi o SSEARCH,
- O PCG gerou como resultado uma matriz de aproximadamente 1 Terabyte (TB), com 4,2 x 10^9 linhas com dados de similaridade de sequências[139],
- Resumindo, o PCG calculou e armazenou índices de similaridade, resultantes da comparação de milhões de sequências de proteína, eliminando, desta forma, a fase de maior custo computacional necessária para análises genômicas comparativas (que é a comparação de sequências[140]).

A primeira produção do grupo LGFB/LaBBio utilizando os dados do PCG foi o PWDB v.1[141]. Este banco de dados foi publicado em [Otto, **Bezerra** *et al.*, 2010], está atualmente funcional e disponível na web através de uma interface gráfica e atende aos requisitos de maior urgência da época em que foi desenvolvido.

Num projeto de sistema de banco de dados, a organização do banco de dados é uma das etapas mais importantes e seu desenho deve ser representado, primeiramente, em um esquema conceitual, que é uma representação gráfica do modelo conceitual. A modelagem conceitual é a fase em que se busca representar, em uma linguagem de alto nível, os conceitos/objetos presentes no domínio do problema; e independe da escolha de software e paradigmas de desenvolvimento, do sistema de computação e da evolução de tecnologia.

Ao longo do projeto de modelagem, objeto desta tese, dois pontos fundamentais foram considerados para responder adequadamente os requisitos do PCG:

[139] Índices de similaridade com um *cut-off* mínimo, estatisticamente significativo.
[140] No caso específico do PCG, os elementos genômicos comparados foram sequências de proteína. Existem diferentes abordagens para comparação de genomas.
[141] Anterior ao esquema conceitual proposto nesta tese.

(a) Para a recuperação de informações preditas para as proteínas em fontes de dados biológicos externas e comparação de diferentes anotações, foram construídas referências cruzadas entre os identificadores das sequências de aminoácidos de cada *hit* com as bases NCBI-RefSeq e UniProt.

(b) Para estudos genômicos, a questão chave foi relacionar as sequências de aminoácidos com suas sequências de nucleotídeos de origem[142] e sua posição genômica.

O esquema conceitual desenvolvido nesta tese foi publicado em [Lifschitz, **Bezerra** *et al.*, 2012]. Uma nova versão (já prevista) do banco de dados PWDB, baseada neste esquema, pode permitir pesquisas mais avançadas e a construção de procedimentos mais complexos.

Devido a característica básica de um esquema conceitual de representar o modelo de dados de maneira independente de especificações computacionais, durante a fase conceitual de um sistema de banco de dados tenta-se adiar decisões de projeto relacionadas a paradigmas computacionais específicos, mantendo-se o foco no processo de compreensão e definição da semântica do domínio do problema. Nessa tese, o esquema conceitual foi representado através de um formato gráfico de fácil compreensão — o diagrama ER, proposto para facilitar o planejamento de consultas e procedimentos por pesquisadores da área de genômica (que normalmente possuem conhecimento limitado em bancos de dados), assim como guiar o desenvolvimento e a implementação física de uma segunda versão do PWDB por profissionais da área de computação (que normalmente possuem conhecimento limitado em genômica).

Poder utilizar fontes distintas e associar diferentes tipos de informação biológica e métodos computacionais é uma interessante opção para pesquisas pontuais e mais específicas, que podem revelar novos dados, além de ressaltar possíveis incoerências ou erros. O esquema apresentado prevê a utilização de recursos variados com a vantagem de considerar a associação de bases de dados biológicos de referência a uma base com índices de similaridade já pré-calculados, resultantes da comparação de sequências de proteína oriundas de genomas completos (e incompletos).

Foram apresentados alguns exemplos para demonstrar a possibilidade de utilização do esquema conceitual para planejamento de consultas e procedimentos, mesmo antes da

[142] Da base de dados Refseq.

existência de um esquema lógico. Com isso, usuários sem conhecimento suficiente para utilizar uma linguagem de banco de dados podem ter uma maior participação no projeto de um sistema, utilizando os elementos do diagrama do esquema para um maior detalhamento de questões de interesse que poderão ser melhor compreendidas pelos desenvolvedores do sistema.

Estudos e implementações utilizando o esquema vêm sendo desenvolvidos pelo grupo do LabBio, e dentre alguns já finalizados pode-se citar [Tristão C, 2012; Viana *et al.*, 2011a; Viana *et al.*, 2011b; Tristão and Lifschitz, 2009].

O esquema conceitual proposto pode ser estendido com facilidade. Módulos anexos podem ser inseridos/removidos para incluir outros projetos baseados em comparação de sequências de proteína que se beneficiem das informações fornecidas pelo módulo CENTRAL do modelo. Da mesma forma, novas bases de dados específicas de diferentes áreas (*-ômicas*, por exemplo) podem ser integradas ao esquema a partir de referências cruzadas não só com a entidade PROTEIN, mas também com as entidades GENE, GENOMIC SEQUENCE e TAXONOMY.

Novas sequências de proteína podem ser comparadas entre si e com as sequências originais do PCG e seus *hits* serem adicionados aos do conjunto original, assim como outros conjuntos de sequências de proteína podem ser comparados, e seus resultados substituírem os *hits* do conjunto original do PCG.

As informações centrais do esquema, que têm como referência as bases de dados RefSeq, Gene e UniProt, podem ser instanciadas e atualizadas a qualquer momento utilizando *scripts* já desenvolvidos e disponíveis para a carga de dados.

Para finalizar, vale ressaltar um ponto fundamental da proposta do sistema de banco de dados PWDB que é permitir a construção de consultas e procedimentos no domínio da genômica comparativa sem a necessidade de comparação de sequências; e pode representar um ganho considerável em processamento, tempo e espaço computacional no momento de execução.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Allen G, Bajaj A, Khatri V, Ram S, Siau K. Advances in Data Modeling Research by Communications of the Association for Information Systems. 2006;17:677-692,

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the Ortholog Conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol. 2012;8(5):e1002514.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990 Oct 5;215(3):403-10.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. 2000;25:25–29.

Augen J. Information technology to the rescue! Nature Biotechnol. 2001;19:BE39–BE40.

Bachman CW. Data Structure Diagrams. ACM SIGMIS Database. 1969;1(2):4-10.

Bagowski CP, Bruins W, te Velthuis AJW. The nature of protein domain evolution: shaping the interaction network. Curr Genomics. 2010 Aug;11(5):368-76.

Bell MJ, Gillespie CS, Swan D, Lord P. An approach to describing and analysing bulk biological annotation quality: a case study using UniProtKB. Bioinformatics. 2012 Sep 15;28(18):i562-i568.

Beringer D. Limits of seamless in object oriented software development. In: Proceedings of the 13th International Conference on Technology of Object Oriented Languages and Systems. Versalles, France. 1994;161-171.

Birney E, Clamp M. Biological database design and implementation. Briefings in Bioinformatics. 2004 March;l5(1):31–38.

Boekhorst J, Snel B. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. BMC Bioinformatics. 2007;8:356.

Bornberg-Bauer E, Paton NW. Conceptual data modelling for bioinformatics. Brief Bioinform. 2002 Jun;3(2):166-80.

Brenner S. Nobel Lecture: Nature´s Gift to Science. Nobelprize.org. 13 Dec 2012. http://www.nobelprize.org/nobel_prizes/medicine/laureates/2002/brenner-lecture.html

Buljan M, Bateman A. The evolution of protein domain families. Biochem Soc Trans. 2009 Aug;37(Pt 4):751-5.

Busch N, Wedemann G. Modeling genomic data with type attributes, balancing stability and maintainability. BMC Bioinformatics. 2009;10:97.

Bushman F. Lateral DNA transfer: mechanisms and consequences. Cold Spring Harbor Laboratory Press; November 13, 2001.

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol. 2006;7:R13.

Chen J, Sidhu AS. Biological Database Modeling. Norwood, MA, USA: Artech House Inc.; 2007

Chen JY, Carlis JV. Genomic data modeling. Information. Special issue: Data management in bioinformatics. June 2003;28(4):287–310

Chen PP, Thalheim B, Wong LY. Future directions of conceptual modeling. Conceptual modeling, Current Issues and Future Directions. Lecture Notes in Computer Science. 1999;1565:287-301

Chen PP. The Entity-Relationship model - toward a unified view of data. ACM Transactions of Database Systems. 1976;1(1):9-36.

Chen X, Zhang J. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol. 2012;8(11):e1002784.

Chena YJ, Carlisb JV. Genomic data modeling. Information Systems. 2003;28:287–310.

Codd EF. A relational model for large shared databanks. Communications of the ACM. 1970;13:377-387.

Dalquen DA, Altenhoff AM, Gonnet GH, Dessimoz C. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PLoS One. 2013;8(2):e56925.

Descorps-Declère S, Lemoine F, Sculo Q, Lespinet O, Labedan B. The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species. Biochimie. 2008 Apr;90(4):595-608.

Dunnen JD, Antonarakis S. Nomenclature for the description of human sequence variations. Human genetics. 2001;109(1):121-124.

Durrens P, Nikolski M, Sherman D. Fusion and Fission of Genes Define a Metric between Fungal Genomes. PLoS Comput Biol. 2008;4(10):e1000200

Eisen JA, Fraser CM. Phylogenomics: intersection of evolution and genomics. Science. 2003;300:1706-1707.

El-Mabrouk N, Sankoff D. Analysis of gene order evolution beyond single-copy genes. Methods Mol Biol. 2012;855:397-429.

Elmasri R, Ji F, Fu J, Zhang Y, Raja Z. Modelling Concepts And Database Implementation Techniques For Complex Biological Data. International Journal of Bioinformatics Research and Applications. 2007;3(3):366–388.

Elmasri R, Navathe S. Sistemas de Banco de Dados. – 6.ed. São Paulo: Pearson Education Br; 2011.

Farrar, M. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics. 2007;23(2):156–161.

Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970;19:99-113.

Fitch WM. Homology a personal view on some of the problems. Trends Genet. 2000 May;16(5):227-31.

Forslund K, Pekkari I, Sonnhammer EL. Domain architecture conservation in orthologs. BMC Bioinformatics. 2011 Aug 5;12:326.

Fulton DL, Li YvonneY, Laird MR, Horsman BGS, Roche FM, Brinkman FSL. Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics. 2006;7:270.

Gillis J, Pavlidis P. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). In Selected proceedings from the Automated Function Prediction Meeting. Vienna, Austria, 15-16 July 2011. BMC Bioinformatics 2013;14(Suppl 3):S15.

Gregory TR. The evolution of the genome. Elsevier/Academic Press; 2005.

Guimarães ACR. Identificação *in silico* de enzimas isofuncionais não-homólogas, um potencial reservatório de alvos terapêuticos [tese]. Rio de Janeiro: Instituto Oswaldo Cruz - FIOCRUZ; 2010.

Halpin TA. Comparing metamodels for ER, ORM and UML data models. In: Keng Siau. Advanced Topics in Database Research Volume 3. Missouri University of Science and Technology, USA; 2004.23-44.

Hammer M, Mcleod D. Database description with SDM: a semantic database model. ACM Transactions on Database Systems. 1981;6(3):351-386.

Harrington JL. Relational database design and implementation: clearly explained. 3rd ed. Morgan Kaufmann Series in Data Management Systems; 2009.

Heger A, Wilton CA, Sivakumar A, Holm L. ADDA: a domain database with global coverage of the protein universe. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D188-91.

Henricson A, Forslund K, Sonnhammer EL. Orthology confers intron position conservation. BMC Genomics. 2010 Jul 2;11:412.

Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems Biology. Annu. Rev. Genomics Hum. Genet. 2001;2:343–72.

Ishak I, Salim N. Database integration approaches for heterogeneous biological data sources: an overview. In: Postgraduate Annual Research Seminar (PARS 2006). Postgraduate Studies Department FSKSM, UTM Skudai; 2006

Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics, 2013;29(7):837-844.

Jun J, Mandoiu II, Nelson CE. Identification of mammalian orthologs using local synteny. BMC Genomics. 2009;10:630.

Juristo, AM. Introductory paper: reflections on conceptual modelling. Data & Knowledge Engineering. 2000;33:103-117.

Kaindls H, Carroll JM. Simbolic modeling in practice. Communications of the ACM. 1999;42(1):28-30.

Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahrén D, Tsoka S, Darzentas N, Kunin V, López-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res. 2005;33:6083–6089

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Res. 2009 Aug;19(8):1404-18.

Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, Thompson LP, Cabello JM, Davidson RS, Goldberg AP, Shasha DE, Coruzzi GM, Gutiérrez RA. VirtualPlant: a software platform to support systems biology research. Plant Physiol. 2010 Feb;152(2):500-15.

Keet CM. Biological data and conceptual modelling methods. Journal of Conceptual Modeling. 2003 October;29.

Khajeh-Saeed A, Poole S, Perot JB. Acceleration of the Smith-Waterman algorithm using single and multiple graphics processors," Journal of Computational Physics. 2010;229(11):4247–4258.

Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrachi I, Pruitt KD, Tatusova T. Solving the Problem: Genome Annotation Standards before the Data Deluge. Stand Genomic Sci. 2011 October 15;5(1):168–193.

Köhler J, Philippi S, Lange M. SEMEDA: ontology based semantic integration of biological databases. Bioinformatics. 2003;19:2420–2427.

Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics. 2005;539: 309–338.

Kummerfeld SK, Teichmann SA. Protein domain organisation: adding order. BMC Bioinformatics. 2009 Jan 29;10:39.

Kummerfeld SK, Teichmann SA. Relative rates of gene fusion and fission in multi-domain proteins. Trends in Genetics. January 2005;21(1):25–30.

Lee ST, Lin CY, Hung CL. GPU-based cloud service for Smith-Waterman algorithm using frequency distance filtration scheme. Biomed Res Int. 2013;2013:721738.

Li ITS, Shum W, Truong K. 160-fold acceleration of the Smith-Waterman algorithm using a field programmable gate array (FPGA). BMC Bioinformatics. 2007;8:185.

Lifschitz S, Viana CJM, Tristão C, Catanho M, Degrave WM, Miranda AB, **Bezerra M**, Otto TD. Design and implementation of ProteinWorldDB. Advances in Bioinformatics and Computational Biology. Lecture Notes in Computer Science. 2012;7409:144-155.

Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227(4693):1435–41.

Liu Y, Wirawan A, Schmidt B. CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions. BMC Bioinformatics. 2013;14:117.

Long M. A New Function Evolved from Gene Fusion. Genome Res. 2000;10:1655-1657

Loucopoulos P, Karakostas V. System requirements engineering. New York, NY, USA: McGraw-Hill, Inc.; 1995.

Manavski SA, Valle G. CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. BMC Bioinformatics. 2008;vol 9,supplement 2,article S10.

Mayordomo AM. Human genome conceptual modeling: an ontological framework for the design and implementation of genomic information systems. Research Challenges in Information Science (RCIS) IEE. 2012;1-6.

Mazumder R, Natale DA, Murthy S, Thiagarajan R, Wu CH. Computational identification of strain-, species- and genus-specific proteins. BMC Bioinformatics. 2005 Nov 23;6:279

Mazza R, Strozzi F, Caprera A, Ajmone-Marsan P, Williams JL. The other side of comparative genomics: genes with no orthologs between the cow and other mammalian species. BMC Genomics. 2009 Dec 14;10:604.

Mendes P, Sha W, Ye K. Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics. 2003;19(2):II122—II129

Navathe SB, Kogelnik AM. The challenges of modeling biological information for genome databases. In: Goos G, Hartmanis J, van Leeuwen J., Chen PP, Akoka J, Kangassalu H, Thalheim B. Selected Papers from the Symposium on Conceptual Modeling, Current Issues and Future Directions. Lecture Notes in Computer Science; 1999. 15651:68-182.

Nelson MR, Reisinger SJ, Henry SG. Designing databases to store biological information. Biosilico. 2003;1(4):134-142.

Olinski RP, Lundin LG, Hallböök F. Conserved synteny between the Ciona genome and human paralogons identifies large duplication events in the molecular evolution of the insulin-relaxin gene family. Mol Biol Evol. 2006 Jan;23(1):10-22.

Ostel JM, Wheelan SJ, Kans JA. The NCBI data model. 2001. In: Baxevanis AD, Ouellette BF. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2nd ed. John Wiley & Sons, Inc.; 2001. Chapter 2.

Otto TD, Catanho M, Tristão C, **Bezerra M**, Fernandes RM, Elias GS, Scaglia AC, Bovermann B, Berstis V, Lifschitz S, de Miranda AB, Degrave W. ProteinWorldDB: querying radical pairwise alignments among protein sets from complete genomes. Bioinformatics. 2010 Mar 1;26(5):705-7.

Pastor O, Casamayor JC, Celma M, Mota L, Ángeles Pastor M, Levin AM. Conceptual modeling of human genome: integration challenges. In: Conceptual Modelling and Its Theoretical Foundations. Lecture Notes in Computer Science. 2012;7260:231-250

Pastor O. Conceptual modeling meets the human genome. In: Qing Li, Spaccapietra S, Yu E, Olivé A. Conceptual Modeling - ER 2008: 27th International Conference on Conceptual Modeling, Barcelona, Spain. Lecture Notes In Computer Science. 2008;1-11

Paton NW, Khan SA, Hayes A, Moussouni F, Brass A, Eilbeck K, Goble CA, Hubbard SJ, Oliver SG. Conceptual modeling of genomic information. Bioinformatics. 2000;16(6):548-557.

Pearson WR. Comparison of methods for searching protein sequence databases. Protein Sci. 1995 Jun;4(6):1145-60.

Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics. 1991 Nov;11(3):635-50.

Pennisi E. How will big pictures emerge from a sea of biological data? Science. 2005;309:94.

Philippi S, Köhler J. Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet. 2006 Jun;7(6):482-8.

Philippi S, Köhler, J. Using XML technology for the ontology-based semantic integration of life science databases. IEEE Trans. Inf. Technol. Biomed. 2004;8:154–160.

Philippi S. Light-weight integration of molecular biological databases. Bioinformatics. 2004 Jan 1;20(1):51-7.

Poptsova MS, Gogarten JP. Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiology. July 2010;156(7):1909-1917.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic Acids Res. 2012 Jan;40(Database issue):D290-301.

Ram, S. Intelligent database design using the unifying semantic model. Information and Management. 1995;29(4):191-206.

Rattei T, Tischler P, Götz S, Jehl MA, Hoser J, Arnold R, Conesa A, Mewes HW. SIMAP--a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. Nucleic Acids Res. 2010 Jan;38(Database issue):D223-6.

Rattei T, Arnold R, Tischler P, Lindner D, Stümpflen V, Mewes HW. SIMAP: the Similarity Matrix of Proteins. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D252-6.

Reeves GA, Talavera D, Thornton JM. Genome and proteome annotation: organization, interpretation and integration. J R Soc Interface. 2009 Feb 6;6(31):129-47.

Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol. 2009;5:e1000431.

Reid AJ, Ranea JAG, Clegg AB, Orengo CA. CODA: Accurate Detection of Functional Associations between Proteins in Eukaryotic Genomes Using Domain Fusion. PLoS One. 2010;5(6):e10908.

Rentzsch R, Orengo CA. Protein function prediction using domain families. In: Selected proceedings from the Automated Function Prediction Meeting 2011. BMC Bioinformatics. 2013;14(3):S5.

Richesson R, Turley JP. Conceptual models: definitions, construction, and applications in public health surveillance. Journal of Urban Health. 2003;80:128.

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science. 2000 Dec 15;290(5499):2105-10.

Rognes T. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. BMC Bioinformatics. 2011 Jun 1;12:221.

Rojas-Mujica I, Bornberg-Bauer E. Database systems for the analysis of biochemical pathways. In Molecular Biology and Pathogenicity of Mycoplasmas (Razin S and Hermann R, eds). 2002:201–220, Kluwer Academic.

Roos DS. Computational biology. Bioinformatics – trying to swim in a sea of data. Science. 2001;291:1260–1261.

Rost B. Enzyme function less conserved than anticipated. J. Mol. Biol. 2002;3318:595–608.

Rubin DL, Shafa F, Oliver DE, Hewett M, Altman RB. Representing genetic sequence data for pharmacogenomics: an evolutionary approach using ontological and relational models. Bioinformatics. 2002;18(Suppl. 1):S207–S215

Rudnicki WR, Jankowski A, Modzelewski A, Piotrowski A, Zadrozny A. The new SIMD implementation of the smith-waterman algorithm on cell microprocessor. Fundamenta Informaticae. 2009;96(1-2):181–194.

Shpaer EG, Robinson M, Yee D, Candlin JD, Mines R, Hunkapiller T. Sensitivity and selectivity in protein similarity searches: a comparison of smith-waterman in hardware to blast and fasta. Genomics. 1996;38(2):179–191.

Siew N, Fischer D. Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins. 2003;53:241-251.

Sjölander K, Datta RS, Shen Y, Shoffner GM. Ortholog identification in the presence of domain architecture rearrangement. Brief Bioinform. 2011 Sep;12(5):413-22.

Smith JM, Smith DCP. Database abstractions: aggregation and generalization. ACM Transactions on Database Systems. 1977;2(2):105-133.

Smith TF, Waterman MS. Comparison of Biosequences. Adv. Appl. Math. 1981;2:482-9.

Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147(1):195-7.

Sokhansanj BA, Fitch JP, Quong JN, Quong AA. Linear fuzzy gene network models obtained from microarray data by exhaustive search. BMC Bioinformatics. 2004;5(1):108.

Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001 Jul;2(7):493-503.

Stein LD. Integrating biological databases. Nature Reviews Genetics. 2003 may;4:337.

Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics. 2000 Feb;16(2):184-5.

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278:631-637.

TianW, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. 2003;333:863–882.

Toll-Riera M, Mar Albà M. Emergence of novel domains in proteins. BMC Evolutionary Biology. 2013;13:47.

Torbjørn R. Faster Smith–Waterman database searches with inter-sequence SIMD parallelization. BMC Bioinformatics. 2011;12:221.

Tristão C, Lifschitz S. Protein World Database: geração do esquema lógico e processo de ETL. [Monografia em Ciência da Computação]. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro; outubro, 2009:28/09.

Tristão C. Uma Abordagem para Modelar, Armazenar e Acessar Sequências Biológicas [tese]. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro; 2012.

Tsichritzis DC, Lochovsky FH. Hierarchical database management: a survey. Computing Surveys. 1976 March;105-124.

Uchiyama I. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res. 2007;35:D343–D346.

Viana CJ, Lifschitz S, Haeusler EH, Miranda AB. Processamento de dados semânticos na cloud: um estudo de caso com o Protein World Database. [Monografia em Ciência da Computação]. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro; abril 2011:03/11.

Viana CJ, Lifschitz S, Haeusler EH, Miranda AB. Protein World Database: definição e implementação de estruturas organizacionais [Monografia em Ciência da Computação]. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro; abril 2011:02/11.

Wieringa R. Requirements engineering frameworks for understanding. Wiley; April 1996.

WR Pearson. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics. 1991 Nov;11(3):635-50.

Xiaohua Zhou, Il-Yeol Song. Conceptual Modeling of Genetic Studies and Pharmacogenetics. Computational science and its applications – ICCSA 2005. Lecture Notes in Computer Science. 2005; 3482:402-415

Yupeng Wang, Haibao Tang, DeBarry JD, Xu Tan, Jingping Li, Xiyin Wang, Tae-ho Lee, Huizhe Jin, Marler B, Hui Guo, Kissinger JC, Paterson AH. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucl. Acids Res. 2012;40(7):e49.

8. ANEXOS

Design and Implementation of ProteinWorldDB

Sérgio Lifschitz[1], Carlos Juliano M. Viana[1], Cristian Tristão[1], Marcos Catanho[2], Wim M. Degrave[2], Antonio Basílio de Miranda[3], Márcia Bezerra[3], and Thomas D. Otto[4]

[1] PUC-Rio - Departamento de Informática - Rio de Janeiro (RJ)
sergio@inf.puc-rio.br
[2] Laboratório de Genômica Funcional e Bioinformática
Instituto Oswaldo Cruz - Rio de Janeiro (RJ)
[3] Laboratório de Biologia Computacional e Sistemas
Instituto Oswaldo Cruz - Rio de Janeiro (RJ)
[4] Wellcome Trust Sanger Institute - Hinxton - United Kingdom

Abstract. This work involves the comparison of protein information in a genomic scale. The main goal is to improve the quality and interpretation of biological data, besides our understanding of biological systems and their interactions. Stringent comparisons were obtained after the application of the Smith-Waterman algorithm in a pair wise manner to all predicted proteins encoded in both completely sequenced and unfinished genomes available in the public database *RefSeq*. Comparisons were run through a computational grid and the complete result reaches a volume of over 900 GB. Consequently, the database system design is a critical step in order to store and manage the information from comparisons' results. This paper describes database conceptual design issues for the creation of a database that represents a data set of protein sequence cross-comparisons. We show that our conceptual schema and its relational mapping enables users to extract relevant information, from simple to complex queries integrating distinct data sources.

Keywords: Database design, Pairwise alignment, Complete genomes.

1 Introduction

The availability of complete genome sequences of numerous organisms, combined with the computational progress occurred in the last few decades, provides an opportunity to use holistic approaches in the detailed study of the genome structure, as well as gene prediction and functional classification.

Among these approaches, we are mainly interested in the comparative genome analysis (or comparative genomics). It consists in the analysis and comparison of genetic material from diverse species (or strains), aiming at investigating their internal organization and evolution of the compared genomes (and the corresponding species). In addition, we are looking forward to revealing the function of genes and non-coding regions in these genomes.

M.C.P. de Souto and M.G. Kann (Eds.): BSB 2012, LNBI 7409, pp. 144–155, 2012.

This work reports results of the PWD[1] research project. It is an initiative dedicated to the comparison of protein information on a genomic scale. The goal is to improve the quality and interpretation of biological data, consequently, our understanding of biological systems and their interactions. Stringent comparisons were obtained after the application of the Smith-Waterman (SW) algorithm[15] in a pair wise manner to all predicted proteins encoded in both completely sequenced and unfinished genomes available at *RefSeq*[2] database.

Rigorous dynamic programming algorithms, such as SW, ensure the determination of the optimal alignment between pairs of sequences. In our case, we have run an implementation of the SW algorithm. However, due to their computational complexity, these algorithms are usually not suitable for the comparison of a large set of sequences. Therefore, we have considered distributed computing resources provided by the World Community Grid[3][12] to determine the sequence similarity level among almost 4 million proteins.

The result data available has reached a huge volume - over 900GB - of data, requiring a database system support. This is a fundamental factor if one wants to maximize the knowledge generation from the results yielded by the PWD project. Indeed, among others, we must consider data persistency, high availability and efficient access, all typical database technology features. Consequently, the database conceptual design becomes a relevant step to achieve the intended goals. Moreover, the corresponding logical (e.g. relational) schema would avoid performance bottlenecks, enable efficient querying and database maintenance.

This paper discusses conceptual modeling issues regarding the PWD project and all related database systems. We propose an extended conceptual schema in order to enable simple and complex queries, involving data obtained in our experiments and other relevant data sources, such as genomic sequences and taxonomies. We give then an overview of the database implementation issues, from the creation of the relational schema into a PostgreSQL DBMS up to samples of queries that define a formal access to our database.

It should be noted that there are many different research initiatives focusing on data modeling for bioinformatics (e.g. [6–8, 10, 11, 13, 16]). However, either there is no actual related project and the solutions proposed are not applicable to specific situations, or the conceptual models are so particular that we could not directly consider here. Furthermore, most approaches in the literature prioritize a systems view rather than a conceptual view. We claim that the modeling choices presented enforce the biological concepts and data integration issues.

This paper is organized as follows: we discuss next the motivation and the actual context of our research work, specifically some processing requirements. In Section 3 we discuss important issues with respect to conceptual data modeling, including some pros and cons of our modeling choices. Section 4 describes our relational schema and presents an overview of implementation and query answering. Finally, Section 5 concludes with future and ongoing work.

[1] `www.proteinworlddb.org`

[2] `http://www.ncbi.nlm.nih.gov/RefSeq/` (version 21).

[3] `http://www.worldcommunitygrid.org`

2 Motivation

Comparative genomics comprehends the comparison of two or more genomes, including genomic sequences and their predicted protein content, the relative positions of their genes and other genomic context features that may be of functional (or regulatory) importance. It also includes the study of gene structure and organization, the presence and location of repetitive sequences, polymorphisms and several other characteristics that may help to differentiate genomes[9].

A detailed analysis of the predicted protein contents of an organism is an important step in genome analysis, and it has been applied to several studies with different objectives. Cancer, for instance, is a class of diseases where modifications in the expression pattern of several genes confer new biological properties to the cell. A better understanding of these alterations may provide new insights for the development of diagnostic and treatment procedures.

An important task is the identification of all protein-coding genes and their location in the genome sequence, as well as the characterization of their functions. Genomic sequences are scanned, searching for protein-coding genes, using computational gene models. For each new genome, each predicted gene is conceptually translated into a protein sequence; the predicted collection of protein sequences is the predicted proteome of the organism. Each predicted protein is used as a query sequence in similarity searches against repositories of biological sequences. Significant matches are added to the genomic sequence together with the gene position and its product description. More sophisticated methods for the search of gene families are also used for annotation. Collectively, these methods provide predictions for the proteome of a newly sequenced organism[9].

Additional information about a proteome can be obtained through the comparison of the set of protein sequences against itself, which identifies *paralogs* (genes originated after duplication events), through the comparison among different proteomes for the identification of *orthologs* (genes originated after speciation events), by studying fusion or fission events or new domain arrangements, and by studying the evolution of cellular, metabolic and regulatory functions.

Genomic analyses presents important computational challenges and one fundamental step is the efficient storage and management of the information derived from DNA and protein sequences, alignments, functions and locations of genes, protein families and domains, relations between genes of different organisms and chromosomal rearrangements, among others. The database system must be logically organized in such a way that all types of information are readily accessible and may be rapidly fetched, even for a large volume of data.

In what follows, we discuss the results of our practical comparisons, our first and straightforward conceptual model to represent it and the extended model that involves many other relevant information that enable a rather complete view of the biological experiment.

3 Database Conceptual Modeling

In our experiments, a set of 3,812,663 proteins from *RefSeq* version 21- consisting of all predicted proteins encoded in 458 completely sequenced and unfinished genomes - and 254,609 proteins from Swiss-Prot version 51.5[4] were compared, in a pair wise manner, with the program SSEARCH[14]. We have configured SSEARCH with standard parameters, and an E-value cut-off equal to one.

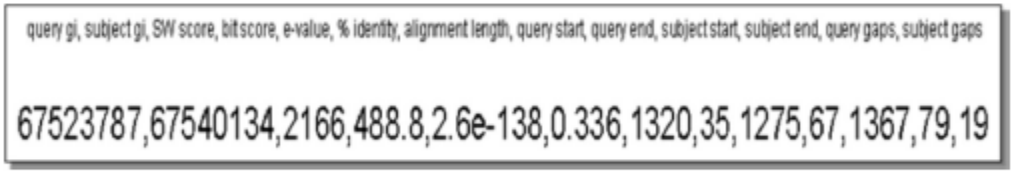


Fig. 1. A PWD match example report. The first line is the header of the listed values. The numbers are the ones that are stored.

For each significant match, a report is generated containing the identification of the pair of sequences compared and the alignment. The output format is given in Figure 1. A pair of protein sequences satisfying the required conditions to be stored was called a *hit*. A hit is defined by identifiers of the two sequences compared (Figure 1 has hit example between sequences *query_gi* = 67523787 and *subject_gi* = 67540134), and stores the validation measures of the pair wise comparison, besides additional information about the alignment, like similarity and coverage. The resulting matrix contains only hit information. The alignment itself was not stored.

Our main problem here was to define a database system that would help us for future querying and general data accesses. The goal is to store the results obtained in such a way that one could use these data together with other external data sources and generate relevant information. However, the whole system must consider the usual impedance mismatch among users offering a simple rather complete way to obtain the required information.

Figure 2 presents a first conceptual schema that can be output directly from the results. We have represented it with a conventional Entity- Relationship (ER) diagram, including min-max cardinalities. There are actually 3 possible combinations of hits involving translated ORFs and Proteins. All minimal cardinalities are zero as not all pair wise comparisons generate significant hits.

Results stored at the initial matrix contain only sequence identifiers and alignment information. The first step of this conceptual design was the creation of an entity that characterizes protein sequences. Information about the catalogued proteins compared in PWD includes the protein definition, its length and organism, and possible external references as protein identifiers (*RefSeq* and/or SwissProt). As the database must be kept up to date and updates occur, we

[4] `http://www.uniprot.org/`

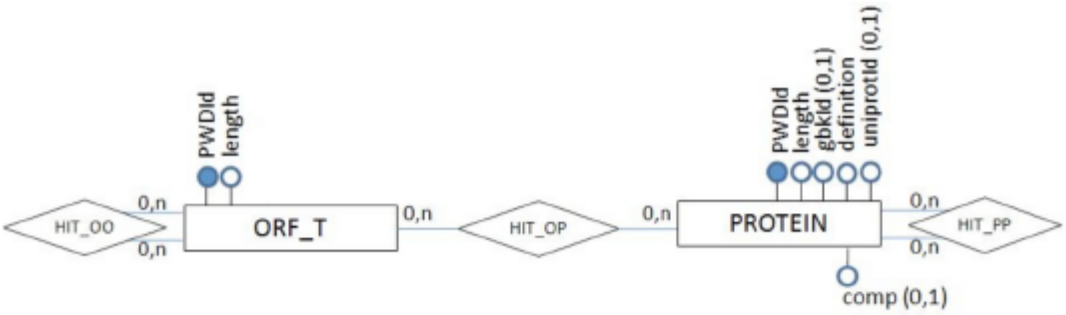


Fig. 2. First approach for a conceptual schema

identify those Proteins that have participated in registered comparisons. These are the main attributes of the Protein entity.

The amino acid sequences *translated_ORF* are represented by another entity *ORF_T* because they do not possess external identifiers. Information about these sequences includes the source organism, genomic location and length. Three types of distinct relationships between, proteins and ORFs are defined:

1. *hit_OO* is result of a comparison between *ORF_T* and *ORF_T*;
2. *hit_OP* is result of the comparison between *ORF_T* and proteins derived from SwissProt (proteins derived from *RefSeq* were not compared with *ORF_T*);
3. *hit_PP* is result of the comparison of *RefSeq* proteins with *RefSeq* and SwissProt proteins.

These relationships possess attributes that specify the pair wise results of PWD: $query_{gi}$, $subject_{gi}$, *SW score* (brute score of the comparison), *bit score* (normalized score), *e-value* (alignment significance), *%identity*, *alignment length*, *query_start*, *query_end*, *subject_start*, *subject_end*, *query_gaps*, *subject_gaps*.

However, we have some general and specific goals with this PWD project that cannot be solved only with the SSEARCH results and the corresponding output. There is a need of external data sources if one wants to check on the availability and feasibility. With respect to comparative genomics, hits represent only the result of protein-related genome comparisons. Further interesting questions depend on the protein coding gene physical position at its genomic context. For instance, the structure, organization and their genes relative position (gene order), and many other genomic features that may be of functional importance.

Therefore, Figure 3 gives an overview of our particular extended conceptual schema. We will discuss some of the data modeling alternatives and our design choices, which have guided us until the current conceptual schema.

We must observe that genes, transcripts, ORFs and genomic sequences are nucleotide sequences, while proteins and translated ORFs are amino acid sequences. The relationships between proteins and nucleotide sequences are constructed based on information from *RefSeq* version21.

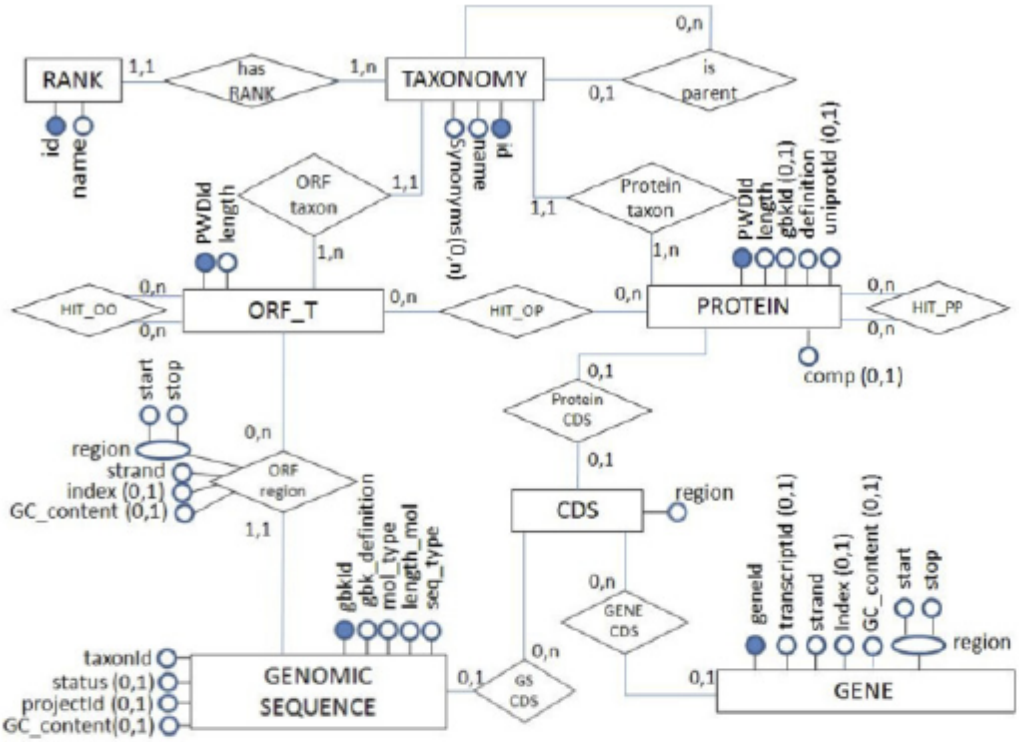


Fig. 3. A Conceptual Diagram for PWD Project

Conceptual Model Objects

We may explain entities, relationships and attributes in Figure 3 in more details. A protein is generated from a gene, which is a genomic sequence region. A protein coding gene is transcribed and produces a primary transcript that, after some processing, generates a mature transcript containing the protein coding sequence (CDS). The mature transcript is formed by the concatenation of sub-sequences containing information for the protein (exons) and untranslated regions (UTRs).

An ORF is a series of nucleotides codons extending up to the first termination codon. ORFs may not code for proteins. This way, all coding sequences (CDS) are ORFs but not all ORFs are coding for proteins.

Entity *Protein* represents the amino acid sequence of the protein that is related with the nucleotide sequence of CDS, and CDS with the gene and the genomic sequence containing it, keeping only an external reference to the transcript. Thus, *CDS* is an entity whose basic property is to keep up with the relationship between entities *Protein, Gene* and *Genomic Sequence*. This is done through the positioning of a given gene coding regions (exons) in the coordinate system of the genomic sequence that contains it. Each exon in a gene corresponds to a CDS sub-sequence, defined by an initial and a final position mapped into the coordinate system of the genomic sequence.

The entity *Gene* possesses an NCBI identifier - Entrez Gene[5], the region of its genomic sequence, defined by a start and a stop position, and a reading sense, its order in relation to the other genes in the genomic sequence, a transcript identifier (from *RefSeq*), and the GC-content. An amino acid sequence *ORF_T* refers to the genomic nucleotide sequence through an *ORF_region* delimited by a start and a stop position, inside the genomic sequence, with the reading sense, its position with respect to neighboring genes and the *GC-content*.

The *Genomic Sequence* entity possesses a *RefSeq* identifier, definition and sequence length, the type of organic molecule (DNA/RNA), status, type of sequence (chromosome, organelle, plasmid), an optional identifier of the respective genome project, GC content and an identifier of the source taxon.

Taxonomies and Classification

The classification used in PWD project is the same as the NCBI Taxonomy Database[2]. Each entry in the NCBI database is a *taxon*, also referred to as a *node*. The *root* node (taxid1) is at the top of the hierarchy. The path from the root node to any other particular *taxon* in this database is called its "lineage"; the collection of all of the nodes beneath any particular taxon is called its *subtree*.

In the conceptual model, the organism from which the genomic sequence was derived is the leaf node, defining the sequenced species (or an inferior rank like strain). It contains the taxID identifier from NCBI (a stable unique identifier for each taxon), the scientific and common names and synonyms. Each tree node has a rank, a parent node and may have descendent nodes. The taxonomic lineage may be obtained through a tree traversal from leaf nodes up to the root.

Conceptual Design Issues

We have first considered a database system exclusively oriented for the PWD project. Thus, the idea was to consider an entity called *Seq_AA* that would represent all compared amino acid sequences including annotated proteins and translated ORFs. This entity would relate with the hits matrix, and we would be able to specialize *Seq_AA* with either *RefSeq* or SwissProt as attributes. This entity would also be limited to sequences compared within the PWD project.

Within this particular conceptual schema, the amino acid sequences would relate with the nucleotide sequence through the entity *Coding_region*, and the latter with its source nucleotide sequence entity *Seq_NT* source. The problem of this representation is that it would be artificially adapting a biological concept, as ORFs were considered even if not coding regions. Moreover, another entity, *Seq_NT* source, also presented a wrong concept by dealing equally with both genomic and transcript sequences.

We have discussed some alternatives and decided to adopt the conceptual model depicted in Figure 3 due to the following reasons and modeling challenges:

- It is important to enable the database system to support updates as new genome sequences become available. It would be an error to limit the database only to the GCP project.

[5] `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene`

– The ORFs-T sequences, as a group of artificial (possible) proteins, brought many design problems. It becomes clear now that the sequences do not share the same characteristics with proteins, and need to be represented by and independent entity.
– Even if proteins with different origins could have distinct characteristics, with a conceptual viewpoint they all could be grouped as a single entity - Protein.

In our project, there are proteins that are originated from genomic projects including gene annotations, mRNA and CDS; those whose origins are only mRNA and proteins that are directly obtained from its sequencing process, without any reference to its original nucleotide. It brings another challenge for conceptual modeling and we have decided to model 3 types of amino acid sequences with specific characteristics and distinct research goals:

1. Protein sequences derived from finished genome projects (possessing the relationships genomic sequence gene transcript - CDS - protein) which will be the sequences considered in the comparative studies of proteomes, because they represent the complete predicted proteome of an organism;
2. Other protein sequences that may have these relationships are useful for the identification and validation of procedures and annotation results;
3. The experimental group of *ORF_T* will be used to evaluate the coding potential of these small ORFs, usually neglected by automatic gene prediction methods. These sequences were derived from complete prokaryotic genomic sequences. Therefore, the relationship *ORF_T* genomic sequence is mandatory, and validated sequences (with a high probability of being coding) may be included in the proteome comparative studies involving the database.

4 Relational Implementation and Queries

We have mapped our conceptual schema to a relational logical schema. Figure 4 shows the *Hits* similarity mapping. The recursive relationship *Hits* among *ORF_T* entity type participants were mapped for the table *hit_oo*. We also mapped recursive relationship *Hits* of *Proteins* (*hit_pp* table) and the relationship *Hits* among *Protein* and *ORF_T* (*hit_op* table).

Figure 5 shows the taxonomy relational mapping. The recursive relationship *is parent* is mapped to *taxonomy* table by the foreign key *taxonomy_id*. The relationships *have* among *ORF_T* and *PROTEIN* are mapped by the attribute *taxonomy_id* into *orf_t* and *protein* tables. We identify the attribute *synonymous* as a composite attribute. Then we have created the *synonymous* table that has different names for the same taxonomy.

Figure 6 shows the resulting logical schema diagram for our central dogma conceptual. Finally, we show in Figure 7 the complete logical schema diagram.

Queries: Validating the Relational Logical Data Model

With our relational schema in mind (Figure 7), it is possible to show how some relevant queries, involving most database objects, may be solved:

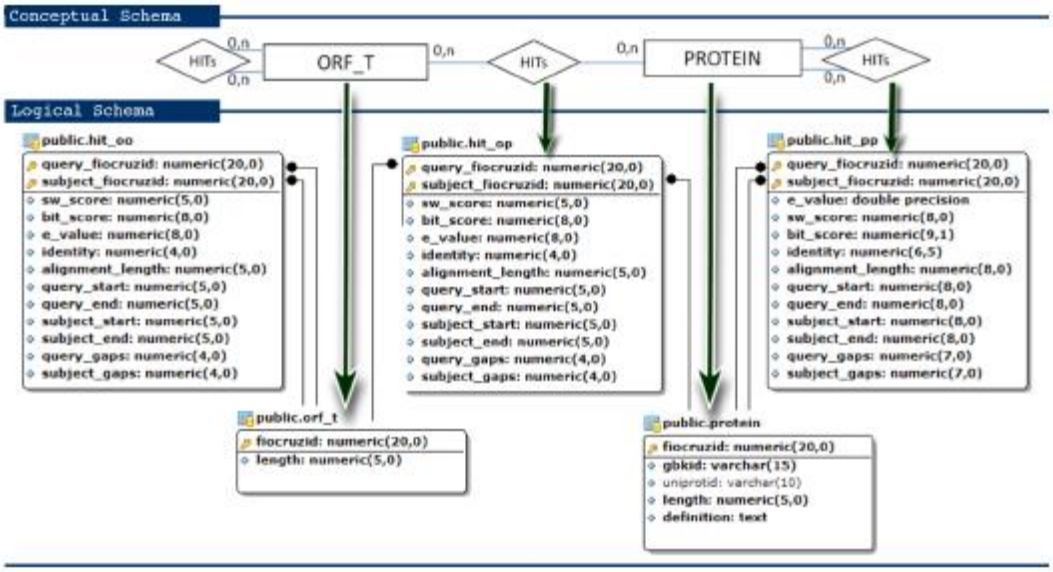


Fig. 4. Hits similarity mapping between translated ORFs and proteins

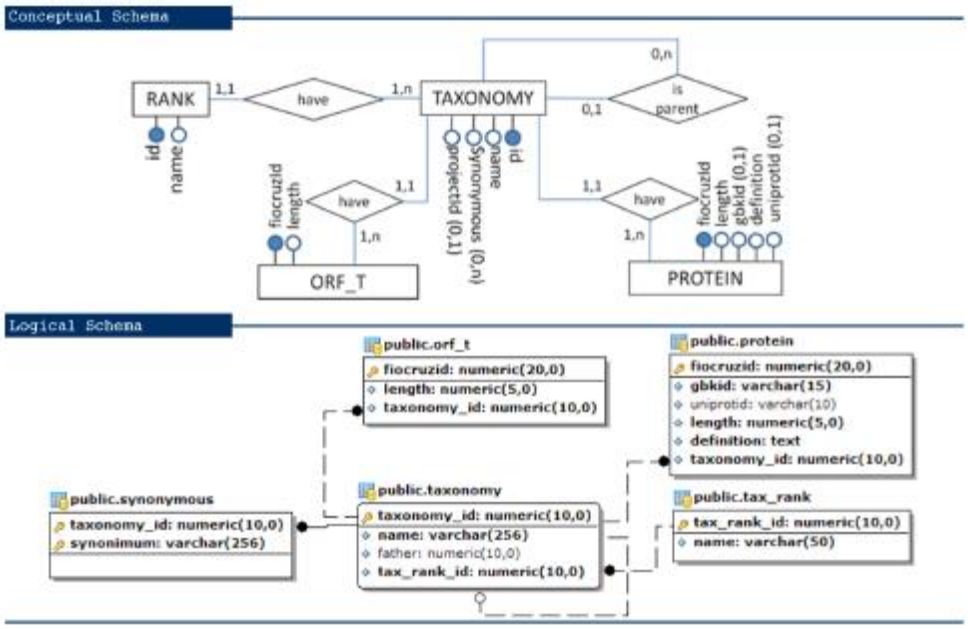


Fig. 5. Taxonomy mapping

1. Counting proteins in the database?

```
SELECT COUNT(DISTINCT p.fiocruzid)
FROM protein p
JOIN hit_pp_qid h ON p.fiocruzid = h.query_fiocruzid;
```

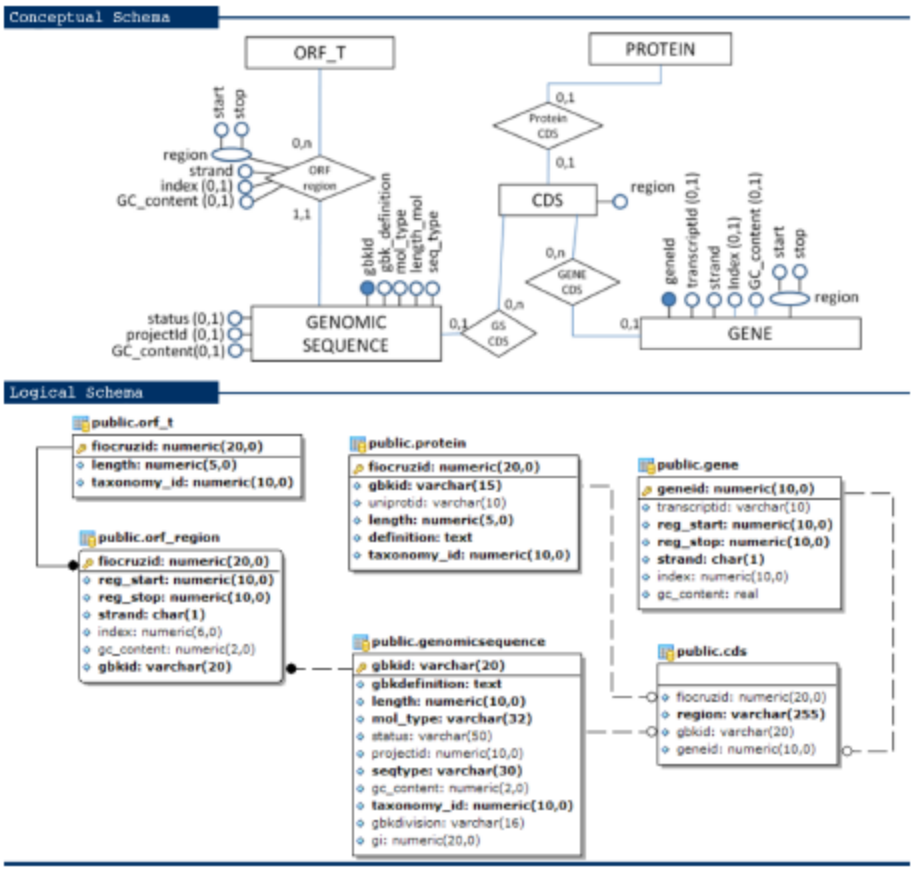


Fig. 6. Central Dogma Mapping

2. Proteins represented with genomic sequences?

```
SELECT p.fiocruzid, p.gbkid, p.definition
FROM genomic_sequence gs
JOIN cds ON gs.gbkid = cds.gbkid
JOIN gene g ON cds.geneid = g.geneid
JOIN protein p ON cds.fiocruzid = p.fiocruzid
JOIN hit_pp h ON h.query_fiocruzid = p.fiocruzid
```

3. How many genomes belong to a given (e.g. Vertebrata) taxonomy group?

```
SELECT gs.gbkid, gs.gbkdefinition, t.name
FROM genomic_sequence gs
JOIN taxonomy t ON gs.taxonomy_id = t.taxonomy_id
WHERE t.name LIKE '%Vertebrata%' ;
```

4. Return all hits above cut-off for a given protein X.

```
SELECT p.fiocruzid, p.definition, h.e_value, h.bit_score, h.sw_score
FROM hit_pp h
JOIN protein p ON h.subject_fiocruzid = p.fiocruzid
```

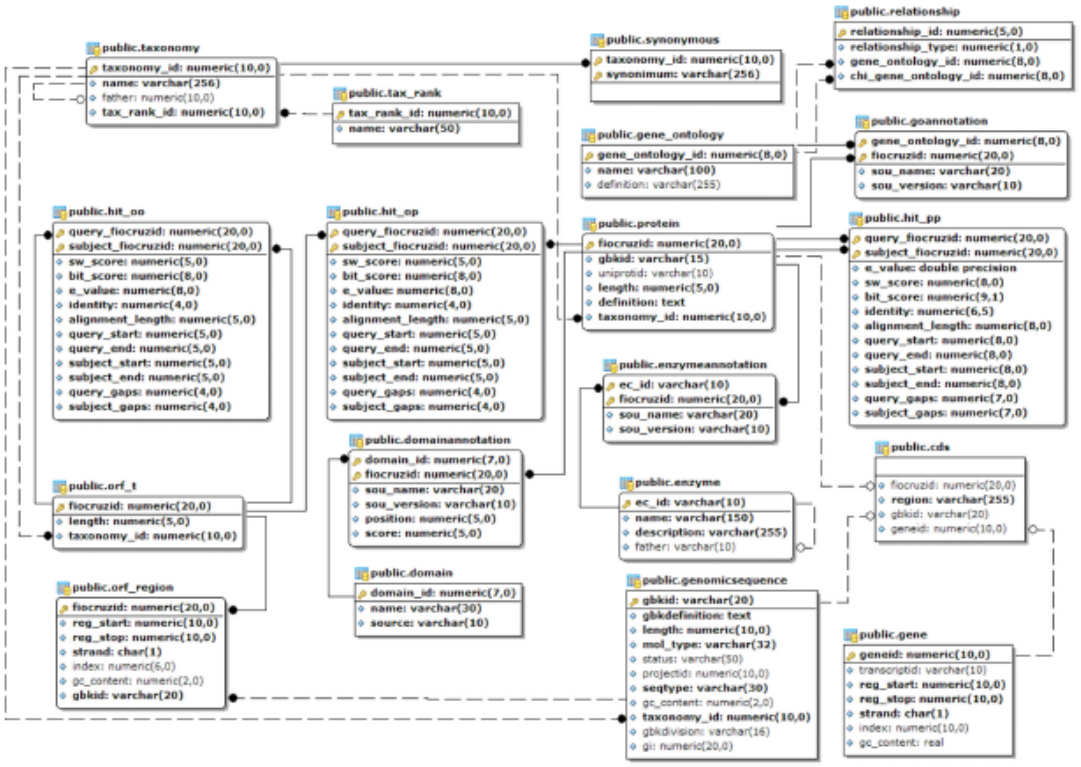


Fig. 7. A Logical Diagram for PWD Project

```
WHERE p.fiocruzid = 10957467
  AND (h.query_fiocruzid = 10957467 OR h.subject_fiocruzid = 10957467)
  AND h.e_value < 1.0e-5;
```

Complex queries are also relatively easy to follow if one considers our logical model depicted in Figure 7. Due to space limitations the reader is invited to check our project website.

5 Conclusions

We have discussed in this paper some conceptual modeling issues, including data modeling and queries, with respect to comparisons of protein information in a genomic scale. Due to the resulting data volume, the database system design becomes a critical step in order to extract significant information. We have discussed also implementation issues regarding our logical relational schema. We have implemented the logical model into PostgreSQL [3] as the underlying DBMS. Indexes access structures have been implemented to optimize some of the requested database queries.

Our main contributions rely on a general database system framework to represent sequence comparisons and the corresponding information with a fundamental and conceptual approach. This paper show an instantiation of our database schema considering data from the ProteinWorldDB project. We claim that many distinct queries, either simple or complex, may be stated in a straightforward

manner. The system has been implemented and its first version is already available to the public.

Many different loading scripts were developed and will become available to the public. Our ongoing and future work involve annotation procedures and external data sources, such as Pfam [5] for protein domains, KEGG [1] (metabolic pathways) and controlled vocabulary based upon GeneOntology[4]. We are also tuning our database system in order to support complex queries and additional procedures, such as the identification of unique genes, *paralogs*, *orthologs* and many others.

Acknowledgements. We wish to thank IBM®, World Community Grid™ for their support.

References

1. KEGG: Kyoto Encyclopedia of Genes and Genomes, `http://www.genome.jp/kegg/`
2. NCBI Taxonomy Database, `http://www.ncbi.nlm.nih.gov/Taxonomy/`
3. PostgreSQL, `http://postgresql.org`
4. The Gene Ontology, `http://www.geneontology.org/`
5. The Pfam Protein Families Database, `http://pfam.sanger.ac.uk`
6. Chen, J.Y., Carlis, J.V.: Genomic data modeling. Information, Special issue: Data Management in Bioinformatics 28, 287–310 (2003)
7. Elmasri, R., Ji, F., Fu, J., Zhang, Y., Raja, Z.: Modelling Concepts and Database Implementation Techniques For Complex Biological Data. International Journal of Bioinformatics Research and Applications 3, 366–388 (2007)
8. Keet, C.M.: Biological Data and Conceptual Modelling Methods. Journal of Conceptual Modeling (2003)
9. Mount, D.: Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press (2004)
10. Navathe, S.B., Kogelnik, A.M.: The Challenges of Modeling Biological Information for Genome Databases. In: Chen, P.P., Akoka, J., Kangassalu, H., Thalheim, B. (eds.) Conceptual Modeling. LNCS, vol. 1565, pp. 168–182. Springer, Heidelberg (1999)
11. Nelson, M.R., Reisinger, S.J., Henry, S.G.: Designing databases to store biological information. BIOSILICO 1, 134–142 (2003)
12. Otto, T.D., Catanho, M., Tristão, C., Bezerra, M., Fernandes, R.M., Elias, G.S., Scaglia, A.C., Bovermann, B., Berstis, V., Lifschitz, S., de Miranda, A.B., Degrave, W.: ProteinWorldDB: Querying radical pairwise alignments among protein sets from complete genomes. Bioinformatics (2010)
13. Pastor, O.: Conceptual Modeling Meets the Human Genome. In: Li, Q., Spaccapietra, S., Yu, E., Olivé, A. (eds.) ER 2008. LNCS, vol. 5231, pp. 1–11. Springer, Heidelberg (2008)
14. Pearson, W.: SSearch. Genomics 11, 635–650 (1991)
15. Smith, T., Waterman, M.: Comparison of Biosequences. Advances in Applied Mathematics 2, 482–489 (1981)
16. Zhou, X., Song, I.Y.: Conceptual Modeling of Genetic Studies and Pharmacogenetics. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005, Part III. LNCS, vol. 3482, pp. 402–415. Springer, Heidelberg (2005)

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 5 2010, pages 705–707
doi:10.1093/bioinformatics/btq011

Databases and ontologies

Advance Access publication January 19, 2010

ProteinWorldDB: querying radical pairwise alignments among protein sets from complete genomes

Thomas Dan Otto[1,2,*], Marcos Catanho[1], Cristian Tristão[3], Márcia Bezerra[3], Renan Mathias Fernandes[4], Guilherme Steinberger Elias[4], Alexandre Capeletto Scaglia[4], Bill Bovermann[5], Viktors Berstis[5], Sergio Lifschitz[3], Antonio Basílio de Miranda[1] and Wim Degrave[1]

[1]Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil, [2]Pathogen Genomics, Wellcome Trust Genome Campus, Hinxton, UK, [3]Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, [4]IBM Brasil, Hortolândia, São Paulo, Brazil and [5]IBM, Austin, TX, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Many analyses in modern biological research are based on comparisons between biological sequences, resulting in functional, evolutionary and structural inferences. When large numbers of sequences are compared, heuristics are often used resulting in a certain lack of accuracy. In order to improve and validate results of such comparisons, we have performed radical all-against-all comparisons of 4 million protein sequences belonging to the RefSeq database, using an implementation of the Smith–Waterman algorithm. This extremely intensive computational approach was made possible with the help of World Community Grid™, through the Genome Comparison Project. The resulting database, ProteinWorldDB, which contains coordinates of pairwise protein alignments and their respective scores, is now made available. Users can download, compare and analyze the results, filtered by genomes, protein functions or clusters. ProteinWorldDB is integrated with annotations derived from Swiss-Prot, Pfam, KEGG, NCBI Taxonomy database and gene ontology. The database is a unique and valuable asset, representing a major effort to create a reliable and consistent dataset of cross-comparisons of the whole protein content encoded in hundreds of completely sequenced genomes using a rigorous dynamic programming approach.

Availability: The database can be accessed through http://proteinworlddb.org

Contact: otto@fiocruz.br

Received on April 25, 2009; revised on January 6, 2010; accepted on January 9, 2010

1 INTRODUCTION

The assignment of biological function predictions and structural features to raw sequence data is typically accomplished by comparing them either to predicted protein sequences or to the corresponding genes. This information is stored in several primary public databases, such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) or EMBL-Bank (http://www.ebi.ac.uk/embl). However, annotations are often incomplete, based on non-standardized nomenclature or might have no value when inferred from previous incorrectly annotated sequences. Hence, secondary databases such as Swiss-Prot (http://www.expasy.ch/sprot/), PFAM (http://pfam.sanger.ac.uk) or KEGG (http://www.genome.ad.jp/kegg), to mention only a few, have been implemented to analyze specific functional aspects and to improve the annotation procedures and results.

Dynamic programming algorithms, or a fast approximation, have been successfully applied to biological sequence comparison for decades, and this class of algorithms comprises the heart of many well-known sequence alignment programs (Batzoglou, 2005). However, because of their quadratic time complexity, rigorous dynamic programming algorithms are usually not suitable for the comparison of a large set of sequences against a database, as they demand exceptionally huge computational power and are very time consuming. For this reason, sequence comparisons are generally performed by heuristics like BLAST (Altschul *et al.*, 1997) and FASTA (Pearson, 1990), which have proved to be quite effective and significantly faster than the dynamic programming algorithms. However, in many instances, these comparisons might lack accuracy, as these heuristics do not guarantee to find a mathematically optimal alignment (Pearson, 1990), therefore affecting all subsequent analytical steps. The Genome Comparison Project (GCP) (http://www.dbbm.fiocruz.br/GenomeComparison) aims to compare protein information on a genomic scale to improve the quality and interpretation of biological data and our understanding of biological systems and their interactions. Stringent comparisons were obtained after the application of the Smith–Waterman (SW) algorithm (Smith and Waterman, 1981) in a pairwise manner to all predicted proteins encoded in both completely sequenced and unfinished genomes available in the public database RefSeq (version 21). The project represents a joint effort involving Fiocruz, PUC-Rio and IBM®, and was executed through World Community Grid™ (WCG), a computational grid on a global scale. We present here the outcome of this joint effort, the ProteinWorldDB, which represents a major effort to create a reliable and consistent dataset of cross-comparisons of the whole protein content encoded in hundreds of completely sequenced genomes using a rigorous dynamic programming approach.

*To whom correspondence should be addressed.

2 METHODS

The core of the ProteinWorldDB comprises the results of all pairwise comparisons accomplished by the GCP. Briefly, a set of 3 812 663 proteins from RefSeq version 21—consisting of all predicted proteins encoded in 458 completely sequenced and unfinished genomes—and 254 609 proteins from Swiss-Prot version 51.5 were compared, in a pairwise manner, with the program SSEARCH (http://fasta.bioch.virginia.edu/), an implementation of the SW local alignment algorithm. The sample was partitioned in blocks containing up to 2000 sequences each, and comparisons were made applying standard parameters, with an E-value cutoff equal to one. To overcome distortions in the E-value and bit score produced by the partitioning of the data, we recalculate the statistical parameters Lambda and K for each aligned pair, taking the entire dataset into account, using four different mathematical models implemented in the SSEARCH algorithm: (i) a weighted regression of average score versus library sequence length, which provides an accurate estimate of whether an alignment score is likely to occur by chance (Pearson, 1998; Pearson and Sierk, 2005), (ii) estimation from the mean and standard deviation of the library scores, without correcting for library sequence length, (iii) maximum likelihood estimates of Lambda and K and the (iv) Altschul–Gish parameters (Altschul and Gish, 1996). For each comparison, a report containing sequence identifiers, alignment length, coordinates of the most similar regions, percentage of identity, number of gaps, raw and bit scores and E-value was returned. These central data were subsequently connected to several third-party annotations, including gene and protein features (RefSeq), taxonomic information (NCBI Taxonomy database), gene ontology (GO), functional classification (Swiss-Prot/TrEMBL), domain and protein family classification (Pfam) and enzymatic activity (KEGG). Additionally, we have clustered all proteins of the dataset. Two or more proteins are included in the same cluster if either their SW score or the combination of identity and overlap is greater than or equal to a certain threshold (Otto *et al.*, 2008). More than 40 complete sets of clusters, using different parameter settings, were generated and stored.

The ProteinWorldDB data are stored and managed using IBM® DB2 database management system, and are publicly accessible via a web-based graphical user interface. Currently, the following analyses are implemented:

(1) Query of annotation features by primary/secondary database identifiers, GO terms, EC numbers or Pfam terms. The records are returned in tabular form, including all aforementioned qualifiers, the genome name and its NCBI taxonomy ID. This is the standard output for most results.

(2) Return of all proteins stored in the database similar to a query sequence according to a certain qualifier. The user can limit the results using the E-value, percentage of identity, overlap area or SW score.

(3) Comparison of protein sequences not included in the database with all proteins in the dataset using BLAST algorithm. The first five hits, including their features, are returned.

(4) Download of the complete comparison data of two (fully sequenced) genomes. The number of hits displayed can be limited as in 3.

(5) Search for unique proteins encoded by each organism. Under a given cluster threshold, these proteins represent the sequences that have not been grouped with any other sequence.

(6) Query of groups of related proteins, based on the primary/secondary database identifiers, GO terms, EC numbers or Pfam terms.

3 RESULTS

ProteinWorldDB hosts a singular core dataset, composed of nearly 4 million proteins compared in a pairwise manner with the rigorous SW algorithm, which guarantees to find a mathematically optimal alignment for a given set of parameters. With the help of the WCG, the processing took ~7 months of calendar time (the equivalent of 3748 computer years, including an average 3-fold redundancy in the grid, which was simultaneously allocating resources to two other projects). The complete result occupies ~1 TB in a tabular form, each line comprising 80 characters for each alignment with an E-value ≤ 1. Of the 16×10^{12} comparisons executed, 4.2×10^9 are currently in the database (comprising 300 GB of data), corresponding to alignments with an E-value ≤ 0.001. Different groups compared subsets of sequences with a SW approach (Kanehisa *et al.*, 2006) or pairs were first filtered with a heuristic method and then compared, after satisfying a certain threshold (Rattei *et al.*, 2008). As previous studies have shown (Pearson, 1990; Uchiyama, 2007), the latter strategy is not guaranteed to find all hits. One should keep in mind that false positives are expected to be found with an E-value threshold of $E \leq 0.001$, as millions of comparisons were done. Nevertheless, function transfer and homology inference should not rely on E-value thresholds alone, since the fraction of identical positions between a pair of sequences, as well as the extension of their overlapping area, among several other sequence properties, play an important role in functional and evolutionary predictions based on sequence similarity (Boekhorst and Snel, 2007; Rost, 2002; Tian and Skolnick, 2003).

Valuable and unique information can be retrieved from ProteinWorldDB. For instance, queries could include: (i) individual or groups of proteins and their similarities with other entries based on the SW algorithm; (ii) download of subsets of the comparison data, i.e. related proteins shared by two particular species (inferred orthologs) or related proteins present in the same organism (inferred paralogs); (iii) genes that are exclusive of a particular species, i.e. taxonomically restricted (unique) genes; (iv) groups of related proteins for particular species using a protein of interest or a shared biological function as reference; and (v) comparison of different annotations for each entry. The ProteinWorldDB will, no doubt, contribute to improve annotation, to studies on genome and protein family evolution and in many other research aspects.

3.1 Further work

At this moment, the database contains similarity information using an E-value cutoff of 10^{-3}. Later on, we will add additional results up to an E-value of one, and comparisons of an experimental set of open reading frames, which have not been predicted as coding. Datasets comprising different phylogenetic experiments, phylogenomics and horizontal gene transfer are in construction. Also, an update can be envisaged with the WCG to compute all the genomes that were included in RefSeq since the end of our experiments. In the future, we hope to develop automatic algorithms to scan differences in annotation between third-party databases, evaluate the confidence of the annotations, add a wiki-like annotation support system, allowing other groups to include their expertise in the database, as well as refine the interface in order to allow more complex queries.

ACKNOWLEDGEMENTS

We wish to thank IBM®, World Community Grid™, Rede Fiocruz, Plataforma de Bioinformática PDTIS, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ),

Programa Estratégico de Apoio à Pesquisa em Saúde (PAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for their support.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.

Altschul,S.F. *et al.* (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Batzoglou,S. (2005) The many faces of sequence alignment. *Brief. Bioinform.*, **6**, 6–22.

Boekhorst,J. and Snel,B. (2007) Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. *BMC Bioinformatics*, **8**, 356.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, **34**, 354–357.

Otto,T.D. *et al.* (2008) AnEnPi: identification and annotation of analogous enzymes. *BMC Bioinformatics*, **9**, 544.

Pearson,W. (1990) Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol.*, **183**, 63–98.

Pearson,W. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.

Pearson,W. and Sierk,M.L. (2005) The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.*, **15**, 254–260.

Rattei,T. *et al.* (2008) SIMAP structuring the network of protein similarities. *Nucleic Acids Res.*, **36**, D289–D292.

Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **3318**, 595–608.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.

Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.*, **35**, D343–D346.

III. PROJETO COMPARAÇÃO DE GENOMAS – PCG

O Projeto Comparação de Genomas (PCG)[143], uma iniciativa do grupo de pesquisas genômicas do Laboratório de Genômica Funcional e Bioinformática (LGFB) do Instituto Oswaldo Cruz / Fiocruz, comparou aproximadamente quatro milhões de sequências de aminoácido e gerou como resultado uma matriz com quase 900 GB, que precisava ser estruturada, armazenada e disponibilizada para a comunidade científica.

Mais detalhadamente, o PCG foi planejado visando:

- A construção de um repositório de referência para a comunidade de anotadores, proporcionando uma fonte de dados confiável para pesquisas. Utilizando-se os índices de similaridade, armazenados na matriz de resultados, juntamente com uma nomenclatura padronizada de genes e seus produtos (Gene Ontology[144]) e de outras bases de dados[145] de proteínas, para dar suporte ao processo de anotação, e fazer um cruzamento de anotações de diferentes tipos;
- Permitir pesquisas avançadas de genômica comparativa, com base na similaridade entre o conjunto total de sequências proteicas de genomas completamente sequenciados ("proteomas preditos").

Para as comparações, utilizou-se uma estrutura distribuída de *grid* fornecida pelo *World Community Grid* (WCG)[146]:

Execução

- Processamento no WCG:
  - Início: 20 de dezembro de 2006
  - Término: 21 de julho de 2007
  - Total: 07 meses de calendário;

[143] http://www.dbbm.fiocruz.br/labwim/bioinfoteam/templates/archives/GenomeComparison/GenomeComparison http://www.worldcommunitygrid.org/research/fcg1/overview.do
[144] http://www.geneontology.org.
[145] Com informações funcionais, de domínios, e classes enzimáticas.
[146] http://www.worldcommunitygrid.org/

- Quantidade de sequências proteicas comparadas:
  - 3.812.663 sequências da base de dados NCBI-RefSeq[147] v.21;
  - 254.609 sequências da base de dados UniProt-Swiss-Prot[148] v.51.5;
  - Um conjunto de quase três milhões de sequências tORFs.

- Comparação:
  - *Par-a-par* e *todas-contra-todas*[149]
  - Programa: SSEARCH[150,151];
  - Parâmetros: padrão do SSEARCH;
  - Valor de corte[152]: E-value = 1;
  - Número de comparações realizadas: 16×10^{12}.

Formato do Resultado

- Na FIGURA III.1 temos o exemplo do resultado de uma comparação gerada pelo SSEARCH:
  - A sequência consulta = *query_gi* = 67523787
  - A sequência subject = *subject_gi* = 67540134

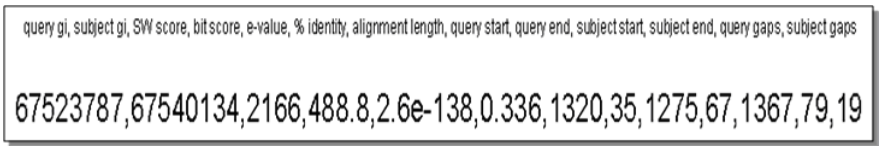


FIGURA III.1. Exemplo do resultado de uma comparação produzido pelo SSEARCH. A matriz de resultados do PCG é composta apenas por registros com este formato, onde somente os valores das comparações são armazenados. A linha superior da figura apresenta apenas os descritores dos valores, que estão especificados na TABELA III.1.

- Resultados armazenados:
  - Apenas as informações dos alinhamentos (FIGURA III.1 e TABELA III.1);

[147] (http://www.ncbi.nlm.nih.gov/refseq/). Todas as proteínas preditas codificadas em 458 genomas (completos e não finalizados).
[148] http://www.uniprot.org/
[149] Comparou-se todas as sequências de proteína presentes em cada genoma de cada organismo entre si e entre todas as sequências de proteína presentes nos genomas de todos os outros diferentes organismos.
[150] [Pearson WR, 1991]; http://fasta.bioch.virginia.edu/.
[151] Uma implementação do algoritmo de Smith-Waterman [Smith and Waterman, 1981], o qual encontra o melhor alinhamento local (do ponto de vista matemático) entre pares de sequências.
[152] Só foram armazenados resultados de similaridades com e-value <= 1.

- Apenas pareamentos com valor de similaridade estatisticamente significativo (e-value <= valor de corte);
- O alinhamento entre o par de sequências comparado não foi armazenado;
- Número de alinhamentos significativos: 4.2×10^9;
- Quantidade de dados da matriz (descompactados): 900 Gb.

Parâmetro	**Descrição**
SW score	Pontuação obtida para o alinhamento de duas sequências, de acordo com uma matriz de substituição particular
Bit score	Bit score. Pontuação normalizada
E-value	Valor esperado ou E-value. Representa o número de alinhamentos com a mesma pontuação ou maior esperado ao acaso
% identity	Fração de posições idênticas para um dado alinhamento.
Alignment length	Comprimento do alinhamento.
Query start	Posição inicial da sequência consulta no alinhamento.
Query end	Posição final da sequência consulta no alinhamento.
Subject start	Posição inicial da sequência no alinhamento.
Subject end	Posição final da sequência comparada no alinhamento.
Query gaps	Nº de gaps introduzidos na sequência consulta durante o alinhamento.
Subject gaps	Nº de gaps introduzidos na sequência comparada durante o alinhamento.

TABELA III.1: Descrição dos parâmetros listados no resultado do SSEARCH da comparação de um par de sequências.

ENFOQUE COMPUTACIONAL

Sob o ponto de vista computacional, podemos ver o problema da quantidade de dados a ser gerenciada como um desafio. 900 Gb referem-se somente à matriz de resultados das comparações, valor que se tornará bem maior com a adição de dados das bases que serão associadas para dar suporte ao processo de anotação e pesquisas genômicas.

Uma utilização eficiente desses resultados está diretamente relacionada com a forma como estes dados serão armazenados e as facilidades para execução de consultas complexas e do tempo de resposta, isto é, do projeto do banco de dados e facilidades de gerenciamento desses dados.

Com relação ao sistema de banco de dados que dará suporte às pesquisas biológicas, deve ser considerado:

- A questão de persistência dos dados (com uma ordem de grandeza de terabytes);
- A questão de acesso e busca eficientes;
- Soluções eficientes para que não ocorram gargalos de processamento, em termos da relação entre os dados em disco e memória RAM;
- O projeto da base de dados de forma que, em termos de software, atenda às necessidades de busca;
- Versões futuras incrementais e atualização do banco de dados.

ENFOQUE BIOLÓGICO

O ponto inicial para as pesquisas biológicas é a matriz de resultados do PCG que contém os índices de similaridades resultantes das comparações entre o conjunto total de sequências proteicas ("proteoma predito") codificado nos diversos genomas considerados.

O segundo ponto, que permitirá o estudo de diferentes questões de genômica comparativa, requisitadas pelos organizadores do PCG, dependerá das bases de dados integradas a essa matriz, definindo e limitando o domínio que fornecerá os contornos básicos para o projeto conceitual do PWDB.

De acordo com a documentação e conversas com a equipe do PCG, algumas das questões a serem analisadas, através de consultas diretas ao banco de dados, ou pela seleção de conjuntos de dados intermediários para utilização em procedimentos posteriores são:

- Obter uma anotação mais precisa das sequências comparadas, de modo a oferecer um repositório de referência para a comunidade de anotadores, proporcionando uma fonte de dados confiável para pesquisas;
- A atribuição de possíveis funções a proteínas hipotéticas de função desconhecida,
- Identificação de proteínas com múltiplos domínios e elementos funcionais,
- Detecção de relacionamentos distantes entre proteínas;
- Relacionamentos evolutivos entre proteínas para uma melhor compreensão da organização genômica, sua evolução e funções celulares;
- Estudos evolutivos de genomas e famílias de proteínas;

- Compreensão do conteúdo protéico total de uma célula, das interações entre as proteínas e das vias bioquímicas e sua regulação;
- Análise da biodiversidade, através do estudo de diferentes aspectos da genética e bioquímica dos organismos;
- Entendimento de relacionamentos parasito-hospedeiro;
- Busca de novos métodos diagnóstico assim como o desenvolvimento de novas drogas e vacinas;
- Tentativa de descobrir padrões incomuns de codificação em sequências genômicas de procariotos, utilizando os resultados do grupo de sequências de aminoácidos derivadas de ORFs[153] que não foram identificadas como codificadoras de proteína através dos métodos computacionais clássicos.

DIFERENCIAL DO PCG

Um ponto importante e inovador desse projeto é a utilização do software SSEARCH para as comparações, e a parceria com o *World Community Grid* – IBM, que disponibilizou a capacidade ociosa de computadores pessoais de voluntários em todo o mundo, através de uma infraestrutura de computação distribuída, sem a qual esse volume de comparações, utilizando este método de comparação em especial, não seria possível.

O PCG é único no que diz respeito à utilização do SSEARH para comparação desse volume de dados. Pode-se citar o SIMAP (*the Similarity Matrix of Proteins* [Rattei *et al.*, 2006, 2010]), como um trabalho similar, mas este utiliza o Blast [Altschul *et al.* 1990, 1997] nas comparações e, inclusive, os autores comentam que a solução ótima para gerar uma matriz de similaridade seria a aplicação do algoritmo de alinhamento Smith–Waterman (SW) e o subsequente armazenamento dos *scores* significativos[154]. E concluem[155] que apesar de existirem implementações eficientes do algoritmo SW os custos computacionais seriam totalmente inviáveis. Assim, no SIMAP, é utilizado o método heurístico Blast para acelerar a busca nos bancos de dados.

[153]O PCG resolveu utilizar essas sequências nas comparações numa tentativa de identificar possíveis sequências candidatas a serem codificadoras, não detectadas pelos métodos automatizados, através do grau de similaridade com sequências de proteína já conhecidas. "Non-coding_ORF" (ncORF) e "translated_ORF" (tORF) são termos definidos para o esquema conceitual do PWDB v.2. Não são proteínas cadastradas em bancos de dados de sequências. Maiores detalhes, anteriormente, em DISCUSSÃO.
[154] Essa foi exatamente a metodologia empregada pelo PCG.
[155] Citando [Rognes *et al.*, 2000].

COMPARAÇÃO DE SEQUÊNCIAS

Buscas por similaridade de sequências em bancos de dados é um processo repetitivo, em que as mesmas buscas são refeitas frequentemente, e, normalmente, é apenas o passo inicial para direcionar estudos posteriores. Assim, pode-se frisar mais um ponto importante sobre o PCG: o projeto comparou todo o conteúdo protéico codificado nos genomas de centenas de organismos, incluindo o homem e diversas outras espécies de interesse médico, comercial, industrial ou de importância em pesquisa como organismos-modelo. Foram comparações "*par-a-par*" e "*todas-contra-todas*", e os resultados de comparações considerados estatisticamente significativos foram armazenados numa matriz.

Isso significa que para a obtenção de informações sobre similaridades e, em última análise, homologias, não é necessário comparar mais de uma vez, por exemplo, o proteoma de um genoma X com o proteoma de um genoma Y, se X e Y foram comparados no PCG. O resultado das comparações (normalmente a parte mais demorada e computacionalmente intensiva da pesquisa) já está armazenado na matriz do PCG. Consultas de similaridade entre proteínas de interesse pertencentes a X e/ou Y, podem ser recuperadas diretamente da matriz do PCG.

BANCO DE DADOS PROTEINWORLDDB (PWDB) V.1[156]

O PCG gerou resultados importantes e bastante úteis para a comunidade científica e resultados de similaridade entre sequências de proteína de centenas de genomas comparados no PCG podem ser facilmente recuperados, além de inúmeras outras consultas, no banco de dados PWDB v.1.

O PWDB v.1 foi desenvolvido através de uma parceria entre o LGFB do IOC/FIOCRUZ e o Laboratório de Tecnologias de Gerência de Dados em Bioinformática (LaBBio) da PUC-Rio e IBM.

[156] (http://157.86.176.108/ProteinWorldDB/default.php). Ver DISCUSSÃO.

IV. BASES DE DADOS

IV.1 NCBI REFSEQ

Parte 1:

- A base de dados RefSeq é uma coleção de sequências, não redundante e ricamente anotada. O projeto fornece padrões de sequências de referência para as moléculas do dogma central, que ocorrem naturalmente na natureza, desde cromossomos, mRNAs até proteínas (TABELA IV.1.1).

RefSeq Release 61 Statistics
Release date **Sep 09, 2013**

Number of Accessions: 41.958.567
==========
Directory: complete

Number of **taxids:** 29.414

molecule type	Number of Accessions
Genomic:	4.291.237
RNA:	4.528.216
Protein:	33.139.114

TABELA IV.1.1. Estatísticas da versão 61 da base de dados RefSeq.

- Inclui sequências de archaea, bactérias, eucariotos, vírus, plasmídeos e organelas.
- É um recurso único pois provê um amplo banco de dados de sequências multi-espécie e curado, com registros separados porém com links explícitos de genomas com seus transcritos e produtos traduzidos (quando apropriado).
- A coleção RefSeq é o resultado da extração de dados de submissões INSDC, curadoria e computação, combinados com uma ampla colaboração com grupos oficiais. Cada molécula é anotada com a maior precisão possível. Se múltiplas submissões do INSDC representam a mesma molécula para um organismo, a "melhor" sequência é escolhida para representar o registro RefSeq.
- Semelhante a um artigo de revisão, um registro RefSeq é uma síntese das informações disiponíveis em múltiplas fontes de dados num determinado momento.
- Registros RefSeq fornecem uma fundamentação, unificando dados de sequências com informação genética e funcional. São gerados para prover padrões de referência com objetivos diversos desde anotação de genomas à descrição da localização de variações de sequências em registros médicos.
- Em oposição à redundância de sequências encontrada em repositórios públicos como o INSDC, a coleção RefSeq busca fornecer, para cada espécie incluída, um conjunto completo

de dados não redundante, com extensivas referências cruzadas, e registros de sequências de ácido nucleico e proteína ricamente anotados.

- A natureza não redundante da coleção RefSeq facilita pesquisas baseadas na localização genômica, sequência ou anotação textual.
- As sequências são validadas para confirmar se a sequência genômica correspondente a um mRNA anotado coincide com o registro da sequência de mRNA, e se a região CDS traduz corretamente a sequência de proteína correspondente.
- A cobertura e finalização de sequências genômicas disponíveis publicamente varia entre organismos e, desta forma, existem registros genômicos intermediários em algumas circunstâncias.
- A coleção RefSeq está disponível sem restrições e pode ser recuperada através de buscas (via Blast), links disponíveis no NCBI[157] e através do RefSeq FTP *site*.
- A coleção RefSeq permite uma base útil para a integração de diversos tipos de dados, incluindo sequências, genética, expressão, e informação funcional, em um consistente *framework* com um conjunto uniforme de convenções e *standards*.
- A coleção RefSeq suporta as seguintes atividades:
  - Anotação genômica,
  - Caracterização de genes,
  - Genômica Comparativa,
  - Descrição de variações de sequências,
  - Estudos de expressão.
- Cada sequência RefSeq tem associado um número de acesso estável (ACCESSION NUMBER), um número para a versão (VERSION) e um identificador inteiro (GI)[158].
- O número GI e a VERSÃO são incrementados quando a sequência é atualizada, enquanto que o ACCESSION permanece o mesmo. O conjunto de identificadores GI e "ACCESSION.VERSION", provê a melhor resolução de referência para uma sequência.
- Registros RefSeq podem ser distinguidos dos outros registros INSDC pela existência de um sinal *underscore* (_) na terceira posição do ACCESSION NUMBER[159].
- O campo COMMENT indica o nível de revisão que um registro possui (FIGURA IV.1.1 e TABELA IV.1.2).
- Versões obsoletas estão geralmente disponíveis se a sequência é atualizada.
- Os prefixos de acesso têm um significado implícito com relação ao tipo de molécula que representam (TABELA IV.1.3 e IV.1.4).

[157] Incluindo PUBMED, Nucleotide, Protein, Gene, and Map Viewer.
[158] O número GI é um identificador único, interno do NCBI. Cada sequência de nucleotídeo e de proteína cadastrada no NCBI tem um número GI associado (fonte: glossário NCBI).
[159] Números de acesso no DDBJ/EMBL/GenBank nunca incluem o símbolo (_).

```
LOCUS       NP_061223         246 aa        linear      ROD 28-JUL-2013
DEFINITION  14-3-3 protein beta/alpha [Mus musculus].
ACCESSION   NP_061223
VERSION     NP_061223.2  GI:31543974

.......
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AL591542.20.
            On Jun 9, 2003 this sequence version replaced gi:9055384.

            Sequence Note: The RefSeq transcript and protein were derived from
            genomic sequence to make the sequence consistent with the reference
            genome assembly. The genomic coordinates used for the transcript
            record were based on alignments.

            Publication Note:  This RefSeq record includes a subset of the
            publications that are available for this gene. Please see the Gene
            record to access additional publications.

            ##Evidence-Data-START##
            Transcript exon combination :: AK004872.1, AK144061.1 [ECO:0000332]
            ##Evidence-Data-END##

....
```

FIGURA IV.1.1. Registro Refseq NP_061223[160]. Pode-se ver: GI, ACCESSION, VERSION, COMMENT.

DESCRIÇÃO DOS CÓDIGOS DE REVISÃO

MODEL The RefSeq record is provided by the NCBI Genome Annotation pipeline and is not subject to individual review or revision between annotation runs.

INFERRED The RefSeq record has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.

PREDICTED The RefSeq record has not yet been subject to individual review, and some aspect of the RefSeq record is predicted.

PROVISIONAL The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff.

REVIEWED The RefSeq record has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature. Some RefSeq records may incorporate expanded sequence and annotation information.

VALIDATED The RefSeq record has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review at which time additional functional information may be provided.

WGS The RefSeq record is provided to represent a collection of whole genome shotgun sequences. These records are not subject to individual review or revisions between genome updates.

TABELA IV.1.2. Código de revisão de registros.

[160] NP_: proteína com accession NM_: mRNA ou NC_: Genomic, Complete genomic molecule, usually reference assembly (TABELA IV.1.).

ACCESSION PREFIX	MOLECULE TYPE COMMENT
AC_	Genomic Complete genomic molecule, usually alternate assembly
NC_	Genomic Complete genomic molecule, usually reference assembly
NG_	Genomic Incomplete genomic region
NT_	Genomic Contig or scaffold, clone-based or WGS[a]
NW_	Genomic Contig or scaffold, primarily WGS[a]
NS_	Genomic Environmental sequence
NZ_[b]	Genomic Unfinished WGS
NM_	mRNA
NR_	RNA
XM_[c]	mRNA Predicted model
XR_[c]	RNA Predicted model
AP_	Protein Annotated on AC_ alternate assembly
NP_	Protein Associated with an NM_ or NC_ accession
YP_[c]	Protein
XP_[c]	Protein Predicted model, associated with an XM_ accession
ZP_[c]	Protein Predicted model, annotated on NZ_ genomic records

[a] Whole Genome Shotgun sequence data
[b] An ordered collection of WGS sequence for a genome,
[c] Computed.

TABELA IV.1.3. Prefixos RefSeq e tipos de moléculas.

Molecule Type	Accession Prefix
protein	NP_; XP_; ZP_; AP_; YP_;
rna	NM_; NR_; XM_; XR_
genomic	NC_; NG_; NT_; NW_; NZ_; NS_; AC_

TABELA IV.1.4. Resumo por tipo de molécula da TABELA IV.1.3.

Parte 2

– Moléculas:

/**mol_type** =

genomic DNA, genomic RNA, mRNA, tRNA, rRNA, snoRNA, snRNA, scRNA, pre-RNA, tmRNA, viral cRNA, other RNA, other DNA, unassigned DNA, unassigned RNA.

/**organelle**: tipo de estrutura intracelular, limitada por membrana, a partir da qual foi obtida a sequência =

mitochondrion, nucleomorph, plastid,

mitochondrion:kinetoplast, plastid:chloroplast, plastid:apicoplast, plastid:chromoplast, plastid:cyanelle, plastid:leucoplast, plastid:proplastid,

/**plasmid** =

Nome do plasmídeo de ocorrência natural a partir do qual foi obtida a sequência. A estrutura plasmídeo é definida como uma unidade genética replicante de forma independente que não pode ser descrita por cromossoma ou segmento.

– Sequenciamento de Genomas:

Status: propriedade que se refere ao estágio atual do projeto de sequenciamento.

- Complete – tipicamente significa que cada cromossomo está representado por apenas uma sequência (scaffold) com montagem de alta qualidade,
- Assembly – tipicamente significa que existem montagens (scaffolds) que ainda não estão no nível de cromossomo e/ou sequências *draft*,
- In Progress – indica que o projeto de sequenciamento está numa fase de pré montagem, ou as sequências montadas/completas ainda não foram submetidas ao GenBank/ EMBL/ DDBJ.

Prefixos

- Prefixo NC_: foram obtidas por procedimento automatizado e revisão de especialista para alguns registros.
- Prefixos NT_, NW_, NZ_: indicam registros que não são individualmente revisados; as atualizações do genoma são liberadas como blocos.

Sequências

- Sequências *Draft*: sequências de DNA que ainda não estão finalizadas mas geralmente possuem alta qualidade (uma acurácia maior ou igual a 90%). Geralmente são fragmentos com 10.000 pares de bases. As posições cromossômicas aproximadas destes fragmentos são conhecidas.
- Sequências *Finished*: sequências com alta qualidade, taxa de erro baixa e sem *gaps*. É permitido apenas um erro a cada 10.000 bases (i.e., uma acurácia de 99.999%).

Referências:

http://www.ncbi.nlm.nih.gov/RefSeq/

RefSeq Help - Bethesda (MD): National Center for Biotechnology Information (US); 2011, http://www.ncbi.nlm.nih.gov/books/NBK50680/

The NCBI Handbook, Editors: Jo McEntyre, Jim Ostell, National Center for Biotechnology Information, Bethesda (MD): National Center for Biotechnology Information (US); 2002, http://www.ncbi.nlm.nih.gov/books/NBK21101/

http://www.jgi.doe.gov/education/genomics_1.html

IV.2 UNIPROT

Características:

- O *Universal Protein Resource* (UniProt) é um recurso abrangente para dados de sequências de proteína e anotação. É uma colaboração entre o *European Bioinformatics Institute* (EMBL-EBI)[161], o *Swiss Institute of Bioinformatics* (SIB)[162] e o *Protein Information Resource* (PIR)[163].
- A missão do UniProt é fornecer à comunidade científica um amplo recurso de sequências de proteína e informação funcional, com alta qualidade e acesso gratuito.
- Os bancos de dados do UniProt são: UniProt Knowledgebase (UniProtKB), o UniProt Reference Clusters (UniRef), e o UniProt Archive (UniParc). E o banco de dados UniProt Metagenomic e Environmental Sequences (UniMES) é um repositório especificamente desenvolvido para dados metagenômicos e ambientais.
- Os dados de sequências padrão do UniProtKB são:
  - Sequências codificadoras (CDS) traduzidas do DDBJ/ENA/GenBank (INSDC)[164],
  - Sequências de estruturas do PDB[165],
  - Sequências do Ensembl[166] e RefSeq[167],
  - Dados derivados de sequências de aminoácido submetidas diretamente ao UniProtKB ou obtidas da literatura.
- A parte central das atividades do Consórcio UniProt – UniProtKb – é um banco de dados de proteína, ricamente curado por especialistas, consistindo de duas seções: UniProtKB/Swiss-Prot e UniProtKB/TrEMBL

  - *UniProtKB/Swiss-Prot*
    - É a seção manualmente anotada e revisada. Um banco de dados de sequências de proteína não redundante e com anotação manual de alta qualidade, que agrupa resultados experimentais, características computadas e conclusões científicas.

[161] http://www.ebi.ac.uk/
[162] http://www.isb-sib.ch/
[163] http://pir.georgetown.edu/
[164] http://www.insdc.org/
[165] http://www.wwpdb.org/
[166] http://www.ensembl.org/index.html
[167] http://www.ncbi.nlm.nih.gov/refseq/

- A anotação manual consiste na análise, comparação e fusão de todas as sequências disponíveis para uma dada proteína, assim como uma revisão crítica de dados associados – experimentais e preditos.
- Os curadores UniProt extraem informação biológica da literatura e executam numerosas análises computacionais.
- O objetivo do UniProtKB / Swiss-Prot é prover todas as informações relevantes sobre uma proteína particular. Ele descreve, num único registro, os diferentes produtos de uma proteína derivados de um certo gene de uma dada espécie, incluindo cada proteína derivada (*splicing* alternativo, polimorfismos e/ou modificações pós-translational).
- As prioridades e processos de curadoria do UniProt estão documentadas em www.uniprot.org/help/biocuration.

- *UniProtKB/TrEMBL*
  - Contém registros analisados computacionalmente de alta qualidade, enriquecidos com anotação automática e classificação.
  - Registros são selecionados para uma anotação manual completa e posteriormente integrados ao UniProtKB/Swiss-Prot de acordo com algumas prioridades definidas para anotação.

– O UniProtKb e, em particular, o UniProtKB/Swiss-Prot é utilizado para acessar informações funcionais de proteínas. Cada registro UniProt contém a sequência de aminoácido, o nome da proteína ou descrição, dados de taxonomia e informação de citações, e além disso são adicionadas o maior número de anotações possíveis. Isso inclui ontologias biológicas, classificações e referências cruzadas, assim como indicações claras sobre a qualidade da anotação na forma de atribuição de evidência dos dados experimentais e computacionais.

– O atibuto PROTEIN EXISTENCE identifica o tipo de evidência que suporta a existência da proteína (TABELA IV.2.1). Não fornece informação de acurácia ou afirmação da sequência estar correta.
  - Apesar de fornecer informação da existência de uma proteína, pode acontecer de a sequência ser ligeiramente diferente, especialmente para sequências derivadas de modelos de genes preditos de sequências genômicas.
  - Apenas os níveis mais altos ou mais confiáveis de suporte de evidência da existência de uma proteína são exibidos para cada entrada. Por exemplo, se a existência de uma

proteína é suportada pela presença de ESTs e sequenciamento direto da proteína, será selecionado o valor Evidence at *protein level*.

Protein level	Indica que existe uma clara evidência experimental para a existência da proteína (sequenciamento de Edman parcial ou completo, clara identificação através de espectrometria de massa, estrutura de raio-X ou NMR, interação proteína-proteína de boa qualidade ou detecção da proteína através de anticorpos).
Transcript level	Indica que a existência da proteína não está provada rigorosamente, mas dados de expressão (como a existência de cdna(s), RT-PCR ou Northern blots) indicam a existência de um transcrito.
Inferred by homology	Indica que a existência da proteína é provável devido a uma clara evidência da existência de ortólogos em espécies relacionadas.
Predicted	O termo é usado para registros sem evidência de proteína, transcrito ou homologia.
Uncertain	Indica que a existência da proteína é incerta.

TABELA IV.2.1. Tipos de evidência para a existência de uma proteína.

Referência:

http://www.uniprot.org/

IV.3 NCBI GENE

Características:

- Um dos principais objetivos dos projetos de sequenciamento de genomas é a identificação e caracterização de genes.
- GENE foi implementado no National Center for Biotechnology Information (NCBI)[168] para organizar informação sobre genes, servindo como um importante nó em relação aos dados de mapa genômico, sequência, expressão, estrutura de proteína, função e homologia.
- Cada registro em GENE recebe um identificador único, GeneID, que pode ser rastreado ao longo dos ciclos de revisão.
- Registros em GENE são estáveis para genes conhecidos ou preditos, os quais são definidos pela sequência de nucleotídeo ou posição no mapa. Nem todos os taxa estão representados, e o atual escopo corresponde ao do NCBI.
- De uma forma regular, bancos de dados de organismos modelo e outros grupos contribuintes são checados por novas informações.
  - Se o registro já existe no GENE, nova informação é adicionada e informações desatualizadas são corrigidas.
  - Caso contrário, um novo registro é criado.
- GENE pode ser considerado curado pois muitos dos bancos de dados contribuintes são curados. Porém, nem sempre o banco de dados tenta reconciliar genes definidos por diferentes *pipelines* de anotação que podem diferir em níveis de regras e revisão curatorial.
- GENE serve com um ponto central de informações para bancos de dados internos e externos ao NCBI.
  - Registros são processados gene-a-gene ou como parte da submissão de um genoma anotado ou cromossoma;
  - Identificadores de Genes e nomes associados, e acesso de sequências, provêm uma estrutura comum de referência para vários bancos de dados.
  - Para alguns genomas (e.g. humano, camundongo, rato, galinha, cachorro), os registros em GENE são atualizados continuamente.

[168] http://www.ncbi.nlm.nih.gov/

- Para outros genomas, atualizações dependem de re-submissão da anotação da sequência genômica por um grupo externo.

– GENE inclui registros para genes confirmados e genes preditos por processos de anotação. A evidência para um gene pode ser inferida a partir do *status* do RefSeq[169] que o define.
  - Por exemplo, RefSeqs definidos como predito ou modelo possuem menos suporte de evidência do que aqueles das categorias validado, provisório ou revisado.

– GENE não pretende ser completo. Ele serve como um guia para informações adicionais em outros bancos de dados.
  - Por exemplo, um gene pode ser representado por múltiplas sequências, mas nem todas são reportadas explicitamente a partir do GENE. Em vez disso, conexões são fornecidas do GENE para o Entrez Protein, Nucleotide e Blink (NCBI), onde mais sequências com similaridade significativa podem ser recuperadas.

– Em adição aos múltiplos *links* dos bancos de dados do NCBI, *LinkOuts* de bancos de dados externos submetidos ao GENE suportam uma imediata navegação a mais informações específicas de genes.

– As funções centrais do GENE são estabelecer identificadores únicos para genes que podem ser rastreados e, através disso, dar suporte a conexões acuradas definindo sequências, nomenclatura e outros descritores. Com essa infraestrutura é possível:
  - Dar suporte ao *pipeline* de anotação do NCBI com base no posicionamento de sequências com GeneID conhecido.
  - Fornecer uma estrutura de referência de genes e todos os seus atributos independente de espécie.
  - Dar suporte a identificação de genes representados por sequências em bases de dados públicas externas.

– Muito do poder de consulta do GENE vem de explorar suas conexões com outros bancos de dados.

– Informações de sequências (*accessions* e *links*) são distribuídas através do registro de GENE.
  - Por exemplo, os diagramas de transcritos e produtos são providos quando um gene foi anotado numa sequência genômica RefSeq, isto é, quando a informação de intron/exon/coding region está disponível nas coordenadas genômicas.

[169] http://www.ncbi.nlm.nih.gov/refseq/

- Cada posição num produto gênico, quando representado por um RefSeq RNA e/ou Proteína, é fornecido relativo à sua sequência genômica de DNA.
- Cada RefSeq ACCESSION NUMBER (genômica, mRNA e proteína) ancora um *link* para diferentes formatos de sequência no Entrez Nucleotídeo ou Entrez Proteína.
- O *link* do ACCESSION NUMBER para a sequência genômica apresenta apenas a região específica do gene.

– A base GENE utiliza várias abordagens para descrever a função de um gene e seus produtos codificados, incluindo:
  - Declarações descritivas explícitas (RefSeq *Summary* e GeneRIF);
  - Nomes de genes, produtos e vias biológicas;
  - Ontologias associadas (GO);
  - Relatórios de interações;
  - Números EC (Enzyme Commission);
  - Inferências a partir do conteúdo de domínios;
  - Descrição de doenças e fenótipos específicos de alelos;
  - *Links* para outros bancos de dados (OMIM, HomoloGene, PubMed etc).
  - Muitas destas categorias incluem *links* para informações adicionais em outros bancos de dados.

Referências:

http://www.ncbi.nlm.nih.gov/gene/

Gene Help - Bethesda (MD): National Center for Biotechnology Information (US); 2005. http://www.ncbi.nlm.nih.gov/books/NBK3839/

Maglott D, Pruitt K, Tatusova T. Gene: A Directory of Genes. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002. Chapter 19. 2005 Mar 3 [Updated 2011 Dec 12]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21085/.

IV.4 PFAM

Características:

- O banco de dados PFAM é uma ampla coleção de famílias de proteína, cada uma representada por alinhamentos múltiplos de sequências e modelos escondidos de Markov (HMM).
- Para cada família na base PFAM é possível analisar múltiplos alinhamentos, avaliar arquiteturas de domínios, examinar a distribuição entre espécies, seguir *links* para outros bancos de dados e visualizar estruturas de proteínas conhecidas.
- São dois os componentes do PFAM: Pfam-A e Pfam-B.
  - Os registros do Pfam-A são de famílias com alta qualidade e curadas manualmente. Apesar do Pfam-A cobrir uma grande proporção de sequências dos bancos de dados de sequências, para oferecer uma cobertura mais ampla de proteínas conhecidas é gerado um suplemento usando o banco de dados ADDA[170] [Heger *et al.*, 2005]. Essas entradas geradas automaticamente são chamadas Pfam-B.
  - Famílias Pfam-B não são anotadas e possuem baixa qualidade pois são geradas automaticamente a partir de clusters não redundantes da versão mais recente do banco de dados ADDA.
  - Apesar da qualidade mais baixa, as famílias do Pfam-B podem ser úteis para identificar regiões conservadas quando não existem registros no Pfam-A.
- PFAM também gera agrupamentos de famílias relacionadas de alto nível, conhecidos como clãs. Um clã é uma coleção de registros Pfam-A que estão relacionados por similaridade de sequências, estrutura ou perfis HMM.
- Os registros PFAM são classificados em uma das quatro opções:
  - *Family*: coleção de regiões de proteína relacionadas.
  - *Domain*: unidade estrutural.
  - *Repeat*: unidade curta que é instável isoladamente mas forma uma estrutura estável quando múltiplas cópias estão presentes.
  - *Motif*: unidade curta encontrada externamente aos domínios globulares.

[170] ADDA é um programa baseado em alinhamentos "todos-contra-todos" para demarcar domínios em sequências de proteína. http://wwwfgu.anat.ox.ac.uk/~andreas/adda/index.html

Diversidade de Proteínas:

- Proteínas são geralmente compostas por um ou mais regiões funcionais, denominados domínios. Diferentes combinações de domínios geralmente dão origem à diversidade de proteínas encontradas na natureza.
- A identificação de domínios em proteínas pode fornecer indícios sobre suas funções.
- Uma dada família Pfam pode ocorrer muitas vezes numa única sequência de proteína, se a família/domínio for uma unidade repetitiva, por exemplo, ou quando um perfil HMM coincide com pequenos segmentos da sequência, mas várias vezes.
- Mais de 79,9% das proteínas do SWISSPROT e TrEMBL (06/2012) têm pelo menos uma correspondência com uma família Pfam-A[171].
- Cada Pfam HMM representa uma família de proteína ou domínio. Fazendo uma busca de uma sequência de proteína contra a biblioteca HMM do PFAM, pode-se determinar quais domínios ela possui, i.e. sua arquitetura de domínios.
- PFAM também pode ser usado para analisar proteomas e questões de arquiteturas complexas de domínios.
- Um HMM é um modelo probabilístico.
  - No PFAM usa-se HMMs para transformar a informação contida num alinhamento múltiplo de sequências num sistema de pontuação específico de posição.
  - Pode-se fazer buscas de HMMs contra o banco de dados de proteína UniProt para encontrar sequências homólogas.

Referências:

http://pfam.sanger.ac.uk/
http://pfam.sanger.ac.uk/help

[171] Pfam 27.0, Março 2013, 14831 famílias

IV.5 GENE ONTOLOGY (GO)

Características:

- O projeto Gene Ontology é um esforço colaborativo que surgiu da necessidade de se elaborar descrições de produtos gênicos que sejam consistentes em diferentes bases de dados.
- O projeto GO desenvolveu três vocabulários controlados (ontologias) que descrevem produtos gênicos (independente do organismo) em termos de sua associação com processos biológicos, componentes celulares e funções moleculares. São três aspectos distintos:
  - O desenvolvimento e manutenção das próprias ontologias
  - A anotação dos produtos gênicos, o que implica fazer associações entre as ontologias e os genes e produtos gênicos nos bancos de dados colaboradores
  - Desenvolvimento de ferramentas que facilitam a criação, manutenção e utilização das ontologias.
- Este vocabulário controlado pode ajudar a análise de fontes de dados diversificadas, facilitar a padronização de anotação, melhorar a elaboração, auxiliar a construção de expressões e processamento de consultas.
- O uso de termos GO por bancos de dados colaboradores facilita a uniformidade de consultas entre eles. Os vocabulários controlados são estruturados de forma que eles podem ser consultados em diferentes níveis, por exemplo, pode-se usar GO para encontrar todos os produtos gênicos no genoma do rato que estão envolvidos em transdução de sinal, ou pode-se focar em todos os receptores tyrosine kinase, por exemplo.
- Esta estrutura também permite que anotadores possam atribuir propriedades de genes ou seus produtos em diferentes níveis, dependendo da profundidade do conhecimento sobre a entidade.
- A ontologia abrange três domínios:
  - Componente celular: partes de uma célula ou seu ambiente extracelular;
  - Função molecular: atividades elementares de um produto gênico no nível molecular, como ligação ou catálise;
  - Processo Biológico: operações ou conjunto de operações de eventos moleculares com um início e fim definido, pertinente ao funcionamento integrado de unidades vivas: células, tecidos, órgãos e organismos.

- Por exemplo, o produto gênico cytochrome c pode ser descrito:
  - Função molecular: pelo termo oxidoreductase activity;
  - Processo biológico: pelos termos oxidative phosphorylation and induction of cell death;
  - Componente cellular: pelos termos mitochondrial matrix and mitochondrial inner membrane
- A ontologia GO é estruturada como um grafo acíclico dirigido e cada termo tem relacionamentos definidos com um ou mais termos no mesmo domínio, e às vezes em outros domínios:
  - Uma sequência de proteína pode ser anotada com zero ou mais nós e em qualquer nível dentro de cada ontologia. A anotação em uma ontologia é independente de sua anotação nas outras ontologias.
- GO não é um banco de dados de sequências gênicas, nem um catálogo de produtos gênicos. Ela descreve como os produtos gênicos se comportam num contexto celular.
- GO não é uma norma imposta, obrigando sua utilização entre bancos de dados. Grupos participam por interesse próprio, e cooperam para atingir um consenso.
- GO é um caminho para unificar bancos de dados biológicos (i.e., GO não é uma "solução federada"). O compartilhamento de um vocabulário é um passo para a unificação, mas não é, por si só, suficiente.
- A anotação é a prática de capturar as atividade e localização de um produto gênico com termos GO, oferecendo referências e indicando que tipos de evidência estão disponíveis para dar suporte às anotações.
- A existência de várias ontologias permite que sejam criados '*cross-products*' que maximizam a utilidade de cada ontologia enquanto evitam redundância.
  - Por exemplo, combinando termos de desenvolvimento em processos GO com uma segunda ontologia que descreve estruturas anatômicas de *Drosophila*, pode ser criada uma ontologia do desenvolvimento de moscas. Pode-se repetir esse processo para outros organismos sem ter que sobrecarregar GO com um grande número de termos específicos de espécies. De forma análoga, pode-se criar uma ontologia de vias de biossíntese combinando termos de biossíntese na ontologia de processos GO com uma ontologia de química.

Referência: http://www.geneontology.org/

IV.6. KEGG – *Kyoto Encyclopedia of Genes and Genomes*

Características:

- É um sistema de bancos de dados para a compreensão de funções e sistemas biológicos em alto nível.
- É atualmente uma proeminente base de conhecimento de referência para a integração e interpretação de conjuntos de dados moleculares de larga escala gerados pelo sequenciamento de genomas e outras tecnologias experimentais *high-throughput*.
- É uma representação computacional de um sistema biológico, que consiste de blocos construtores de genes e proteínas (informação genômica) e substâncias químicas (informação química), integradas com o conhecimento de diagramas de ligação de interações, reações e redes de relações (informação de sistemas).
- Consiste de 16 bancos de dados principais (TABELA IV.6.1) que podem ser categorizados, de forma ampla, em informação de sistemas, genômica e química.
- Estes bancos de dados contêm vários objetos de dados para a representação computacional de sistemas biológicos. Desta forma, um registro do KEGG é chamado de objeto KEEG para cada banco de dados.
- Estatística em 29/08/2013 (TABELA IV.6.2)

Para o projeto conceitual desta tese, dois módulos foram considerados inicialmente: KEGG PATHWAY e, dentro do KEGG LIGAND[172], a base ENZYME.

- O KEGG PATHWAY é uma coleção de mapas manualmente traçados de forma a representar o conhecimento das redes de reações e interações moleculares para metabolismo, processamento de informação genética, processamento de informações ambientais, processos celulares e doenças humanas.
- O KEGG LIGAND contém o conhecimento do universo de substâncias químicas e reações relevantes para a vida.
- O esquema, atualmente, faz referência cruzada apenas com a base de dados ENZYME, derivada da Nomenclatura de Enzimas[173] IUPAC-IUBMB.

[172] Consiste das bases de dados: COMPOUND, DRUG, GLYCAN, REACTION, RPAIR e ENZYME

[173] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they catalyze (http://www.chem.qmul.ac.uk/iubmb/enzyme/).

Category	Database	Content
Systems information	KEGG PATHWAY	KEGG pathway maps
	KEGG BRITE	BRITE functional hierarchies
	KEGG MODULE	KEGG modules of functional units
	KEGG DISEASE	Human diseases
	KEGG DRUG	Drugs
	KEGG ENVIRON	Crude drugs and health-related substances
Genomic information	KEGG ORTHOLOGY	KEGG Orthology (KO) groups
	KEGG GENOME	KEGG organisms with complete genomes
	KEGG GENES	Gene catalogs in complete genomes
	KEGG SSDB	Sequence similarity database for GENES
Chemical information	KEGG COMPOUND	Metabolites and other small molecules
	KEGG GLYCAN	Glycans
	KEGG REACTION	Biochemical reactions
	KEGG RPAIR	Reactant pair chemical transformations
	KEGG RCLASS	Reaction class defined by RPAIR
	KEGG **ENZYME**	Enzyme nomenclature

TABELA IV.6.1. Lista dos 16 bancos de dados principais.

KEGG PATHWAY	Pathway maps, reference (total)	448 (262,304)
KEGG BRITE	Functional hierarchies, reference (total)	147 (88,012)
KEGG MODULE	KEGG modules, reference (total)	582 (195,843)
KEGG DISEASE	Human diseases	1,301
KEGG DRUG	Drugs	10,018
KEGG ENVIRON	Crude drugs and health-related substances	845
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	17,046
KEGG GENOME	KEGG Organisms	2,822
KEGG GENES	Genes in high-quality genomes (192 eukaryotes, 2452 bacteria, 160 archaea)	11,228,989
KEGG SSDB	Best hit relations within GENES Bi-directional best hit relations within GENES	144,053,385,301 3,323,814,251
KEGG DGENES	Genes in draft genomes (18 eukaryotes)	432,488
KEGG EGENES	Genes as EST contigs (99 eukaryotes)	3,792,883
KEGG MGENES	Genes in metagenomes (716 samples)	90,754,418
KEGG COMPOUND	Metabolites and other small molecules	17,084
KEGG GLYCAN	Glycans	10,985
KEGG REACTION	Biochemical reactions	9,398
KEGG RPAIR	Reactant pair chemical transformations	14,218
KEGG RCLASS	Reaction class	2,831
KEGG ENZYME	Enzyme nomenclature	6,043

TABELA IV.6.2. Estatística em 29/08/2013.

Referência:

http://www.genome.jp/kegg/