MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

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

MODELAGEM ESTATÍSTICA DO FENÔMENO DE TROCA HIDROGÊNIO/DEUTÉRIO EM PROTEÍNAS ATRAVÉS DE PROPRIEDADES ESTRUTURAIS E DINÂMICAS

LUCAS DE ALMEIDA MACHADO

Rio de Janeiro

Julho de 2016

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

Lucas de Almeida Machado

MODELAGEM ESTATÍSTICA DO FENÔMENO DE TROCA HIDROGÊNIO/DEUTÉRIO EM PROTEÍNAS ATRAVÉS DE PROPRIEDADES ESTRUTURAIS E DINÂMICAS

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

Orientador: Prof. Dr. Paulo Ricardo Batista

RIO DE JANEIRO

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

M149 Machado, Lucas de Almeida

Modelagem estatística do fenômeno de troca hidrogênio/deutério em proteínas através de propriedades estruturais e dinâmicas / Lucas de Almeida Machado. – Rio de Janeiro, 2016.
xii, 51 f. : il. ; 30 cm.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2016.
Bibliografia: f. 76-85

1. Modelagem estatística. 2. Troca hidrogênio/deutério. 3. Estrutura e dinâmica de proteínas. 4. Análise de modos normais. I. Título.

CDD 572.65

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: LUCAS DE ALMEIDA MACHADO

MODELAGEM ESTATÍSTICA DO FENÔMENO DE TROCA HIDROGÊNIO/DEUTÉRIO EM PROTEÍNAS ATRAVÉS DE PROPRIEDADES ESTRUTURAIS E DINÂMICAS

ORIENTADOR: Prof. Dr. Paulo Ricardo Batista

Aprovada em: 29 / 03 / 2016

EXAMINADORES:

Prof. Dr. Mauricio Garcia de Souza Costa – **Presidente** (Fiocruz)
Prof. Dr. Fabio Ceneviva Lacerda Almeida (UFRJ)
Prof. Dr. Marcelo Ribeiro Alves (Fiocruz)
Prof. Dra. Viviane Silva de Paula (UFRJ)
Prof. Dr. Francisco Gomes Neto (Fiocruz)

Rio de Janeiro, 29 de março de 2016

AGRADECIMENTOS

Gostaria de agradecer a todos os companheiros do Programa de Computação Científica e do Programa de Pós-graduação em Biologia Computacional e Sistemas que colaboraram direta ou indiretamente com este trabalho. Os nomes são muitos, por isso não vou me arriscar a listar, mas é certo que sem a colaboração de cada uma dessas pessoas, esse trabalho teria sido uma tarefa bem mais difícil. Dentre esses, gostaria de agradecer especialmente ao meu orientador, Paulo Ricardo Batista, por todos os ensinamentos ao longo desses dois anos.

Agradeço ao meu pai, Irismar Machado, por despertar a curiosidade científica em mim e em meu irmão, por estimular atividades criativas desde os primeiros anos e por todo o estímulo e apoio desde sempre, sem isso eu não teria optado pelo caminho que sigo hoje. À minha mãe, Lucinda Almeida, por todo o suporte, pela compreensão em cada momento difícil e por cada empurrão quando eu me via em um dilema importante, sem isso eu não teria conseguido trilhar o caminho. Ao meu irmão, Davi Machado, pela companhia e pelas conversas produtivas sobre as diversas curiosidades do mundo.

Agradeço à Aline Oliveira pelo exemplo como pesquisadora, assim como seu marido Marcos da Costa Alves pelo suporte, pelo incentivo e por todos os momentos divertidos até aqui, sem essas duas pessoas eu não teria chegado até aqui. À minha tia Lucy Almeida e seu marido Marcelo Esteves pelo apoio ao longo de cada etapa registrada nessa dissertação, e ao meu padrinho André Almeida pelo exemplo de determinação.

Acima de tudo, agradeço a todos os pesquisadores que vieram antes de mim e compartilharam suas obras com a comunidade científica, tornando esse trabalho possível.

“O verdadeiro prazer está em descobrir, não em saber”

- Isaac Asimov

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

MODELAGEM ESTATÍSTICA DO FENÔMENO DE TROCA HIDROGÊNIO/DEUTÉRIO EM PROTEÍNAS ATRAVÉS DE PROPRIEDADES ESTRUTURAIS E DINÂMICAS

Lucas de Almeida Machado

RESUMO

O estudo da estrutura e da dinâmica de proteínas é de suma importância para a compreensão dos mecanismos funcionais das mesmas. Dentre os métodos experimentais disponíveis para realizar esse tipo de estudo, está a utilização da troca hidrogênio/deutério (HX). Este método consiste em expor a proteína à água deuterada e analisar através de ressonância magnética nuclear (NMR) ou espectrometria de massa (MS) quais dos hidrogênios amídicos foram trocados por deutérios do solvente, permitindo assim, inferir grau de exposição ao solvente, presença de ligações hidrogênio e flexibilidade da proteína. Diversos modelos foram criados nos últimos anos afim de explicar e predizer dados de HX, porém, nenhum deles foi capaz de explicar completamente o fenômeno. No presente trabalho foram construídos modelos estatísticos para explicar dados de troca obtidos por MS, utilizando parâmetros estruturais (número de contatos e ligações hidrogênio) e parâmetros que descrevem a dinâmica: como fatores B, flutuações obtidas por análise de modos normais (NMA) e por modelos de redes elásticas (ENM). Empregando parâmetros estruturais, dinâmicos e informações acerca das condições experimentais, também foram construídos modelos preditivos lineares e baseados em *machine learning* para dados de troca obtidos por NMR. Observamos que a adição das variáveis dinâmicas aos modelos que utilizam apenas parâmetros estruturais aumenta as correlações entre os valores ajustados e os dados experimentais obtidos por MS. Além disso, o modelo preditivo baseado em *machine learning* construído para a predição de dados de HX obtidos por se mostrou eficaz na predição dos dados de diversas proteínas. Os resultados aqui mostrados realçam a influência dos movimentos de grande amplitude sobre os dados de HX, e a importância da dinâmica na modelagem desse tipo de dado, assim como a utilização de informações acerca das condições experimentais.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

STATISTICAL MODELING OF HYDROGEN/DEUTERIUM EXCHANGE IN PROTEINS THROUGH DYNAMICAL AND STRUCTURAL PROPERTIES

LUCAS DE ALMEIDA MACHADO

Abstract

The study of protein structure and dynamics is an important step to understand its functional mechanisms. Hydrogen/deuterium exchange (HX) is one of the methods available for this kind of investigation. This method consist of exposing the protein to heavy water and analyzing through mass spectrometry (MS) or nuclear magnetic resonance (NMR) which amidic hydrogens exchanged with water's deuterons, thus allowing to infer solvent exposure, presence of hydrogen bonds and protein flexibility. In the last years, several models were built in order to explain and predict HX data. However, none of them was able of explaining the data. In the present work, we built statistical models to explain HX data probed through MS, using structural parameters (number of contacts and hydrogen bonds) and dynamical parameters, such as B-factors, fluctuations obtained through normal mode analysis (NMA) and through elastic network model (ENM). Using information of experimental conditions in conjunction with structural and dynamical parameters, we built machine learning based models and linear models to predict HX data obtained through NMR. Here we observed that the inclusion of dynamical parameters in models built purely with structural parameters enhances the correlations between experimental data and the fitted values. Besides that, machine learning based predictive models for HX data obtained through NMR was efficient in predicting data of several proteins. The results shown here highlight the influence of large amplitude motions in the HX data and the importance of dynamics when modeling this kind of data, as well as the use of experimental condition information.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

%D – porcentagem de deuteração

AIC – *Akaike information criteria*

CA – carbono-α

Cryo-EM – crio-eletromicroscopia

ENM – modelos de Redes Elásticas

GNM – modelo de redes Gaussianas

HX – troca hidrogênio/deutério

K_{ch} – constante de troca

K_{cl} – constante de fechamento

K_{int} – constante Intrínseca

K_{op} – constante de abertura

MD – dinâmica molecular

MS – espectrometria de massa

NMA – análise de modos normais

SASA – área da superfície acessível ao solvente

NMR – ressonância magnética nuclear

PDB – *Protein Data Bank*

PF – fator de proteção

RF – *Random forest*

RMSE – *root mean square error*

RMSF – *root mean square fluctuation*

SNase – nuclease estafilocócica

CAPÍTULO I – INTRODUÇÃO

1. ESTRUTURA E DINÂMICA DE PROTEÍNAS

A compreensão da estrutura e dinâmica de proteínas é um dos grandes desafios da biologia moderna, visto que diversas funções em organismos vivos dependem tanto da estrutura quanto do comportamento dinâmico dessas moléculas. Nos últimos anos diversos avanços nessa área foram realizados graças à utilização de métodos como cristalografia por difração de raios-X (1) e a ressonância magnética nuclear (NMR) (2, 3).

Proteínas são polímeros cujas subunidades básicas são resíduos de α-aminoácidos ligados através de ligações peptídicas (4). Estas subunidades são moléculas compostas por um grupamento amina, um grupamento carboxila, um hidrogênio e uma cadeia lateral ligados a um carbono (por convenção chamado de carbono α, CA). A diversidade desses blocos de construção (que em eucariotos apresentam 20 diferentes cadeias laterais) faz com que seja grande o número de possíveis combinações de aminoácidos em uma proteína. Durante a síntese proteica, os aminoácidos são ligados covalentemente em um arranjo linear, onde o grupamento amina de um aminoácido reage com o grupamento carboxila do aminoácido seguinte, resultando na formação da ligação peptídica e na liberação de uma molécula de H_2O. Uma vez que os grupamentos amina e carboxila são perdidos na formação da ligação entre os α-aminoácidos, essas subunidades do polímero passam a ser chamadas de resíduos de aminoácidos.

Em proteínas existe uma hierarquia quanto à classificação estrutural (Figura 1). A estrutura primária é o arranjo linear dos resíduos, ou seja, a sequência de resíduos ordenada do N ao C-terminal (5, 6). Porém, existem outros níveis estruturais. Logo após o início da tradução, os resíduos recém-sintetizados interagem com outros formando estruturas locais, que são estabilizadas por ligações hidrogênio no esqueleto peptídico, tendo o hidrogênio amídico como doador, e o oxigênio da carbonila como aceptor. A disposição

regular destas ligações hidrogênio pode originar padrões estruturais, como as alfa-hélices e as folhas beta, que são classificados como estruturas secundárias (1, 6, 7).

A estrutura terciária, por sua vez, é o arranjo espacial das estruturas secundárias de uma cadeia polipeptídica, que pode ser estabilizado por interações intramoleculares fracas (ex. ligações hidrogênio, pontes salinas, etc); e/ou por ligações covalentes, no caso das pontes dissulfeto (que ligam cadeias laterais de resíduos de cisteína). Alguns arranjos de estruturas terciárias são encontrados frequentemente em proteínas e são chamados de domínios estruturais, e estão relacionados com funções específicas em proteínas (8).

Algumas proteínas são monoméricas, possuindo apenas uma cadeia polipeptídica. No entanto, diversas proteínas em sua forma madura são formadas por duas ou mais cadeias. Ao arranjo de mais de uma cadeia polipeptídica de uma proteína damos o nome de estrutura quaternária (5, 6).

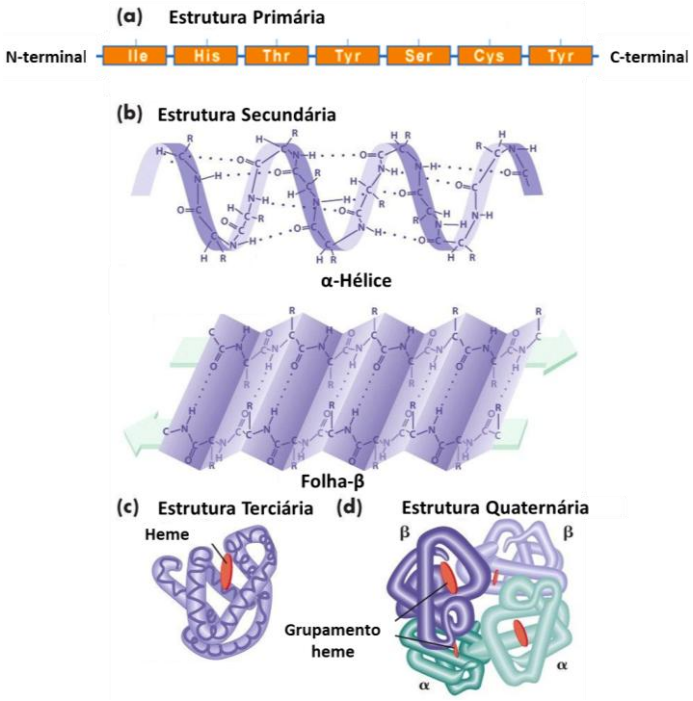


Figura 1 - Representação dos níveis estruturais das proteínas.
a) Estrutura primária – sequência de resíduos de aminoácido; b) Estruturas secundárias (α-hélices e folhas β) – as ligações hidrogênio estão representadas nas estruturas por linhas pontilhadas. C) Representação da estrutura terciária de uma das cadeias da Hemoglobina, sendo representada a presença do grupamento prostético Heme; e C) arranjo espacial das quatro cadeias do tetrâmero da hemoglobina, a estrutura quaternária. Adaptado de (9).

As principais evidências sobre estruturas de proteínas surgiram nos anos 50: *i.* com as estruturas secundárias postuladas por Pauling, Corey e Brandson (7); *ii.* com os experimentos de Linderstrom-Lang, que visaram, em um primeiro momento, verificar experimentalmente os padrões de ligação hidrogênio em proteínas; e, *iii.* finalmente o trabalho de John Kendrew, que resolveu a primeira estrutura cristalográfica de uma proteína, a mioglobina (1).

Com o avanço das técnicas de biologia molecular e estrutural, atualmente são conhecidas sequências de proteínas de vários organismos, e em uma menor escala suas estruturas. Com base nesses dados foi possível estabelecer relações entre sequência, estrutura e função em proteínas (10), tornando claro o fato de que a estrutura é mais conservada que a sequência (8).

No entanto, proteínas não são entidades estáticas, possuindo diversos graus de liberdade conformacional (3N, onde N é o número de átomos). Sendo assim, faz-se importante o estudo não apenas da estrutura, mas também da dinâmica para o entendimento da função das proteínas (11).

1.1. Métodos experimentais para determinação da estrutura de proteínas

Desde a elucidação da estrutura da mioglobina houve um crescimento exponencial da aplicação de métodos experimentais para o estudo de estruturas de proteínas. Atualmente o banco de dados de estruturas de proteínas (*Protein Data Bank* - PDB) contém mais de 100.000 estruturas depositadas, sendo a grande maioria determinada por cristalografia por difração de raios-X. O segundo método mais utilizado é a ressonância magnética nuclear (NMR), seguida pela microscopia eletrônica e métodos híbridos (12).

O estudo da estrutura de proteínas pode ser o ponto de partida para o entendimento dos mecanismos moleculares pelos quais estas desempenham suas funções (13, 14). Nesta dissertação serão discutidas algumas das metodologias mais utilizadas para o estudo de estrutura e dinâmica de proteínas.

1.1.1. Cristalografia por difração de Raios-X

Em 1912, o primeiro padrão de difração de raios X foi obtido utilizando como alvo um cristal de sulfato de cobre (15). Mas foi nos anos 50 que a difração de raios-X teve um importante papel na elucidação da estrutura do DNA e de proteínas, tornando-se uma das principais técnicas para a determinação da estrutura de biomoléculas (1, 16). Resumidamente, este método consiste em incidir um feixe de raios-X através de um cristal da molécula alvo. Este feixe interage com os átomos, ocorrendo o fenômeno de difração (flexão das ondas ao redor de um obstáculo). Devido à simetria do cristal, à partir do padrão de intensidade dos raios difratados, aplicando-se a lei de Bragg, pode-se então obter as densidades eletrônicas dos átomos do sistema. A partir destas densidades é possível ajustar computacionalmente as posições dos átomos de cada resíduo, construindo modelos estruturais de forma a satisfazer os dados experimentais (17, 18). As principais etapas deste processo estão descritas na Figura 2.

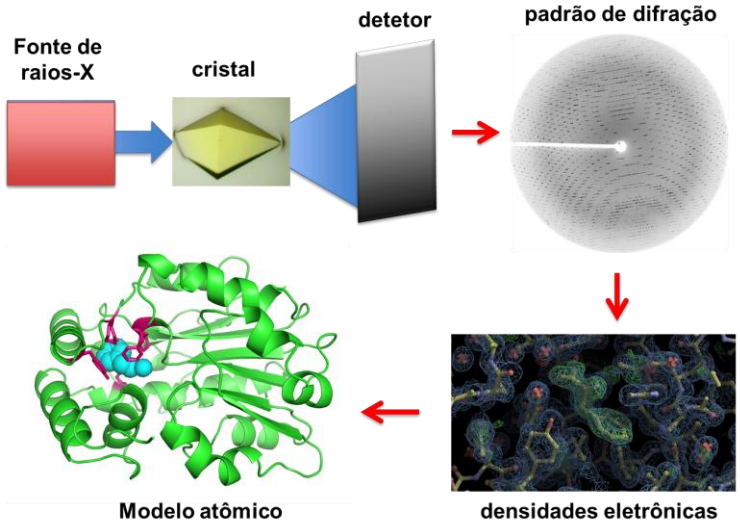


Figura 2 - Principais etapas para a determinação da estrutura de uma proteína por cristalografia por difração de raios X
O esquema representa a obtenção dos padrões de difração de raios X no cristal, a partir dos quais se obtém as densidades eletrônicas dos átomos do sistema estudado. De posse das densidades eletrônicas, são construídos modelos que satisfaçam as restrições impostas pelas mesmas.

Diversos fatores podem influenciar a qualidade das estruturas obtidas, como a qualidade do cristal, flexibilidade da proteína, etc. Os dados de difração resultam de uma média de todas as conformações dos átomos do cristal ao longo do tempo, e embora os átomos da proteína tenham movimento restrito em

ambiente cristalino, os mesmos não estão estáticos durante o experimento. Assim como um objeto em movimento aparece borrado em uma fotografia, certas regiões da proteína com liberdade conformacional podem não gerar densidades eletrônicas com uma resolução satisfatória (1, 19, 20). Além disso, para grande parte das proteínas, a obtenção dos cristais pode ser uma etapa difícil e custosa, como comentado por Dale *et al*. (21), menos de 20% das proteínas expressas formam cristais propícios para a determinação de estruturas.

A cristalografia pode ser utilizada no estudo de grandes complexos proteicos. Por outro lado, um de seus vieses está nas restrições conformacionais impostas pelo ambiente cristalino. Em cristais, existe um arranjo periódico de várias unidades da mesma proteína (ou complexo proteico). Sendo assim, as unidades interagem entre si formando os chamados contatos cristalográficos. Diversas destas interações não são observáveis em condições fisiológicas. As condições de pH, temperatura e a presença de agentes estabilizantes também podem gerar artefatos, favorecendo uma conformação não condizente com a estrutura da molécula em solução (22-24).

3.2.1. Fatores B

O fator B ou fator de Debye–Waller é um valor calculado a partir de dados cristalográficos e está associado à liberdade conformacional dos átomos no cristal (22). Essa medida trata da incerteza quanto à posição de um átomo em relação a sua respectiva densidade eletrônica. Assim, átomos com menores valores de fator B estão em regiões mais ordenadas do cristal, enquanto átomos com maiores valores de fator B estão em regiões mais flexíveis (25). Dessa forma, o fator B vem sendo utilizado na literatura como uma forma aproximada de se representar a flexibilidade de uma proteína (22, 25). Os fatores B utilizados nesta dissertação foram obtidos diretamente das estruturas cristalográficas.

1.1.2. Ressonância Magnética Nuclear (NMR)

Embora a cristalografia seja o método mais utilizado para a determinação de estruturas, a ressonância magnética nuclear (NMR – *nuclear magnetic resonance*) é uma poderosa abordagem para estudos tanto de estrutura quanto de dinâmica de proteínas. Uma das grandes diferenças entre a NMR e a cristalografia no que diz respeito às amostras é o fato de a primeira poder ser

realizada em solução, onde a proteína purificada se encontra livre de restrições espaciais e em contato com o solvente.

A NMR trata de um fenômeno em que os núcleos atômicos expostos a um campo magnético absorvem e reemitem radiação. A NMR foi pela primeira vez demonstrada por Isidor Rabi em 1938 (26-28), e posteriormente aplicada por Richard R. Ernst e Kurt Wüthrich (3, 29-31) ao estudo de proteínas, culminando na determinação da estrutura completa de uma proteína globular utilizando este método pela primeira vez entre 1982-5 (2).

O cerne da metodologia reside no fato de que idealmente, cada núcleo atômico está inserido em um ambiente químico diferente, e consequentemente, as influências das diferentes vizinhanças geram comportamentos distintos frente a um campo magnético. No entanto, para grandes polímeros (como proteínas) há diversas sobreposições de sinais em espectros unidimensionais, o que leva à necessidade da utilização de espectros multidimensionais.

Experimentos como COSY (*correlation spectroscopy*), TOCSY (*total correlation spectroscopy*) e NOESY (*Nuclear overhauser effect spectrocopy*) são formas amplamente utilizadas de espectros bidimensionais que são obtidos por sinais gerados por um tipo de núcleo (usualmente ^{1}H), sendo chamadas de metodologias homonucleares. COSY e TOCSY são métodos baseados na transferência de magnetização através das ligações químicas de prótons adjacentes, sendo possível observar quais núcleos estão acoplados, e desta maneira determinar sua proximidade na cadeia do polímero. O experimento de NOESY por outro lado, é baseado na transferência de magnetização através do espaço, sendo utilizado para estudar o acoplamento entre núcleos que podem estar distantes na cadeia polipeptídica (32).

A análise do espectro de NOESY é capaz de gerar informação sobre distâncias espaciais máximas entre dois núcleos atômicos, essas distâncias podem ser utilizadas como restrições espaciais para a construção de modelos que as satisfaçam. Além das restrições de distância, também podem ser introduzidas restrições dos diedros, uma vez que a geometria dos átomos em relação aos carbonos-α afeta seus valores de deslocamento químico. Existem diversos programas que são utilizados para gerar modelos que satisfaçam as diversas restrições espaciais introduzidas pelos dados experimentais (33, 34),

sendo possível gerar conjuntos de estruturas (*ensembles*). Os modelos gerados devem então ser validados, normalmente utilizando métodos estatísticos, tais como os presentes nos servidores WHATIF (35) e PROCHECK (36).

A principal limitação da NMR está no tamanho das moléculas estudadas, proteínas grandes geram diversas sobreposições de picos, fazendo com que a metodologia seja aplicada ao estudo de proteínas pequenas (37).

Figura 3 - Representação de um *ensemble* de estruturas de uma proteína determinado por NMR
Ensemble de estruturas sobrepostas gerado a partir dos dados de NOE, representando diversos estados possivelmente explorados para a proteína ALG13 [adaptado de (38)].

1.1.3. Crio-eletromicroscopia

Outra alternativa que vem ganhando espaço é a crio-eletromicroscopia (crio-EM), que é baseada na passagem de feixes de elétrons em espécimes congelados a temperaturas muito baixas para a produção de imagens (39) (Figura 4). Tradicionalmente a microscopia eletrônica vem sendo utilizada para o estudo de vírus, tecidos e outras estruturas tratadas com metais pesados (40-42). No entanto, o advento da crio-EM permite ir além da estrutura de tecidos, possibilitando a determinação de estruturas de grandes complexos macromoleculares. A crio-EM é aplicada para o estudo de complexos heterogêneos e grandes demais para serem estudados tanto por difração de raio-X quanto por NMR. Ao contrário da cristalografia, a determinação de estruturas por crio-EM não requer a proteína cristalizada (42). Um campo aparentemente promissor para a crio-EM é a elucidação de estruturas de proteínas de membrana, devido a dificuldade de obter-se cristais destas (42, 43).

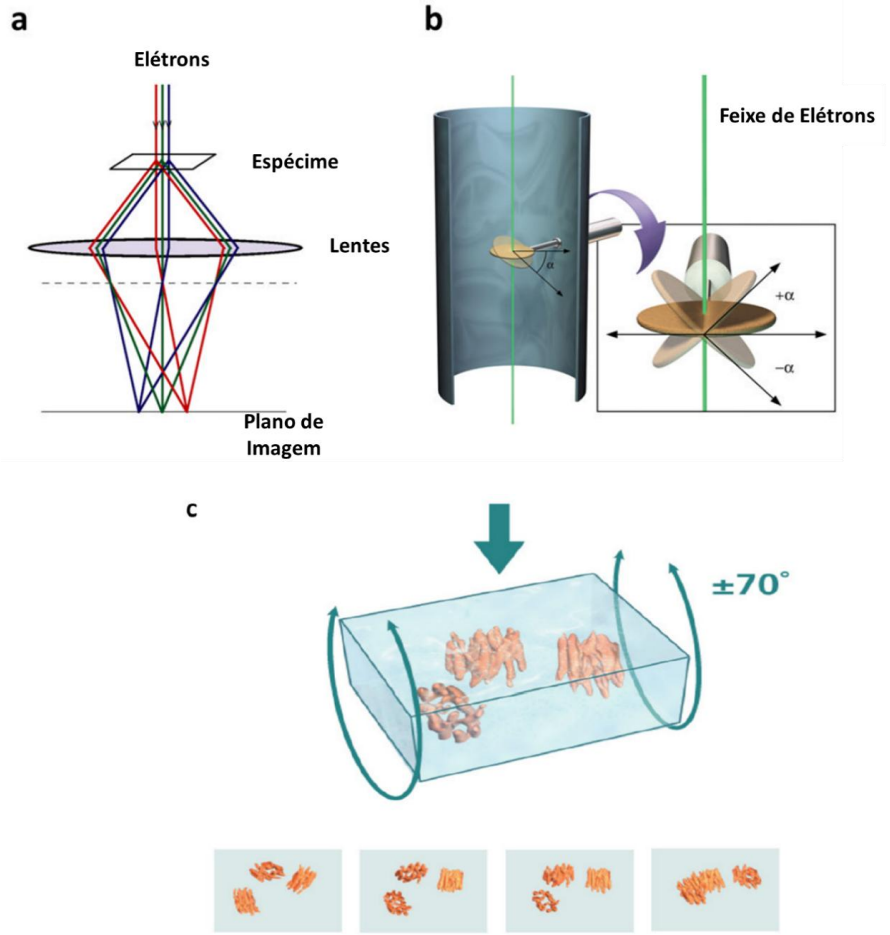


Figura 4 - Formação de uma imagem de crio-EM

Representação esquemática do equipamento e do posicionamento do espécime em relação ao feixe de elétrons e as lentes. b) Esquema da aquisição de dados, que se dá enquanto o espécime é inclinado em relação ao feixe de elétrons. c) as diversas imagens obtidas combinadas computacionalmente para a obtenção das distribuições de densidade dos objetos. Adaptado de (44).

1.2. Métodos computacionais para a predição de estrutura

Dentre os grandes desafios da biologia estrutural computacional, destaca-se a predição da estrutura tridimensional de proteínas. Este desafio começou a ser considerado após a descoberta de Anfinsen em 1961, de que toda a informação necessária para o enovelamento da maioria das proteínas está presente somente na estrutura primária. Essa hipótese foi confirmada com um experimento onde a enzima ribonuclease após ser desnaturada por mercaptoetanol e ureia, era capaz de recompor sua estrutura/atividade quando o agente desnaturante era removido (45). Esse conceito ficou conhecido como

hipótese termodinâmica, ou Dogma de Anfinsen. Existem porém, casos onde proteínas conhecidas como chaperonas são necessárias para conduzir o processo de enovelamento para melhor eficiência (46).

Embora Anfinsen tenha demonstrado que a informação necessária para o enovelamento está presente na estrutura primária, a predição de estruturas através do conhecimento da sequência não é algo trivial. Uma importante questão foi levantada sobre o processo de enovelamento proteico, que ficou conhecida como o Paradoxo de Levinthal. Ele afirmou que uma cadeia polipeptídica com 100 resíduos de aminoácidos levaria mais do que a idade do universo para se enovelar, caso o enovelamento fosse realizado através da exploração aleatória de todas as conformações possíveis (considerando apenas 2 conformações/resíduo/picossegundo, $t= 2^{100}x10^{-9}s$, $\sim10^{16}s$) (47). Contudo, como é sabido, proteínas se enovelam em escalas de tempo muito menores. Sendo assim, o enovelamento só seria possível se fosse um processo dirigido, não aleatório. Partindo da hipótese termodinâmica, surge o modelo de funil de energia de Onuchic e Wolynes (48, 49). Neste modelo, o enovelamento é um processo direcionado e a cadeia polipeptídica é dirigida a explorar conformações cada vez mais termodinamicamente favoráveis.

Embora o modelo do funil restrinja o número de conformações exploradas pelas cadeias polipeptídicas enquanto se enovelam, até hoje só foi possível a simulações atomísticas do enovelamento *in silico* de peptídeos e pequenas proteínas (50, 51) ou através da utilização de métodos simplificados (52, 53). Isto porque o enovelamento proteico ocorre em escalas de tempo dificilmente acessíveis por simulações de dinâmica molecular. Alternativamente, a maioria dos métodos computacionais atuais para a predição de estruturas de proteínas demandam informações de estruturas já conhecidas.

Com o aumento do número de sequências e estruturas determinadas experimentalmente, observou-se que a estrutura de uma proteína é mais conservada que sua sequência (8). Desta maneira, assumindo que em proteínas a estrutura é mais conservada que a sequência, é possível utilizar como moldes (*templates*) estruturas conhecidas para construir modelos da estrutura desejada através da criação de restrições espaciais. Esta é a base da modelagem comparativa (anteriormente conhecida como modelagem por homologia) (54,

55). Outras metodologias, como a modelagem por *threading,* utilizam-se do reconhecimento de padrões de enovelamento (56), aproveitando-se do fato de que existe um número limitado de padrões conhecidos na natureza (57). Algumas pequenas regiões das sequências podem ser modeladas por métodos *ab initio,* ou seja, sem utilizar estruturas de referência, [como revisado em (58)]. No entanto, esta última abordagem possui limitações quanto ao tamanho das sequências para as quais se deseja predizer a estrutura, devido ao grande número de graus de liberdade (59).

1.3. Métodos para o estudo da estrutura e dinâmica de proteínas

O estudo das estruturas das proteínas é uma das partes centrais das pesquisas em biologia estrutural. No entanto, como discutido nesta dissertação, proteínas são entidades dinâmicas e a compreensão de seus movimentos também é fundamental para entender suas funções (11). Inúmeras abordagens experimentais para o estudo da dinâmica de proteínas foram desenvolvidas nas últimas décadas, como NMR, espectroscopia de fluorescência, espectrometria de massa (MS) e outros (60-62). Porém, o constante avanço na capacidade de processamento dos computadores também propiciou o surgimento de métodos computacionais para a exploração dos movimentos de macromoléculas, sendo possível realizar cálculos teóricos partindo de dados estruturais obtidos experimentalmente (63).

1.3.1. Métodos experimentais

1.3.1.1. Ressonância Magnética Nuclear

Uma das formas de se investigar a dinâmica através da NMR, é através do estudo da relaxação. O fenômeno de relaxação consiste na deterioração dos sinais ao longo do tempo, descrevendo como os estados excitados retornam ao equilíbrio após a perturbação (64). Os tempos de relaxação são sensíveis à dinâmica das moléculas, por este motivo é possível estudar movimentos das mesmas - tanto os que ocorrem em escalas de tempo pequenas (ps a ns), quanto movimentos lentos (que ocorrem em escalas de μs a ms) – através de métodos que exploram esse fenômeno (65). Para investigar as flutuações dos sinais dos

núcleos de ^{15}N ou ^{13}C, utiliza-se a técnica de HSQC (*heteronuclear single quantum coherence*), onde estuda-se a transferência de magnetização de um próton para um núcleo como ^{15}N ou ^{13}C. Porém, como os átomos de nitrogênio na maioria dos resíduos estão localizados apenas no *backbone* da proteína, os movimentos identificados desta maneira não refletem a dinâmica das cadeias laterais, o que pode ser alcançado através do estudo de isótopos como C^{13} e deutério (60, 66).

É possível também analisar a flexibilidade de proteínas através da análise de *ensembles* de estruturas geradas através das restrições espaciais obtidas por experimentos de NOESY (67).

O estudo da dinâmica de proteínas por NMR pode ser feito ainda através da avaliação da troca hidrogênio-deutério (HX). Este tópico será tratado em detalhes nas próximas seções, devido à sua relevância no escopo deste trabalho.

1.3.1.2. Espectrometria de Massa

A espectrometria de massas (MS), em um primeiro momento, foi aplicada principalmente no estudo de pequenas moléculas. Porém na década de 1980, com o surgimento das tecnologias de MALDI (ionização e dessorção a laser assistida por matriz**)** e ESI (ionização por *electrospray*) (68, 69), a MS passou a ganhar espaço nos estudos de proteômica, sendo amplamente aplicada tanto na identificação de proteínas como no estudo da estrutura das mesmas (70, 71).

Independentemente das possíveis variações das aplicações da MS, seu cerne reside na análise da relação massa-carga (m/z) do analito. A Figura 5 apresenta um esquema básico de um espectrômetro de massas. De uma forma geral, todos os espectrômetros possuem: *i.* uma fonte de íons (ionizador) onde as moléculas são ionizadas (sendo por MALDI, ESI, etc); seguida de: *ii.* um (ou mais de um) analisador de massas, que separa os íons por sua relação massa-carga; e finalmente *iii.* o detector, que por sua vez detecta os sinais elétricos a partir da corrente de íons gerada pela chegada dos íons (72).

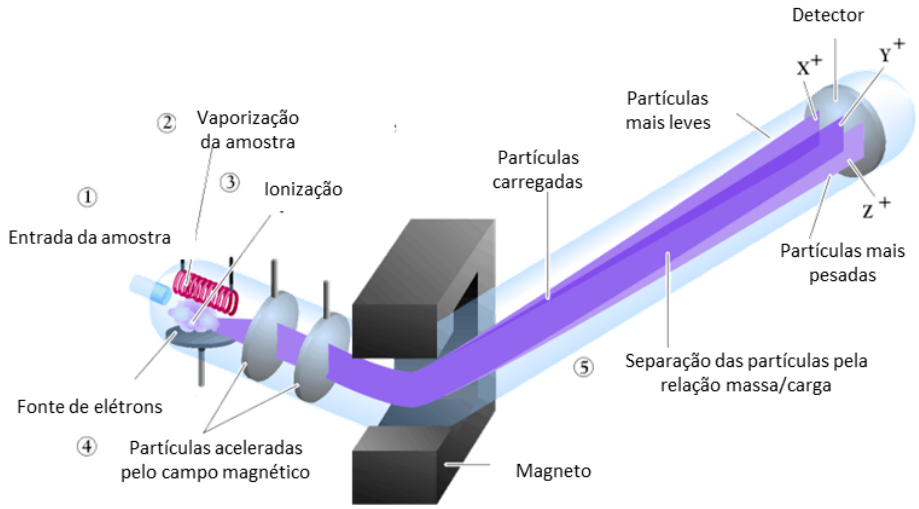


Figura 5 - Representação esquemática de um espectrômetro de massa
Representação esquemática básica de um espectrômetro, desde a introdução de amostras no equipamento, até a separação e detecção das partículas ionizadas. Adaptado de ref. (73).

Quando se trata de proteínas, estas são normalmente analisadas após uma etapa de digestão enzimática (classicamente a tripsina), sendo reduzidas a peptídeos (74). Embora a MS seja amplamente aplicada em estudos de proteômica para a identificação de proteínas, também vem sendo aplicada em estudos relacionados à estrutura de proteínas. Neste contexto, pode ser aplicada de três principais formas: *i.* utilizando agentes que causam *crosslinks* (ligações cruzadas) entre resíduos específicos, para determinar a proximidade espacial desses resíduos em uma proteína ou complexo proteico (75); *ii.* marcação oxidativa induzida por laser e (76) *iii.* o estudo da troca hidrogênio/deutério (HX) (77).

1.3.1.3. Troca Hidrogênio/Deutério

O estudo da troca hidrogênio/deutério (HX) baseia-se em um fenômeno que ocorre naturalmente nas proteínas em água. Trata-se da troca dos hidrogênios da proteína com os hidrogênios da água. O método foi inicialmente aplicado em por Linderstrom-Lang e colaboradores (78, 79) para o estudo das estruturas secundárias propostas por Pauling (7), visto que a troca não ocorre da mesma forma para todos os hidrogênios da proteína. Observou-se, por exemplo, que hidrogênios comprometidos em ligação hidrogênio seriam trocados com o solvente com menos frequência, sendo estes ditos “protegidos”

(80). Nos experimentos de HX, as proteínas de interesse são expostas à água deuterada (D_2O). Uma vez que o deutério é um isótopo mais pesado (possui um próton e um nêutron), enquanto o hidrogênio possui apenas um próton. Desta maneira a troca dos hidrogênios amídicos da cadeia principal pelos deutérios da água deuterada pode ser monitorada através de MS ou NMR. No caso da NMR, o próton e o deutério apresentam diferentes características magnéticas; o deutério não pode ser detectado frente ao mesmo campo magnético que o próton. Sendo assim, perde-se o sinal quando o próton é trocado por deutério, fenômeno que é acompanhado através da utilização do método de HSQC (78).

Com os avanços dos métodos de MS, foi possível utilizar essa metodologia para estudar a HX em proteínas de alto peso molecular (77). Nesse caso, as proteínas são expostas ao D_2O e a reação de troca é realizada; em seguida, a taxa da reação de troca é reduzida pela diminuição do pH para cerca de 2.5 (onde é a troca é mínima). As proteínas são então digeridas por pepsina (capaz de funcionar em baixo pH) e os peptídeos gerados são analisados por MS. Assim, é possível calcular o número de deutérios incorporados em cada peptídeo ao comparar com os resultados de uma proteína não exposta ao D_2O (77, 80) (Figura 6).

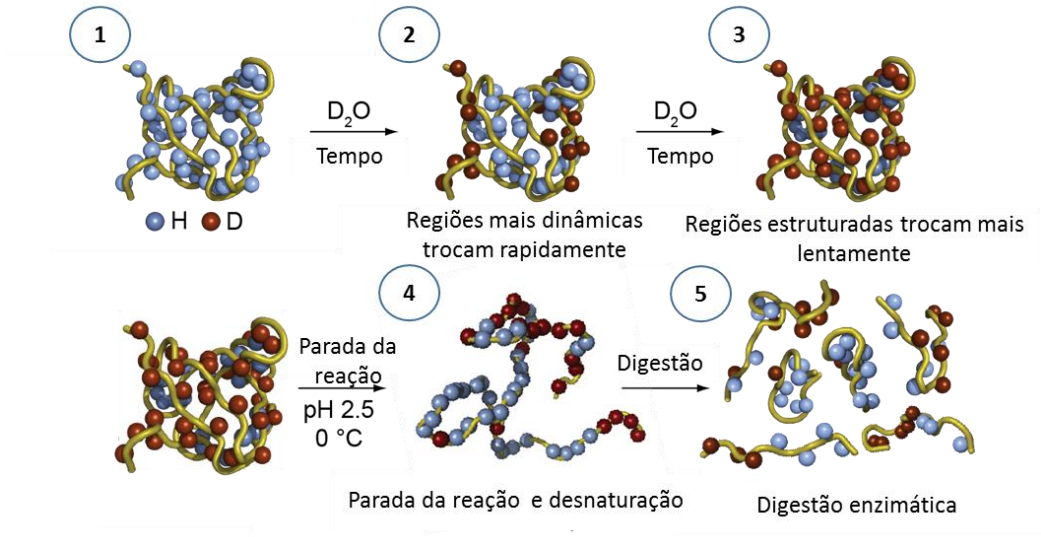


Figura 6 - Esquema simplificado do experimento de MS-HX
Representação esquemática das etapas do experimento de MS-HX. 1) A proteína é exposta à solução de água deuterada. 2) Após algum tempo de exposição as regiões mais expostas têm seus hidrogênios rapidamente trocados por deutérios. 3) As regiões estruturadas da proteína mantem os hidrogênios após algum tempo de reação. 4) Após a exposição, a reação de troca é parada através da diminuição do pH e da temperatura, levando também à desnaturação das proteínas. 5) As proteínas são digeridas por pepsina para a análise espectrométrica.

No caso da utilização de MS, a detecção da deuteração se dá pela comparação da massa de peptídeos deuterados com a massa daqueles que não foram expostos à água pesada. A cobertura dos peptídeos é variável e pode ser observada na Figura 7 (77, 79).

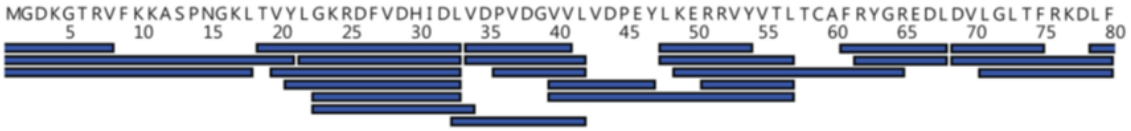


Figura 7 - Representação da cobertura dos peptídeos obtidos em experimentos MS-HX
Representação da cobertura dos peptídeos obtidos por digestão enzimática em um experimento de MS-HX, são mostradas as representações de uma letra para os resíduos de aminoácido da proteína, as posições destes na sequência e os retângulos azuis representam a extensão dos peptídeos obtidos. Os dados da figura correspondem à proteína β-arrestina 1, do resíduo 1 ao 80, adaptado de (81).

Em ambas as abordagens experimentais de HX podem ser estudadas mudanças conformacionais ocasionadas por alterações em condições experimentais como pH, temperatura, mutações e presença ou ausência de ligantes, fazendo com que a técnica tenha um amplo espectro de aplicações na elucidação de diversos fenômenos (82, 83).

Ao longo do desenvolvimento da técnica, Linderstrom-Lang postulou as equações descrevendo o processo de HX, essas equações são utilizadas até os dias de hoje para a interpretação dos resultados experimentais. Assume-se que um dado hidrogênio amídico possui dois possíveis estados, um estado onde possui competência para a troca (estado aberto), e um estado onde não pode trocar (estado fechado). Os dois estados existem em um equilíbrio regido por duas constantes, uma constante k_{op} que descreve o processo de abertura (transição de não competente para competente) e uma constante k_{cl} que descreve o fechamento (transição de competente para não competente) (84).

Além do estudo da dinâmica de proteínas por HX no estado nativo, também podem ser estudados intermediários de *folding* (enovelamento). Dessa forma, assume-se a existência de dois regimes de troca. Um dos regimes, EX1, descreve a troca dos hidrogênios na presença de agentes desnaturantes, e é frequentemente utilizada para o estudo de intermediários de *folding*. O regime

de troca no estado nativo é chamado EX2. Cada regime é caracterizado por suas relações com as constantes de fechamento k_{cl} e a constante que rege a reação de troca (k_{ch}) (84).

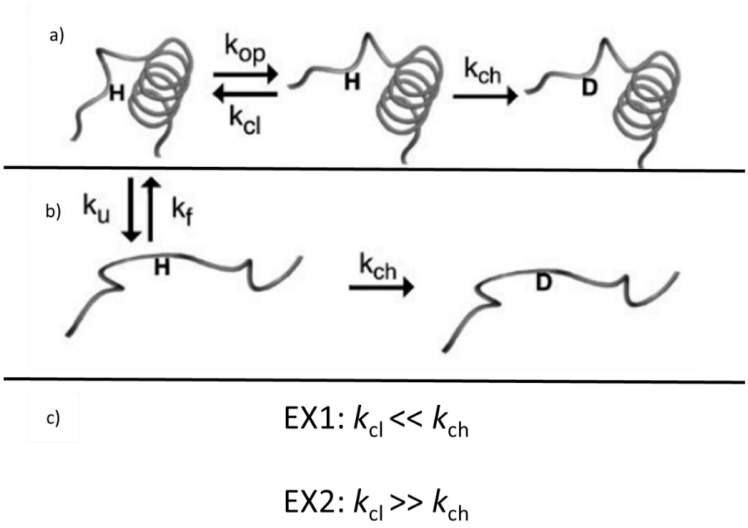


Figura 8 - Representação esquemática dos regimes de troca EX1 e EX2
a) Representação do mecanismo EX2, onde as flutuações estruturais de um domínio da proteína em estado nativo expõem um dado hidrogênio amídico que é posteriormente trocado. b) Representa o mecanismo EX1, K_u e K_f representam constantes de enovelamento e desenovelamento (*folding* e *unfolding*). c) Relações entre as constantes de troca e as constantes de fechamento em cada regime. Adaptado de (84).

No que diz respeito à catálise da reação de troca, uma vez que um hidrogênio está exposto e livre de ligações hidrogênio, a reação pode ocorrer por catálise básica ou catálise ácida. O mecanismo de catálise básica se dá quando o OH^- da solução sequestra o hidrogênio amídico, e posteriormente um átomo de deutério de uma molécula de água deuterada (D_2O) se liga ao nitrogênio amídico (Figura 9). A catálise ácida pode ocorrer por dois diferentes mecanismos: *i.* onde ocorre a protonação do nitrogênio amídico por um átomo de D^+, seguida do sequestro do H^+ pelo solvente; e outro *ii.* onde ocorre uma etapa intermediária de protonação do oxigênio da carbonila, seguida da transferência deste para o nitrogênio amídico. Este último também é chamado de mecanismo do ácido imídico (85). É importante observar que a taxa de reação mínima fica próxima ao pH 2.5 (85).

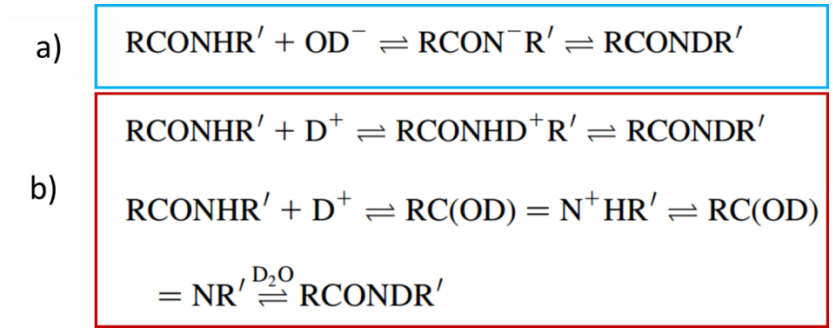


Figura 9 - Equações químicas dos mecanismos de catálise da reação de troca hidrogênio/deutério

a) Mecanismo de catálise básica onde o próton é perdido para o solvendo e o deutério e incorporado b) os dois possíveis mecanismos de catálise ácida, sendo o primeiro a incorporação do D^+ seguida pelo sequestro do hidrogênio pelo solvente, e o segundo o mecanismo do ácido imídico, onde existe uma etapa intermediária em que o D^+ se liga ao oxigênio da carbonila. Adaptado de (85).

Em alguns estudos foram utilizados dipeptídeos sintéticos para determinar as influências da estrutura primária sobre a reação de troca. Os experimentos foram realizados de forma a determinar as constantes de troca em dipeptídeos sintéticos em diversas condições de pH, estabelecendo a influência das que as diferentes cadeias laterais sobre um resíduo vizinho. Desta maneira foram mostradas as relações entre a estrutura primária e pH sobre a reação de troca em dipeptídeos, tornando a constante de reação da troca (referida como constante intrínseca ou k_{int}) calculável uma vez que se possui dados sobre a estrutura primária e as condições experimentais (considerando que este se encontra em uma região desestruturada), como já implementado em servidores como clntX e *Sphere* (86).

A constante intrínseca portanto, rege a reação de troca em dipeptídeos desestruturados, e é constantemente utilizada na interpretação de dados de NMR, através do cálculo do fator de proteção (PF) que trata da razão entre a constante intrínseca e a constante de troca observada no experimento – na proteína em estado nativo ou em intermediários de *folding* – expresso como PF= k_{int}/k_{obs} (87).

1.3.2. Métodos Computacionais.

1.3.2.1. Dinâmica Molecular

Desenvolvida nos anos 50 e 60 (88-90), a dinâmica molecular (*molecular dynamics* – MD) consiste em um método computacional para a simulação de sistemas de átomos com o intuito de estudar a evolução destes ao longo do tempo (91-93). O primeiro estudo de MD com o intuito de investigar o movimento de proteínas enoveladas foi realizado em 1977 (94).

Para o estudo de MD de uma proteína é necessária uma conformação inicial da mesma. Usualmente, utiliza-se modelos obtidos por cristalografia e difração de raios-X ou NMR para a simulação de proteínas no estado nativo. As propriedades dos átomos e de suas ligações e interações são representadas pelo chamado “campo de forças”, que consiste em um conjunto de parâmetros empíricos ou provenientes de cálculos quânticos que descrevem as propriedades dos átomos e moléculas do sistema. As forças exercidas sobre cada átomo são descritas por uma função de energia potencial (91-93).

O método de MD despreza a existência de partículas subatômicas, utilizando um modelo onde os átomos são representados por esferas com massa, carga e raio definidos. Esse modelo onde os prótons, elétrons e nêutrons não são considerados é baseado na aproximação de Born-Oppenheimer (95), que assume que os elétrons se adaptam instantaneamente a uma nova posição do núcleo.

A MD possui limitações quanto à exploração da superfície de energia potencial, uma vez que a função que descreve a energia potencial em cada uma das conformações é complexa e dependente da posição de cada um dos *N* átomos da proteína. Isto faz com que a superfície de energia potencial seja uma hipersuperfície *N* dimensional. A dificuldade para a exploração de todos os movimentos na MD representa uma barreira para o estudo de fenômenos que ocorrem em grandes escalas de tempo, tornando a técnica computacionalmente custosa (96).

1.3.2.2. Análise de Modos Normais

A análise de modos normais (*Normal Mode Analysis* – NMA) representa uma alternativa interessante quando existe a necessidade de estudar

movimentos que ocorrem em escalas de tempo dificilmente acessíveis pela MD. A NMA trata de movimentos oscilatórios intrínsecos do sistema, que estão contidos em sua organização estrutural e podem ser decompostos em um número de movimentos (ou modos) iguais ao número de graus de liberdade conformacional (97). Cada movimento possui direções e frequências próprias, sendo independente dos outros modos de movimento. Em proteínas, o número de movimentos internos é igual a 3N-6, sendo N o número de átomos que compõem o sistema. Desta maneira é possível decompor os movimentos internos de uma molécula de proteína em 3N-6 modos, permitindo também que estes sejam ordenados de acordo com sua frequência associada. Usualmente modos de baixa frequência representam movimentos mais coletivos – ou seja, envolvendo mais átomo se movendo de forma correlacionada - e de maior amplitude (como por exemplo, movimentos de abertura de domínios). Por outro lado, modos de alta frequência descrevem movimentos menos coletivos e de menor amplitude (como estiramentos de ligações) (97-99).

3.2.2. Fundamentação teórica

Enquanto a MD trata da resolução numérica das equações de movimento de Newton para a obtenção das posições dos átomos ao longo do tempo, a NMA é uma abordagem que utiliza uma resolução analítica para estas equações, levando em consideração uma superfície de energia potencial aproximada (quadrática), para um sistema que se encontra em um mínimo de energia (Figura 10).

A NMA é uma técnica que permite explorar movimentos que acorrem em escalas de tempo usualmente não acessíveis pela MD. Para isso, utiliza-se os mesmos parâmetros dos campos de forças empregados em simulações de MD, porém ao invés de obter as trajetórias dos átomos ao longo tempo, tem como resultado as frequências e direções de cada um dos modos normais de vibração da molécula (98, 99).

A NMA se baseia no estudo de estruturas em mínimos locais de energia potencial, onde a forma do potencial é relativamente simples. Dado que a molécula está em uma região de mínimo de energia (q_0), e este potencial pode ser expandido em uma série de Taylor, consideramos uma aproximação quadrática – desprezando os termos de segunda ordem ou de ordens superiores

da série de Taylor –, sendo assim, a energia potencial aproximada *V* de um sistema com coordenadas internas q_i é descrita por:

$$V = \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)\eta_i\eta_j = \frac{1}{2}V_{ij}\eta_i\eta_j \qquad [2]$$

Onde o termo η_i representa o desvio da posição de equilíbrio ($\eta_i = q_i - q_{0i}$).

Assim como a energia potencial, a energia cinética *T* também é tratada como uma aproximação quadrática, desta forma define-se a função Lagrangiana como *L=T – V,* que leva a *n* equações diferenciais lineares de movimento:

$$T_i\ddot{\eta}_i + V_{ij}\eta_j = 0 \quad [3]$$

Assumindo-se uma solução oscilatória para a equação acima, obtém-se:

$$A^T V A = \lambda \quad [4]$$

Onde *A* representa uma matriz de amplitudes e *V* representa uma matriz contendo as segundas derivadas da energia potencial (referida como matriz Hessiana) e λ representa a matriz diagonal.

Com a diagonalização da matriz hessiana, é possível obter seus autovetores (A_k) e seus autovalores (λ_k) associados. Esses correspondem respectivamente às direções dos movimentos de cada modo normal *k* e às frequências (ω_k) dos mesmos, sendo $\omega_k = \sqrt{\lambda_k}$. De posse dos 3N modos, desconsidera-se os 6 modos de rotação e translação do sistema, e utiliza-se os 3N-6 movimentos internos do mesmo. Dentre estes modos, os de mais baixa frequência tendem a ser movimentos mais coletivos, normalmente relacionados com funções das proteínas, enquanto os modos de mais alta frequência representam movimentos menos coletivos.

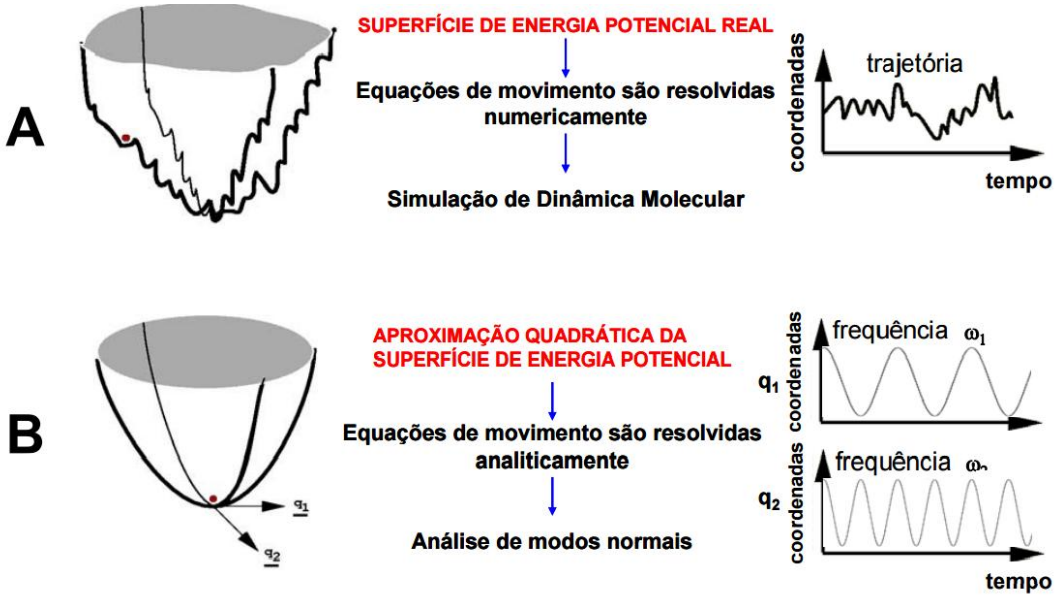


Figura 10 - Comparação entre MD e NMA
Em A temos a representação de uma superfície de energia potencial hipotética, que apresenta diversas irregularidades e é definida por uma função extremamente complexa. O gráfico da direita representa esquematicamente como as coordenadas se modificam ao longo do tempo. Em B está representada a superfície de energia potencial aproximada utilizada para o cálculo de NMA, onde as equações de movimento serão resolvidas de forma analítica, assumindo que próximo ao mínimo a energia potencial pode ser representada por uma aproximação quadrática, os gráficos da esquerda demonstram como as coordenadas variam periodicamente (100).

1.3.2.3. Métodos Estocásticos

Métodos estocásticos como o método de Monte Carlo são amplamente utilizados para a geração de *ensembles* de estruturas, minimização de energia e até mesmo em estudos de enovelamento (101-103). Métodos de Monte Carlo consistem em abordagens que utilizam amostragens aleatórias. No caso de uma molécula é possível gerar alterações conformacionais de forma aleatória adotando critérios para a aceitação ou não de cada mudança gerada (101). É possível também utilizar dados experimentais como critérios para a geração de *ensembles* para criar estruturas condizentes com experimentos (104).

Outro método desta categoria é a abordagem de *simulated annealing*, para busca de máximos ou mínimos locais utilizando uma busca probabilística em ciclos (105). O método é uma analogia ao processo metalúrgico de aquecer e esfriar os metais para que os átomos a cada ciclo tenham energia para buscar

uma posição ótima (106). Esta abordagem é utilizada para a elucidação de estruturas utilizando restrição de posição obtidas por NOE (107).

1.4. Modelos computacionais para a predição de dados HX

Diversos autores tentaram empregar ferramentas computacionais para predizer dados experimentais de HX (62, 108-111). Porém, ainda hoje a predição dos dados de troca utilizando estruturas de proteínas e cálculos computacionais continua sendo um problema, o que leva constantemente ao questionamento dos fatores determinantes do fenômeno de troca.

Uma vez que os hidrogênios amídicos precisam estar expostos para serem trocados, assume-se uma relação entre a área de superfície acessível ao solvente (*solvent accessible surface area* - SASA) obtida pelo método de Shrake-Rupley (112), análogo a rolar uma esfera de 1.4 Å sobre a superfície da estrutura de uma proteína, calculando desta maneira a área da superfície que estaria em contato com a água. A correlação entre SASA e dados de HX foi demonstrada para Thrular *et al.* (113) na enzima metilesterase, esse mesmo estudo demonstrou também uma correlação entre os fatores B e os dados de HX para essa mesma proteína.

Os modelos atuais também assumem a proteção dos hidrogênios comprometidos em ligações hidrogênio, uma vez que é preciso que eles estejam livres para trocar com o solvente (62, 104). O modelo mais utilizado para explicar a troca através de dados estruturais considera que o PF é determinado pelo número de contatos deste resíduo e pela presença ou não de ligações hidrogênio como visto na equação 1, onde existe um termo N_{hb} que representa o número de ligações hidrogênio, e um termo N_c que representa o número de contatos, os coeficientes (β) de cada termo são obtidos ao ajustar o modelo a um *dataset* por regressão linear. Best *et al.* (62) Vendruscolo *et al.* (104) utilizaram esse modelo fenomenológico ajustado a um grupo de proteínas para gerar *ensembles* de estruturas por métodos estocásticos, utilizando os dados experimentais como restrições.

$$PF = \beta_{hb} N_{hb} + \beta_c N_c \quad [1]$$

Alguns modelos recentes utilizam simulações de dinâmica molecular para realizar as predições dos dados de HX. Park *et al.*(114) utilizou conformações obtidas por MD para construir um modelo capaz de predizer dados de HX obtidos

por MS, o modelo utiliza como informação a presença de ligações hidrogênio nos resíduos em cada conformação obtida ao longo da simulação e mostrou fortes correlações com os dados experimentais, porém o uso de MD para a exploração de mudanças conformacionais acarreta grande custo computacional.

Assumindo o modelo da Equação 1, foi desenvolvido um método para a predição de HX utilizando apenas informação da estrutura primária (111), esse método foi aplicado em um servidor não mais existente intitulado camP. Embora a predição seja realizada utilizando apenas a estrutura primária, uma rede neural foi treinada utilizando um banco de dados de 2000 estruturas descrevendo o dado de troca através do modelo fenomenológico da equação 1. As correlações entre os dados preditos e experimentais variaram entre 0.5 e 0.7.

Também foram utilizadas informações de estrutura primária para criar modelos estatísticos capazes de prever o grau de proteção de um determinado resíduo de aminoácido de uma proteína (109). Para isso, foi calculada a propensão de cada resíduo de uma dada sequência a estar envolvido em ligações hidrogênio, assim como a densidade de contatos, com base em um banco de dados de estruturas proteínas globulares. O algoritmo leva em consideração ambas as informações para predizer se o resíduo está protegido ou não. Porém nesse estudo, não foi possível determinar o quão protegido está um resíduo, e o algoritmo se baseia em um valor de corte para determinar se um determinado resíduo está ou não protegido.

Bahar e colaboradores (115) utilizaram um abordagem diferente, aplicando o modelo de redes gaussianas (*Gaussian Networks Model* – GNM) que trata de um modelo simplificado para calcular a flexibilidade da molécula a partir de aproximações semelhantes às da NMA. Os dados obtidos nesse trabalho indicaram qualitativamente relações entre os dados de flutuações calculadas por GNM e os dados HX.

Skinner *et al.* (116) utilizou dados de HX obtidos para a proteína nuclease estafilocócica (SNase) para testar dois modelos preditivos, um baseado na geração de *ensembles* para determinar a estabilidade de proteínas (117), e o modelo de Vendruscolo *et al.* (104), demonstrando que esses modelos falham e demonstram baixas correlações com os dados experimentais da SNase, como é possível observar na Figura 11.

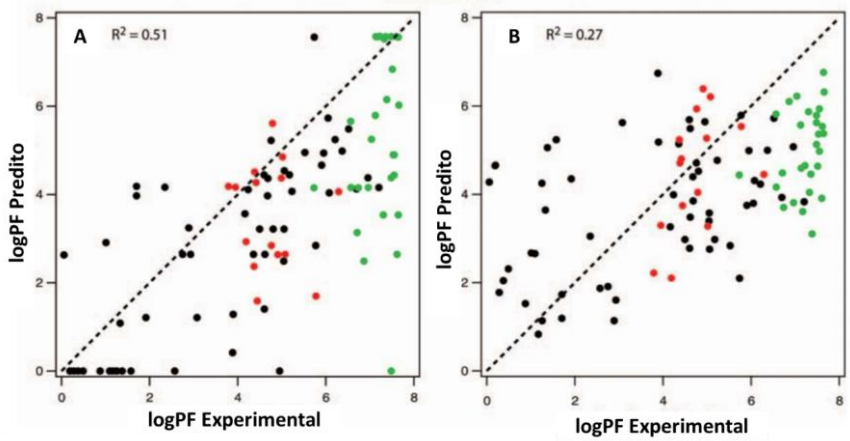


Figura 11 - Correlações entre os dados de HX experimentais e preditos para a enzima SNase.

O modelo descrito por Best *et al* (62) foi utilizado em A e o modelo de Hilser *et al.* (117) em B. As cores dos pontos representam mecanismos de troca de cada um dos resíduos da enzima, classificados por Skinner *et al.* (116) como flutuações locais (vermelho), grandes desenovelamentos (verde) e desconhecidos (preto).

Além de testar os modelos existentes, também foram discutidos outros fatores que influenciam o fenômenos da troca dos hidrogênios, demonstrando que em alguns casos a exposição de um hidrogênio ao solvente não necessariamente implica em troca, uma vez que ele pode estar envolvido em ligações hidrogênio com aceptores de prótons da proteína ou do solvente – como observado em dados cristalográficos - discutindo também que a proteção dos resíduos na superfície da proteína pode se dar por potenciais eletrostáticos dos resíduos adjacentes. Nesse mesmo trabalho, afirma-se que a troca pode ocorrer por diferentes mecanismos, e que um algoritmo preditivo para o fenômeno de troca deveria levar em conta o mecanismo pelo qual a troca ocorre em cada hidrogênio para que a predição pudesse ser mais acurada.

Assim, ainda restam muitas perguntas em aberto e uma ampla discussão na literatura sobre os determinantes do fenômeno de HX, visto que a compreensão dos detalhes por trás da troca implica diretamente em uma melhor interpretação dos dados e em novas possibilidades para seu uso, tal como a geração de *ensembles* baseados em dados experimentais.

CAPITULO II - OBJETIVOS

Este trabalho tem como objetivo geral a investigação de parâmetros estruturais e dinâmicos capazes de explicar a troca hidrogênio/deutério em proteínas em estado nativo (mecanismo EX2), visando a construção de modelos preditivos e explicativos baseado nos fatores supracitados.

2.1 . Objetivos Específicos

2.1.1. Modelagem dos dados de MS-HX

- Investigar a influência de parâmetros estruturais (número de contatos e ligações hidrogênio) e das flutuações obtidas por NMA e modelo de redes elásticas na troca hidrogênio/deutério, através da utilização de modelos lineares ajustados aos dados de troca de cada proteína contida em um *dataset*.
- Investigar as influências dos fatores B cristalográficos na predição do fenômeno de troca.

2.1.2. Modelagem dos Dados de NMR-HX

- Investigar a influência de parâmetros estruturais (número de contatos, ligações hidrogênio, acessibilidade ao solvente e estrutura secundária), assim como os diferentes critérios geométricos para o cálculo do número de contatos em cada proteína contida em um *dataset* de NMR-HX.
- Verificar a eficácia da utilização de informações de estrutura secundária e acessibilidade ao solvente calculados pelo algoritmo *dssp* na predição dos dados de HX.
- Estudar o efeito da temperatura e do pH em modelos lineares ajustados a dados de HX obtidos por NMR

- Criar modelos para a predição de dados de HX através de *i.* regressão linear e *ii.* através de um algoritmo de aprendizado de máquina (*random forest)* e avaliar os modelos através de validação cruzada.

CAPITULO III - MATERIAIS E MÉTODOS

3.1. Construção do *dataset*

O *dataset* de proteínas para a modelagem do fenômeno de HX foi dividido em duas partes: *i.* proteínas com dados experimentais obtidos por MS-HX e *ii.* por NMR-HX. Estes dados foram obtidos diretamente de artigos científicos da literatura ou foram fornecidos pelos autores, após demanda. Existem algumas diferenças principais entres os dados proveniente dessas duas metodologias. No caso da MS-HX, como antes da análise a proteína alvo é digerida por uma enzima (no caso a pepsina), as análises são feitas nos peptídeos resultantes dessa clivagem e são apresentados na forma de porcentagem de troca hidrogênio/deutério (%D) para cada um dos peptídeos obtidos pela digestão enzimática. No entanto, recentemente foi possível obter os dados de MS-HX em resolução de resíduo, para a enzima SNase [cedidos por Kan *et al*. (118)]. Já no caso da NMR-HX, o dado experimental pode ser representado em algumas formas: K_{ch} (constante de troca) ou PF (fator de proteção), ou logPF (logarítimo do PF). Para fins de uniformização, os dados experimentais foram convertidos, se necessário, para logPF.

A estrutura tridimensional correspondente a cada uma das proteínas do *dataset* foi obtida no PDB. A matriz de identidade entre as sequências das proteínas do *dataset* foi calculada usando o servidor MUSCLE (119), para evitar vieses devido à semelhança entre as proteínas usadas nos modelos.

3.3. Cálculos de parâmetros estruturais e dinâmicos

3.3.1. Preparo das estruturas

De posse das estruturas tridimensionais obtidas no PDB, foi utilizado o programa *pdb2pqr* (120) para adicionar os hidrogênios a cada estrutura, de acordo com as predições do programa *propka* (121) (que é implementado internamente no pdb2pqr). Este *software* determina os estados de protonação mais prováveis dos resíduos tituláveis levando em conta o pH correspondente às condições de cada experimento de HX.

A seguir, as topologias referentes a cada proteína foram geradas utilizando o *software* CHARMM e o campo de forças CHARMM 27 (122). Para moléculas não proteicas que não estão parametrizadas no campo de forças, – tais como ligantes presentes em algumas das proteínas do *dataset* – o servidor CHARMM-GUI foi utilizado para gerar as topologias dos ligantes a partir do CHARMM *General Force Field* (123).

3.3.2. Cálculo de Parâmetros Estruturais

Diversos modelos explicativos/preditivos descritos na literatura utilizam frequentemente o número de contatos e de ligações hidrogênio como fatores descritores/preditores para o estudo da HX em proteínas. Dentre eles, destaca-se o modelo linear proposto por Vendruscolo *et al*. (104), descrito pela Equação 1. Nesta dissertação, investigar-se-á a influência destes parâmetros estruturais (assim como a adição de outros que representam propriedades dinâmicas) nos modelos estatísticos aqui criados.

3.2.2.1. Número de contatos e de ligações hidrogênio

O número de contatos (N_c) de um resíduo *i* foi calculado considerando como um contato cada resíduo vizinho com ao menos um átomo do *backbone* dentro de um raio de corte (r_c) de 6.5 Å do nitrogênio amídico do resíduo *i,* conforme Vendruscolo *et al.* (104).

Para os dados de NMR-HX, diversos valores de r_c foram levados em consideração ($r_c = d/2$ Å, com *d* variando de 12 a 17), afim de determinar o raio ótimo para o cálculo do N_c. Outros critérios para o N_c também foram testados, como por exemplo: considerar apenas os átomos da cadeia principal, ou todos os átomos dos resíduos. Além disso, duas formas de contabilização dos contatos foram adotadas: número de átomos em contato com o nitrogênio amídico do resíduo *i;* ou o número de resíduos em contato com o resíduo *i* (considerando ao menos um átomo do resíduo dentro do raio de corte). A Figura 12 mostra uma representação esquemática dos critérios para o cálculo do N_c.

Para determinar o número de ligações hidrogênio (N_{hb}) foi considerado o seguinte critério geométrico: a presença do aceptor de próton em um raio de corte de 2.4 Å a partir ao próton ligado ao nitrogênio amídico, como proposto por Best *et al.* (62).

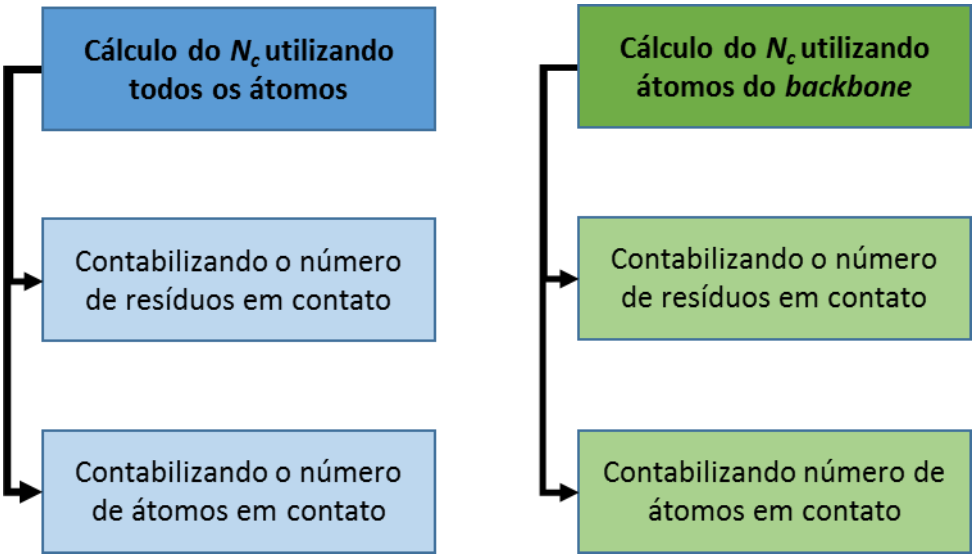


Figura 12 - Representação esquemática dos critérios utilizados para o cálculo do N_c para a modelagem de dados de NMR-HX
O esquema representa os quatro critérios utilizados para o cálculo da variável N_c durante a modelagem dos dados de NMR-HX. Estes foram: 1) considerando todos os átomos do sistema, contabilizando o número de resíduos em contato; 2) todos os átomos do sistema, contabilizando o número de átomos em contato; 3) apenas átomos do *backbone,* contabilizando o número de átomos em contato e 4) apenas átomos do *backbone,* contabilizando o número de átomos em contato.

3.2.2.2. Estrutura secundária

Para calcular a que tipo de estrutura secundária pertence cada resíduo, foi utilizado o programa *dssp* (*Define Secondary Structure of Proteins*)(124) implementado no pacote bio3d do *software* R (125). Este algoritmo realiza a predição de ligações hidrogênio através de um critério energético. Após a predição das ligações hidrogênio, o programa utiliza os padrões de ligação e outros critérios geométricos para classificar as estruturas secundárias em uma dentre 8 classes, sendo essas: hélices (Hélice 3_{10}, α hélice e hélice π), ponte *β* e folha *β* , *turns* (regiões onde existe ligação hidrogênio entre $CO_{(i)}$ to $NH_{(i+n)}$ sendo n=3, 4 ou 5), *bends* (regiões de alta curvatura, onde os ângulos envolvendo 3 carbonos α são inferiores a 70º) e alças (regiões que não se encaixam em outras classes) (124).

3.2.2.3. Área da superfície acessível ao solvente

A área de superfície acessível ao solvente (SASA) foi calculada utilizando o programa *dssp*, implementado no pacote bio3d para o *software* R. Para realizar o cálculo, utiliza como sonda uma esfera de 1.4 Å ao longo da superfície da

proteína, calculando para cada resíduo a área de superfície acessível à esfera e portanto, também considerada acessível ao solvente. Os pontos que tocam a esfera são considerados expostos ao solvente, uma vez que a esfera possui dimensões semelhantes às de uma molécula de água (112).

3.4. Parâmetros dinâmicos

Para a criação dos modelos dinâmicos, três parâmetros foram selecionados para representar a flexibilidade das proteínas: flutuações calculadas a partir de NMA, flutuações calculadas a partir de ENM e os fatores B cristalográficos.

3.4.1. Análise de Modos Normais

3.4.2. Cálculo dos modos normais

Para o cálculo dos modos normais de um sistema é necessário que este esteja em uma região de mínimo na superfície de energia potencial. Para isto, após o preparo das estruturas (como descrito no item 3.2.2.1), estas foram submetidas à minimização de energia por otimização das geometrias moleculares utilizando o programa CHARMM. Foi utilizado o método de gradiente conjugado, adotando como critério de parada variação menor que 10^{-5} kcal/mol/Å^2. Partindo das estruturas otimizadas, os 200 modos de mais baixa frequência foram calculados usando os módulos DIMB e VIBRAN, implementados no CHARMM (126, 127). Foi utilizado um raio de corte de 11 Å para a definição dos pares de átomos não ligados, sendo a partir de 5 Å de distância aplicada uma função de *switch* para assegurar que os potenciais eletrostáticos e de Van der Waals alcancem valor zero em distâncias de 9 Å ou superiores. O valor utilizado para a constante dielétrica foi de 2 F/m.

3.4.3. Flutuações dos Modos Normais

A raiz da flutuação quadrática média (*root mean square fluctuation* – RMSF) representa as flutuações dos átomos do sistema ao longo de uma trajetória (no caso de análises de uma simulação de MD). Já as flutuações dos

modos normais ($RMSF_{NMA}$), correspondem à flutuação dos átomos quando deslocados ao longo das direções dos modos normais.

Para a utilização no modelo, as flutuações de cada proteína foram calculadas a partir dos 100 modos internos de mais baixa frequência. As flutuações dos modos normais são calculadas de acordo com a equação 3.

$$\langle \Delta r_i^2 \rangle = k_b T \sum_{j=1}^{n} \sum_{\alpha=1}^{3} \frac{q_{i\alpha,j}^2}{\omega_j^2} \qquad [3]$$

onde K_b é a constante de Boltzmann, T é a temperatura absoluta, Δr_i é o deslocamento do átomo i com relação à posição de referência da estrutura minimizada, $q_{i\alpha,j}$ é o elemento correspondente ao i-ésimo átomo no j-ésimo vetor de modos normais. O índice α (1, 2, 3) indica o eixo de coordenadas cartesianas e ω_j indica a frequência do modo normal.

Os deslocamentos ao longo dos vetores q_j são expressos na forma de uma função de raiz quadrática média ponderada pela massa (MRMS - *mass weighted root mean square*), como mostrado na equação 4.

$$d_{j\ MRMS} = \frac{1}{\sqrt{M}} \sum_{i=1}^{3N} \sqrt{m_i}\left(r_i - r_i^0\right) q_{ij} \quad [4]$$

onde i corresponde a um dado grau de liberdade relacionado a um átomo de massa m_i. M é a massa total e q_{ij} é o i-ésimo elemento do j-ésimo vetor de modos normais.

3.4.4. Modelo de redes elásticas

O modelo de redes elásticas (ENM – *Elastic Networks Model*) trata os resíduos de aminoácidos de uma forma simplificada, representando apenas o carbono-α. Os resíduos representados são então conectados por molas a outros resíduos que estejam dentro de um raio de corte. Com isso, a partir das forças exercidas sobre os pseudo-átomos é possível calcular os modos normais. Embora simples, o cálculo dos modos normais partindo da abordagem de ENM possui correlações com dados experimentais de fatores B cristalográficos (128).

Devido às aproximações inerentes a esse modelo, os ligantes das proteínas não foram considerados. Neste trabalho as flutuações do ENM foram calculadas utilizando a biblioteca *bio3d* para o *software* R. Foi utilizado o método de Hinsen *et al.* implementado no pacote *bio3d* (129, 130) que aplica molas com constantes de força dependentes da distância, assumindo 2,9 Å como raio de corte mínimo para considerar interações entre átomos, as flutuações foram obtidas levando em consideração todos os modos calculados para cada proteína.

3.5. Modelagem Estatística

3.4.1. Modelagem dos dados de MS-HX

A etapa de modelagem estatística se divide em duas partes: *i.* modelagem explicativa dos dados de HX e *ii.* modelagem preditiva dos dados (aplicada somente aos dados de NMR-HX).

Para tal, foram utilizados modelos lineares (*lm*) utilizando regressão pelo método dos mínimos quadrados, isto é, de forma que os coeficientes obtidos para cada parâmetro e o intercepto da função retornem o menor valor do somatório dos quadrados dos erros, sendo o erro igual à diferença entre os valores ajustados do modelo e os dados experimentais.

Além dos modelos lineares, os dados de NMR-HX também foram modelados através do método *Random Forest* (131) para explorar relações não lineares entre as variáveis utilizadas e os dados modelados. Todas as etapas de criação e análise dos modelos foram realizadas utilizando o *software* R.

Em cada um dos modelos foram estudados os coeficientes associados a cada parâmetro, também foram analisadas as correlações entre os valores ajustados dos modelos e os dados experimentais. O RMSE (equação 5) foi empregado afim de avaliar o erro de ajuste do modelo, onde *n* corresponde à *n*-ésima observação, e $\hat{y}$ e y são respectivamente a variável ajustada do modelo e o valor do dado experimental.

$$RMSE = \sqrt{\frac{\sum_1^n (\hat{y}-y)^2}{n}} \quad [5]$$

A fim de avaliar o efeito da introdução de novas variáveis nos modelos lineares, foi utilizado o critério de informação de Akaike (*Akaike Information Criterion* - AIC) de cada modelo. O AIC é uma função para avaliação de modelos que inclui o valor máximo da função de verossimilhança do modelo (L), e uma penalidade para a inclusão de um novo parâmetro (K). Assim, ao incluir um novo parâmetro no modelo, este pode ser comparado com um modelo anterior onde o parâmetro não estava incluído. Valores menores de AIC representam melhores modelos (Equação 6) (132).

$$AIC = 2k - 2\ln(L) \quad [6]$$

Os modelos *Estrutural* e *Estrutural + variável dinâmica* foram comparados utilizando ANOVA. Uma visão geral da modelagem dos dados de MS-HX pode ser vista na Figura 13.

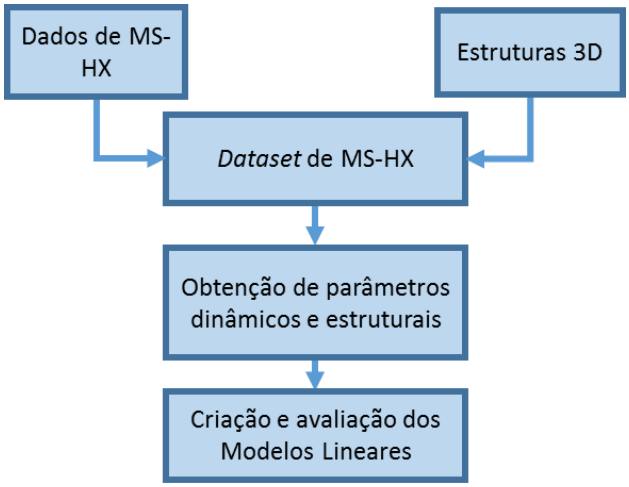


Figura 13 - Representação Esquemática da metodologia para a modelagem de dados de MS-HX
O esquema mostra a os passos tomados desde a obtenção dos dados da literatura até a criação dos modelos lineares para explicar os dados de MS-HX.

3.4.2. Modelagem Estatística dos Dados de NMR-HX

Em um primeiro momento foram analisados os critérios ótimos para o cálculo do N_c, a fim de definir quais valores da variável em questão são melhores preditores para os fatores de proteção. Para isso, o modelo *Estrutural + NMA* foi criado para cada proteína empregando cada um dos critérios utilizados no cálculo do N_c descritos no tópico *3.2.2.1.* Após definir o critério ótimo para o

cálculo, as proteínas cujos modelos ajustados tiveram as maiores correlações com os dados experimentais foram agrupadas em um *dataset* menor. Partindo do *dataset* reduzido, foram criados modelos empregando apenas uma das variáveis calculadas de cada vez. Em seguida as variáveis pH e temperatura foram adicionadas afim de introduzir as condições experimentais nos modelos.

Foram analisados os valores de AIC de cada um dos modelos, assim como os valores de *RMSE* e as correlações entre os valores ajustados e os dados experimentais. Dentre esses modelos criados para o *dataset* reduzido, aquele que gerou o ajuste com os menores valores de AIC e RMSE e maiores valores de correlação, foi selecionado para a criação de modelos preditivos nesse mesmo conjunto de dados.

Os modelos preditivos foram criados através de regressão linear pelo método dos mínimos quadrados ou *random forest,* e avaliados por validação cruzada utilizando o método *leave-one-out*. Nesta metodologia de validação, uma proteína é retirada do dataset e os modelos de são ajustados ou treinados com as proteínas restantes no dentro do conjunto de dados, em seguida é realiza uma predição dos dados da proteína que havia sido retirada. O RMSE e coeficiente de correlação de Pearson foram calculados para cada teste da validação. A representação esquemática dos passos para a modelagem dos dados de NMR-HX pode ser vista na Figura 14.

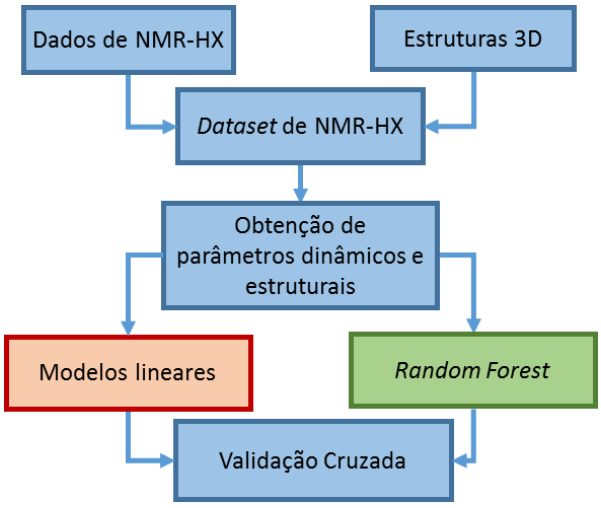


Figura 14 - Representação Esquemática da metodologia para a modelagem de dados de NMR-HX
Representação dos passos para a modelagem dos dados de NMR-HX desde a montagem do *dataset* até a construção dos modelos (tanto lineares quanto construídos por random forest) e finalmente a validação cruzada.

3.5. Random Forest

O método de *Random forest* (RF) é uma abordagem estatística utilizada para regressão ou classificação e se baseia nos métodos de árvores de decisão.

O método de árvores de decisão é utilizado para classificação ou regressão de um grupo de dados. Em um primeiro passo o algoritmo utiliza uma variável independente que divide o grupo de dados em dois subgrupos, gerando assim dois nós filhos a partir do nó pai. São utilizadas então outras variáveis independentes para separar os próximos nós até que sejam gerados os nós terminais da árvore (ou folhas), onde estarão contidos os resultados da classificação (Figura 15). Assim, ao treinar uma árvore com um determinado grupo de dados, é possível classificar um novo dado.

A abordagem de RF consiste em criar *ensembles* de árvores de regressão de forma que cada nó dessa árvore seja gerado utilizando uma decisão randômica para escolher entre as variáveis candidatas à cada ramificação. Assim para um número grande de árvores, as variáveis que são fortes preditores estarão presentes em mais árvores do que outras variáveis. Após o treinamento do *ensemble* de árvores, a predição para um novo dado é realizada através das médias das predições de todas as árvores (131, 133). No presente trabalho *ensembles* foram gerados com 500 árvores e o número de variáveis por árvore foi definido como o número de variáveis disponíveis dividido por 3, seguindo a configuração *default* do modelo de RF do pacote *RandomForest* para o *software* R (131).

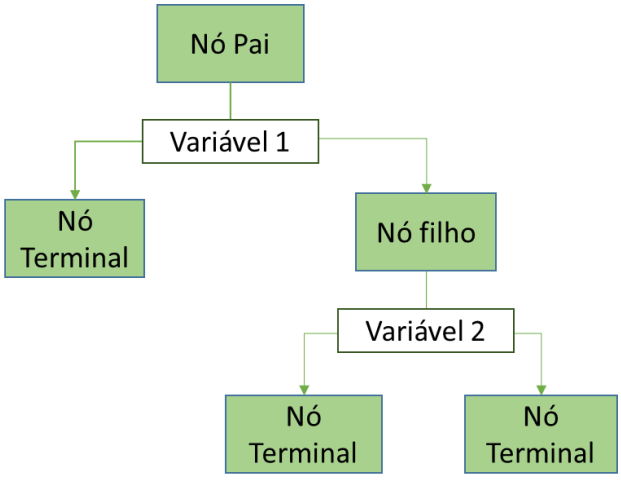


Figura 15 - Esquema de árvore de classificação
Os nós da árvore são representados na forma de quadros verdes, em cada ramificação são representadas as variáveis que dividem o nó pai em dois nós filhos. Adaptado de Lemon *et al* (133).

CAPÍTULO IV. MODELAGEM DE DADOS DE MS-HX

Neste capítulo serão tratados os resultados relacionados à modelagem estatística dos dados de HX obtidos pela técnica de MS. A maior parte desses dados é representada por peptídeos resultantes da digestão pela pepsina. No entanto, para a proteína SNase, os dados experimentais disponíveis são para cada resíduo. Um artigo científico com os resultados referentes a este capítulo encontra-se em processo de finalização e está em processo de submissão.

4. RESULTADOS E DISCUSSÃO

4.1. Construção do *dataset*

O *dataset* contendo os dados de MS-HX e as respectivas estruturas 3D de cada proteína foi composto de 10 proteínas, algumas em diferentes condições (*apo/holo*), totalizando 12 sistemas com dados em resolução de peptídeos e uma proteína (SNase) com os dados em resolução de resíduo. A Tabela 1 mostra todas as proteínas contidas no *dataset,* assim como informações sobre classificação funcional, número de resíduos, condições experimentais da MS-HX e número de peptídeos obtidos (ou resíduos no caso da SNase) de cada uma das proteínas.

As proteínas selecionadas para este *dataset* apresentam estruturas e funções bastante diversas, sendo divididas em diferentes classes: hidrolases, redutases, proteínas estruturais, transportadoras e reguladoras. Alguns sistemas apresentam ligantes ou exibem diferentes estados oligoméricos. Por exemplo 1AQT e 2E5Y: onde o primeiro trata-se de um dímero ligado a uma molécula de ATP, enquanto o segundo é um monômero da mesma proteína. Já a proteína de código 1NFI apresenta-se em dois estados, ligada ou não ao NF-κB. No caso da hemoglobina, as cadeias alfa e beta foram analisadas separadamente, porém todas as simulações foram realizadas com a estrutura tetramérica, uma vez que o estado oligomérico tem influência sobre a dinâmica e sobre os parâmetros estruturais calculados.

Tabela 1 - Proteínas contidas no dataset de MS-HX

PDB ID	Proteína	Classificação Funcional	# de res.	Condições Experimentais	*N*	Ref.
1AQT (134)	ATP Sintase (dímero +ATP)	Hidrolase	138	pH = 7.0; 298 K; 10 min	6	(135)
2E5Y (136)	ATP sintase		133			
1EY8 (137)	SNase	Hidrolase	149	pH = média (8.6, 8.3, 5.6, 4.2); 293K;	8	(118)
1JSY (138)	arrestina-2	Sinalização	418	pH = 7.4; 298 K; 17 min	10	(139)
1NFI (140)	*apo IκBα*	Controle da transcrição	213	pH = 7.5; 298 K; 2 min	6	(141)
	holo IκBα					
1PU0 (142)	Superóxido dismutase	Oxido-redutase	153	pH = 7.2; 277 K; time = média (0.25, 0.8, 2.5 e 8.3 min)	8	(143)
2BBO (81)	NBD1 humana com Phe508	Transporte	291	pH = 7.0; 298 K; 77 min	5	(81)
2EYI (144)	*Apo* α-actina domínio CH_2	Estrutural	234	pH = 2.5; 277 K; tempo = média (0.25, 0.5, 1, 2, 5 e 15 min)	6	(145)
2NT1 (146)	*apo* GCase	Hidrolase	497	pH = 7.8; T = 296 K; 0.8, 1.6, 5, 16.6 e 50 min	8	(147)
2NSX (146)	*Holo* Gcase (isofagomina)					
2QSS (148)	Hemoglobina bovina	Transporte de oxigênio	141	pH = 7.2; 298 K; 120 min	30	(149)

Na coluna PDB ID as referências de cada estrutura encontram-se entre parênteses, a coluna *Ref*. apresenta as referências dos dados experimentais. As colunas mostram respectivamente: O código do PDB de cada proteína com a respectiva referência, o nome do sistema, a classificação funcional de acordo com o PDB, o número de resíduos de cada estrutura, as condições dos experimentos de HX, o número de peptídeos obtidos por MS-HX ou resíduos no caso da SNase e por último as referências bibliográficas de onde foram obtidos os dados de HX.

No intuito de avaliar se as proteínas incluídas no *dataset* eram suficientemente diferentes para representar de forma robusta uma diversidade significativa, comparamos suas sequências. Para isso, as sequências foram alinhadas e a porcentagem de identidade foi calculada utilizando o servidor

MUSCLE. Desta forma, foi construída uma matriz de identidade entre as proteínas contidas no *dataset.* As proteínas deste apresentaram identidades entre 5,5 e 41%, (Figura 16). Este resultado está de acordo com o critério utilizado no trabalho de Tartaglia *et al*, onde identidade inferior a 50 % com outras proteínas do *dataset* foi o critério de inclusão (111).

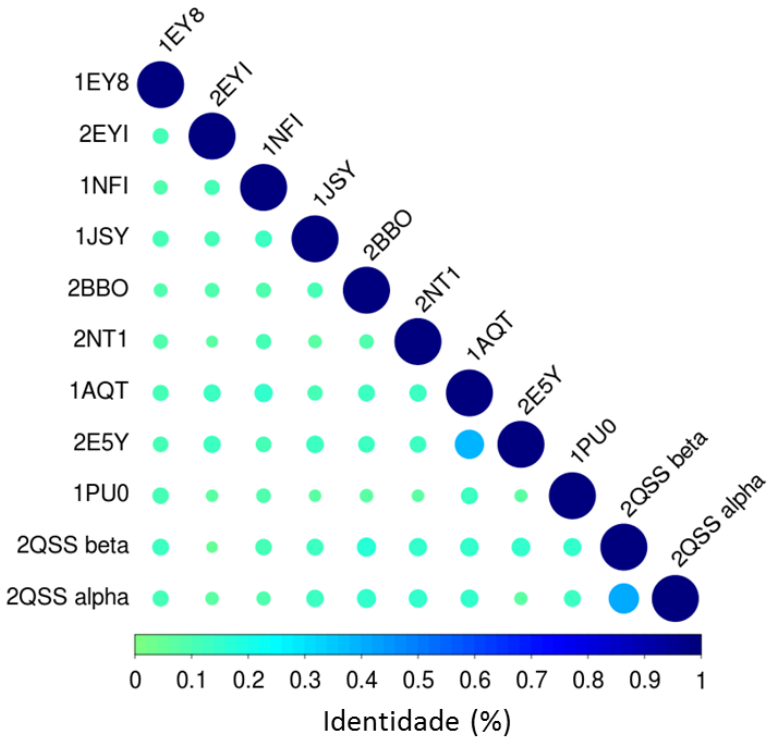


Figura 16 - Matriz de identidade entre as proteínas do *dataset* de HX-MS
As proteínas do *dataset* estão representadas por seus códigos do PDB. Os percentuais de identidade estão representados em esferas – esferas maiores em tons mais escuros representam valores mais próximos de 1, esferas menores e mais claras representam valores mais próximos de 0.

4.2 Construção e Análise dos Modelos

Com o intuito de analisar a influência de cada fator (estrutural ou dinâmico) no fenômeno de HX, foram construídos 6 modelos diferentes (como mostrado na Tabela 2) para cada um dos sistemas descritos na Tabela 1. Dois parâmetros estruturais foram calculados utilizando-se a estrutura 3D de cada proteína: Número de contatos e ligações hidrogênio; sendo estes designados aqui respectivamente como N_c e N_{HB}. A partir de simulações computacionais

utilizando as estruturas como ponto de partida, foram calculados os seguintes parâmetros dinâmicos: flutuações obtidas a partir de NMA, as flutuações de ENM e os fatores B cristalográficos, aqui designados respectivamente como: *$RMSF_{NMA}$, $RMSF_{ENM}$ e BFAC.*

Tabela 2 - Descrição dos modelos criados para os dados de MS-HX

Modelo	Equação
Contatos	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + e_i$
hbond	$\mathrm{HX}_i = \beta_0 + \beta_{HB} N_{HB_i} + e_i$
Estrutural	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{HB} N_{HB_i} + e_i$
Estrutural + NMA	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{HB} N_{HB_i} + \beta_{NMA} RMSF_{NMA_i} + e_i$
Estrutural + ENM	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{HB} N_{HB_i} + \beta_{ENM} RMSF_{ENM_i} + e_i$
Estrutural + BFAC	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{HB} N_{HB_i} + \beta_{BFAC} BFAC_i + e_i$

São representados os modelos e suas respectivas estruturas: intercepto, coeficientes e variáveis associadas e erro.

4.2.1. Modelos Estruturais

Os três primeiros modelos aqui tratados, empregam apenas variáveis estruturais. É possível observar na Tabela 3 uma visão geral da modelagem dos dados de MS-HX. São mostradas as correlações entre valores ajustados e experimentais, RMSF e AIC de cada modelo são apresentados para cada um dos sistemas estudados. As correlações são mostradas em negrito, o RMSE entre colchetes e o AIC de cada modelo entre parênteses, os asteriscos denotam os resultados do ANOVA quanto à significância da diferença entre o modelo *Estrutural* e o modelo *Estrutural* quando uma das 3 variáveis dinâmicas é introduzida ao mesmo.

Tabela 3 - Correlações, AIC, RMSE e análises ANOVA de cada modelo

Proteína	***Contatos***	***Hbond***	***Estrutural***	***Estrutural + NMA***	***Estrutural + ENM***	***Estrutural + Bfac***
2QSS α	[23,8] **0,36** (152,8)	[24,8] **0,22** (154,2)	[23,3] **0,4** (154,2)	[22,1] **0,5** (154,5)	[23,3] **0,4** (156,1)	[22,1] **0,49** (154,5)
2QSS β	[19,3] **0,7** (128,6)	[24,4] **0,42** (135,1)	[19,2] **0,7** (130,4)	[17,4] **0,76** (129,8)	[13,8] **0,86** (123,1)**	[13,8] **0,86** (123,3)**
1AQT	[20,2] **0,26** (147,6)	[20,9] **0,03** (148,7)	[20] **0,3** (149,2)	[19,5] **0,37** (150,4)	[19,7] **0,34** (150,8)	[19,9] **0,31** (151,2)
1JSY	[7,2] **0,39** (74)	[6,2] **0,62** (70,7)	[6,1] **0,62** (72,7)	[6] **0,65** (74)	[5,9] **0,67** (73,8)	[6,1] **0,63** (74,6)
2BBO	[20,9] **0,72** (139,7)	[29,5] **0,21** (150,1)	[20,9] **0,72** (141,7)	[18,1] **0,8** (139,5)*	[20,2] **0,74** (142,8)	[16,2] **0,84** (136,1)**
2E5Y	[29,6] **0,4** (159,8)	[30] **0,37** (160,3)	[27,9] **0,51** (159,9)	[22,1] **0,73** (154,5)**	[25,6] **0,61** (159,2)	[23,4] **0,69** (156,3)**
1PU0	[12,6] **0,29** (148,4)	[12,8] **0,24** (148,9)	[12,1] **0,41** (148,7)	[10,8] **0,58** (146,6)*	[11,7] **0,46** (149,8)	[11,1] **0,55** (147,7)
2NSX	[20,7] **0,46** (522,4)	[22,5] **0,27** (531,7)	[20] **0,52** (519,9)	[15,8] **0,74** (494,7)****	[16,3] **0,72** (498,3)****	[18] **0,64** (509,6)****
2EYI	[10,4] **0,5** (126,3)	[11,8] **0,16** (130,5)	[10,4] **0,5** (128,3)	[8,8] **0,68** (125)**	[9,1] **0,65** (126)*	[9,3] **0,63** (126,9)
2NT1	[21,3] 0,48 (525,3)	[23,3] **0,28** (535,7)	[20,4] **0,54** (522,3)	[18,6] **0,64** (513,8)***	[17,9] **0,67** (509,3)****	[18,8] **0,63** (514,8)***
1NFI *apo*	[24] 0,75 (61,2)	[26,8] **0,68** (62,5)	[20,6] **0,82** (61,3)	[9,9] **0,96** (54,6)	[19,6] **0,84** (62,7)	[19,7] **0,84** (62,8)
1NFI *holo*	[14,8] 0,56 (55,4)	[17,8] **0,01** (57,6)	[14,7] **0,56** (57,3)	[4,7] **0,96** (45,6)*	[13,4] **0,66** (58,2)	[12,7] **0,7** (57,6)
1EY8	[1,8] 0,63 (355,8)	[1,9] **0,56** (367,9)	[1,6] **0,73** (335,7)	[1,5] **0,78** (322,4)****	[1,5] **0,75** (331,9)**	[1,6] **0,74** (334)*

* $< 0,1$ *** $< 0,01$
** $< 0,05$ **** $< 0,001$

As correlações entre os valores ajustados e experimentais encontram-se em negrito, os valores de AIC encontram-se entre parênteses e os valores de RMSE entre colchetes. Os asteriscos denotam a significância da diferença entre os modelos quando testados utilizando o método ANOVA

Ao analisar os resultados obtidos para o modelo *contatos* na Tabela 3 é possível notar que os valores ajustados deste modelo nos sistemas 2BBO, 2QSS *β* e 1EY8 (que será discutido adiante), e nos sistemas 1NFI *holo* e *apo* apresentam correlações com os dados experimentais (respectivamente R = 0.75 e R = 0,56). É importante observar que no caso do último sistema citado, as correlações são ainda maiores na forma *apo*, aparentemente pelo fato de o N_c não levar em consideração as restrições conformacionais impostas por ligantes, representando apenas um estado particular. Embora existam correlações nesses 5 casos citados, os valores de RMSE são altos e deve-se também levar

em consideração que o sistema 1NFI possui apenas 6 peptídeos com dados de troca determinados.

Em seguida foi analisado o modelo *hbond*, que foi construído utilizando a variável N_{HB}. Por se tratar de uma variável binária (presença ou ausência da ligação hidrogênio), é esperado que não haja variações suficientes nos valores para que os dados sejam modelados corretamente. As correlações observadas entre os valores ajustados e os dados experimentais são fracas para a maioria dos modelos apresentados com exceção de 1EY8, 1JSY e 1NFI *apo*, neste último novamente as fortes correlações podem ser explicadas pelo pequeno número de peptídeos com dados de troca observados, e pelo efeito do intercepto.

O passo seguinte foi utilizar ambas as variáveis (N_c e N_{hb}), assim como realizado por Vendrusolo *et al*. (104) para modelar o fenômeno, este modelo foi designado aqui como *Estrutural.* Neste caso é possível observar aumentos nas correlações na maioria dos sistemas quando se compara com o modelo *contatos* ou *hbond.* Vendruscolo *et al.* descreveram uma observação similar, mostrando que a utilização do número de contatos ou ligações hidrogênio individualmente não é suficiente para descrever o fenômeno. Nesse mesmo trabalho, os modelos que empregam apenas N_{HB} não alcançaram correlações superiores a 0,4 (com exceção do caso do sistema 1NFI *apo*), e aqueles que utilizaram apenas N_c não tiveram correlações superiores a 0,5. Porém, foi demonstrado que existe um aumento considerável nas correlações quando ambas as variáveis são utilizadas em conjunto (104), corroborando os dados aqui apresentados.

4.2.2. Inclusão de Parâmetros Dinâmicos

Uma vez que o fenômeno de HX descreve o equilíbrio dinâmico das proteínas em solução, e estas por sua vez exploram não apenas uma, mas diversas conformações ao longo do tempo de experimento, o próximo passo foi a inclusão de variáveis que representam a dinâmica da proteína. Assim, os modelos *Estrutural + NMA*, *Estrutural + ENM* e *Estrutural + BFAC* foram criados, utilizando respectivamente as flutuações calculadas a partir da NMA, flutuações calculadas a partir de ENM e os fatores B cristalográficos em conjunto com o N_c e N_{HB}.

Na Tabela 3 é possível observar o aumento nas correlações entre os valores ajustados e os dados experimentais quando os parâmetros dinâmicos são introduzidos. Além disso, no caso da adição das flutuações de NMA ou ENM (com exceção do sistema 1JSY) observou-se diminuição de todos os valores de AIC, demonstrando que existe melhoria nos modelos *Estrutural* + *NMA* e *Estrutural* + *ENM* quando comparado ao modelo *Estrutural*. No caso do modelo *Estrutural* + *BFAC,* não houve diminuição dos valores de AIC em 4 dos sistemas, e nos demais a diminuição foi ínfima quando comparada com a dos modelos *Estrutural* + *ENM* e *Estrutural* + *NMA.* Também é importante salientar a diminuição nos valores de RMSE ao inserir a variável NMA. No que tange a comparação entre a utilização das flutuações obtidas por NMA e por ENM, a maioria dos modelos que utilizam NMA exibem maiores correlações com os dados experimentais do que os modelos que utilizam ENM, além disso, quando os modelos *Estrutural* + *NMA* e *Estrutural* + *ENM* são comparados com o modelo *Estrutural* pelo método ANOVA, a diferença é mais significativa quando se utiliza NMA.

Foram estudadas também as variações dos coeficientes associados às variáveis N_c, N_{HB}, $RMSF_{NMA}$ e $RMSF_{ENM,}$ em cada um dos modelos (Figura 17), é possível observar que salvo um *outlier*, os valores se agrupam ao redor de um mesmo ponto, como é possível observar em todos os painéis da figura.

Na análise seguinte, os valores ajustados dos peptídeos de todas as proteínas obtidos através dos modelos *Estrutural*, *Estrutural* + *NMA*, *Estrutural* + *ENM* e *Estrutural* + *BFAC* foram concatenados e comparados com os seus respectivos dados de porcentagem de troca, gerando apenas um valor de correlação para cada modelo (Figura 18). Nessa análise, é possível observar que embora a NMA tenha menores valores de RMSE e AIC para a maioria dos casos, as correlações obtidas ao concatenar os dados são bastante parecidas com os modelos que utilizam BFAC e ENM.

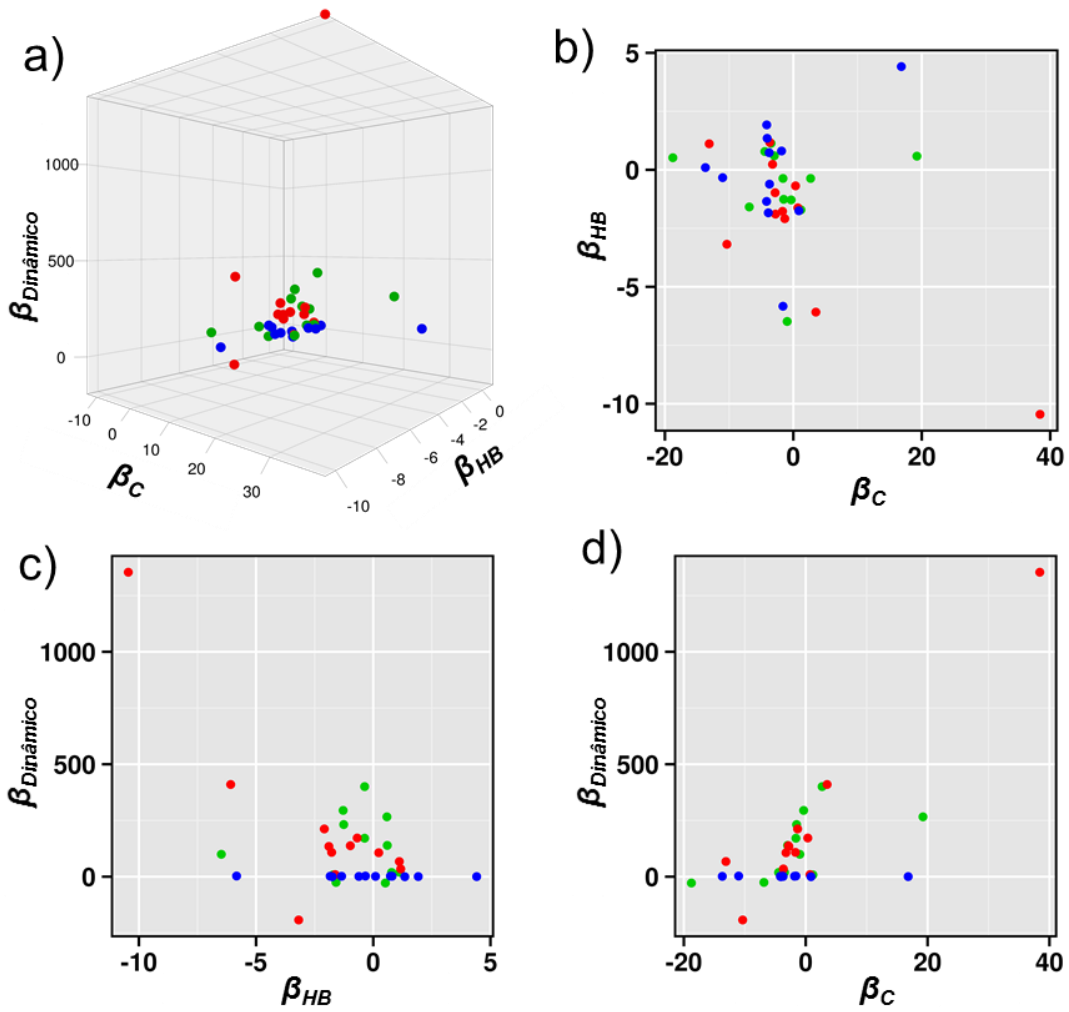


Figura 17 – Análise dos coeficientes das variáveis em cada modelo
Os quadros mostram a variação dos coeficientes β_{HB}, β_C e o coeficiente das variáveis dinâmicas nos modelos que foram construídos utilizando B-factor (azul), NMA (vermelho) e ENM (verde), sendo cada ponto uma das proteínas do *dataset*. a) Representação tridimensional dos coeficientes em cada um dos modelos; b) β_{HB} em função de β_c em cada um dos modelos construídos; c e d mostram o $\beta_{dinâmico}$ respectivamente em função de β_{HB} e β_c

As diferenças mostradas até aqui entre NMA e as outras variáveis escolhidas para representar a dinâmica de proteínas podem ser explicadas pelas limitações dos fatores B como indicativos de flexibilidade, uma vez que são medidas de cristais e os contatos cristalográficos das proteínas podem causar falsas impressões de estabilidade (22) e pelas aproximações inerentes ao método de ENM.

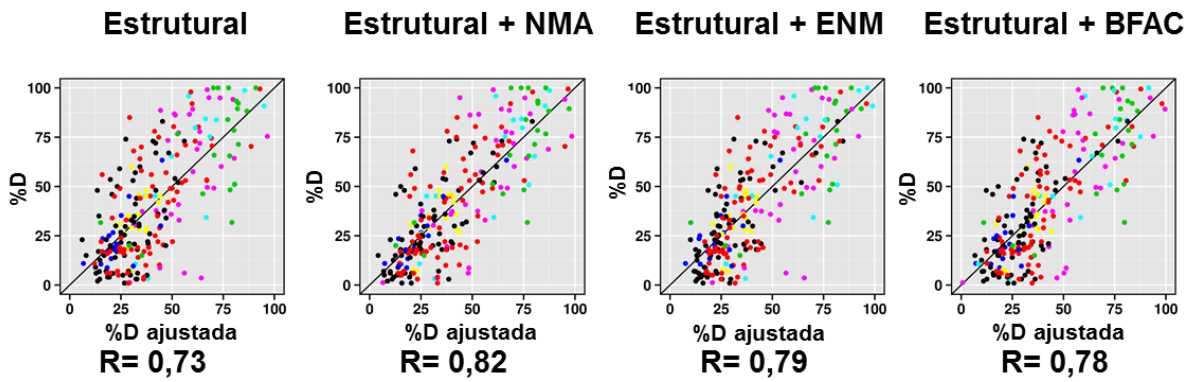


Figura 18 - Correlação entre os valores ajustados concatenados de todos os peptídeos e seus respectivos valores de %D
Os modelos lineares foram ajustados individualmente (um pra cada uma das proteínas) e representados simultaneamente no gráfico. Estes foram construídos utilizando apenas variáveis estruturais (*Estrutural*) ou incluindo cada um dos parâmetros dinâmicos (flutuações de NMA e ENM ou BFAC). Os valores ajustados para cada modelo de cada uma das proteínas (onde cada cor representa uma proteína) são representados no gráfico (concatenados) em relação ao dado experimental. Abaixo são mostradas as correlações entre os dados concatenados de todos os modelos contra seus respectivos dados experimentais

Uma vez determinada a importância das flutuações de NMA, foram realizadas comparações visuais entre os dados experimentais e os valores ajustados dos modelos *Estrutural* e *Estrutural + NMA* utilizando uma representação dos dados nas estruturas das proteínas (Figura 19). É possível observar que as representações dos valores ajustados nas estruturas são mais parecidas com a representação dos dados experimentais quando se utiliza a variável NMA em conjunto com o modelo *Estrutural*, porém, apesar de a maioria dos casos serem visualmente idênticos, a modelagem de algumas regiões ainda é dificultosa (como por exemplo as regiões terminais dos sistemas 2QSS *α* e 2QSS *β* e regiões de *loop* em 2EYI).

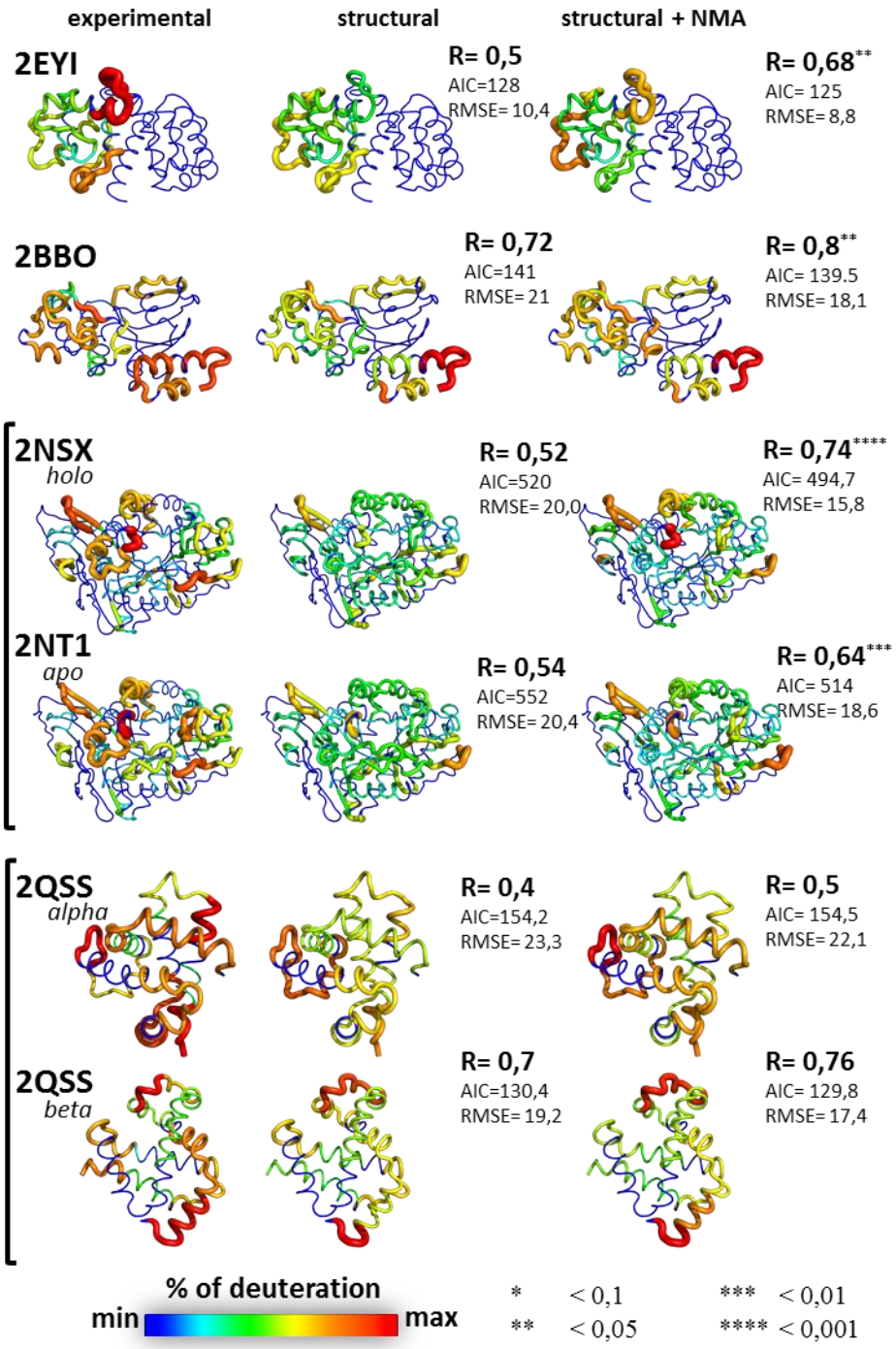
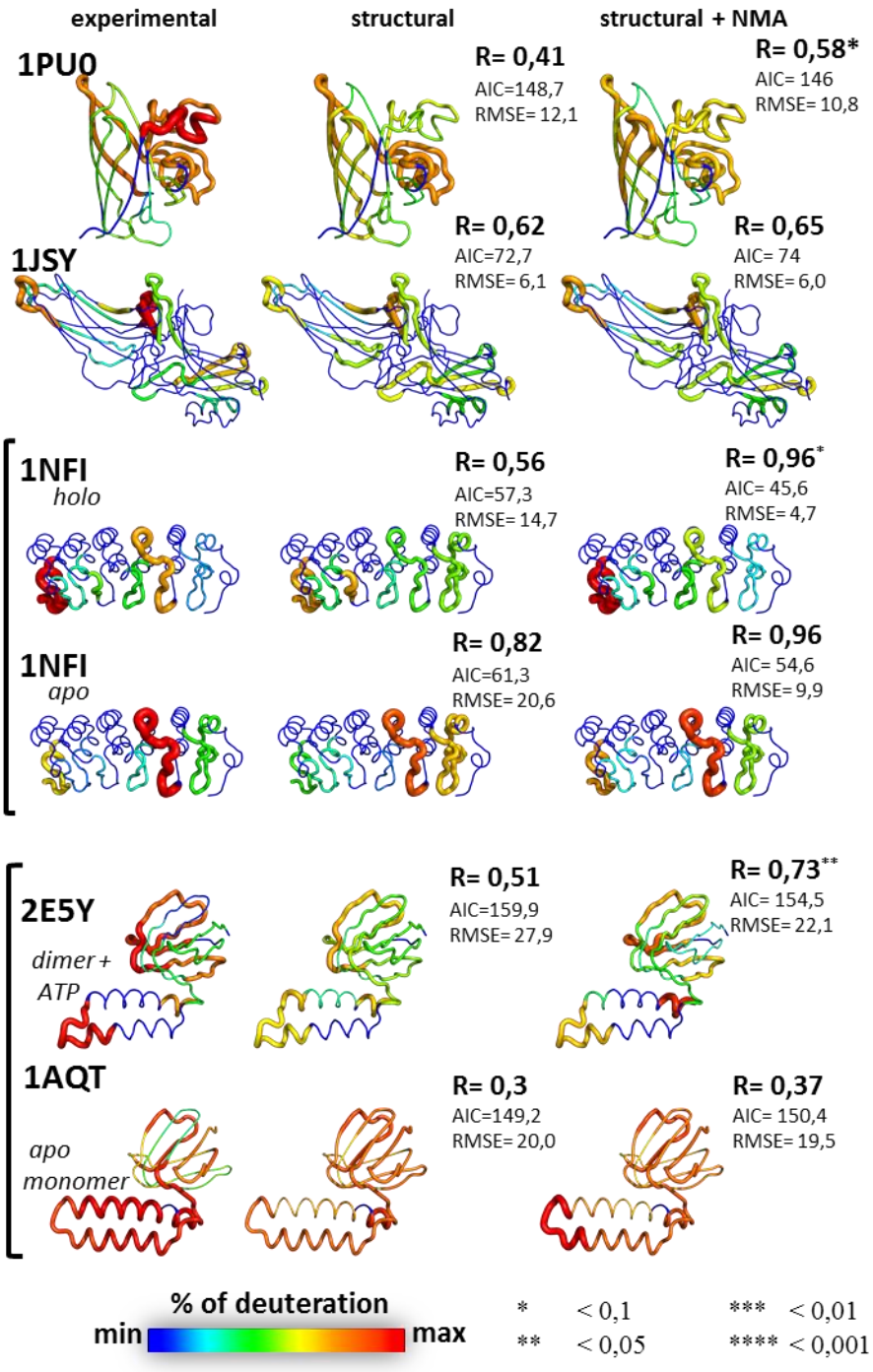


Figura 19 - Representação dos valores ajustados e experimentais nas estruturas das proteínas

Representação das porcentagens de troca, e valores ajustados dos modelos Estrutural e Estrutural + NMA de cada proteína representados nas estruturas de cada uma das proteínas em uma escala de cores (vermelho representa valores maiores, e azul, valores menores, sendo os valores mínimos da escala iguais a zero ou representando ausência de dados experimentais).



Continuação da Figura 19.

4.2.3. Modelo único ajustado a todas as proteínas

Com o intuito de construir um modelo único ajustado para todas as proteínas, foram criados 3 modelos ajustados simultaneamente aos dados de todas as observações de todas as proteínas. Foram considerados os modelos *Estrutural*, *Estrutural + NMA* e um terceiro modelo, em que foi adicionado ao modelo *Estrutural + NMA* e uma variável γ para cada proteína, resultando em um ajuste individual (Figura 20). O modelo *Estrutural* mostra baixas correlações com os dados experimentais (R=0,4), mostrando um aumento significativo quando a variável NMA é adicionada (R=0,54), o que ressalta a importância de informações sobre a dinâmica no modelo geral, porém, as correlações observadas para esse modelo ajustado a todas as proteínas, ainda não se equiparam às correlações observadas nos modelos individuais de cada proteína- como mostrado nas etapas anteriores. Tendo isso em vista, foi adicionado uma variável γ identificando cada proteína no modelo *Estrutural + NMA,* e correlações semelhantes aos casos individuais foram obtidas (R=0,73), nota-se também a diminuição dos valores de RMSE ao adicionar as a variável NMA e posteriormente ao adicionar a variável γ, é possível que a introdução desta tenha vindo a ajustar o intercepto para cada proteína compensando as diferenças de amplitude dos dados de HX, uma vez que as escalas de porcentagem de troca variam não apenas em função de fatores relacionados a estrutura e dinâmica, mas também em função das condições experimentais – que por sua vez, são diferentes em cada um dos sistemas.

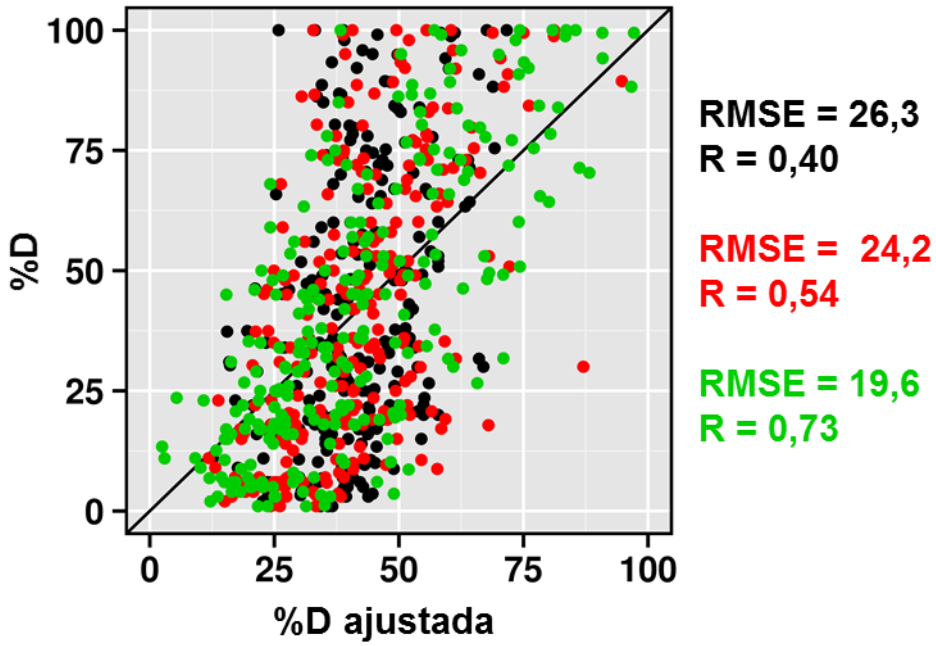


Figura 20 - Modelos ajustados para todo o *dataset*.

Ajuste de três modelos – *Estrutural* (preto), *Estrutural* + *NMA* (vermelho), *Estrutural* + *NMA* com a utilização da variável γ (verde). Os valores de correlação estão indicados na parte inferior do gráfico.

4.2.4. Apo x holo

Ao observar o modelo *Estrutural* das formas *apo* e *holo* do IκBα (1NFI), é possível notar que a forma *holo* possui baixas correlações com os dados experimentais (R=0,56) quando comparada com a forma *apo* (R=0,75) (Figura 19). Esse fato pode ser justificado pelas restrições de movimento impostas pelos ligantes presentes em cada uma das proteínas, restrição que veio a ser representada no modelo através da introdução das flutuações de NMA, uma vez que o modelo *Estrutural* + *NMA* apresentou fortes correlações em ambos os sistemas (R=0,96). Chama atenção o fato de que a NMA foi mais efetiva para o sistema 1NFI do que o BFAC ou ENM, possivelmente pelo fato desses últimos não representarem as restrições conformacionais impostas pelo ligante. Os sistemas sob os códigos 2NSX e 2NTI (beta-glucosidase *apo* e *holo*) apresentaram correlações semelhantes quando utilizando o modelo *Estrutural,* no entanto, a adição tanto das flutuações obtidas por NMA quanto por ENM resulta em diferentes valores de correlação entre os dados ajustados e experimentais, mostrando que os sistemas diferem quanto à dinâmica,

possivelmente por diferenças sutis na estrutura inicial considerada no cálculo, uma vez que a ENM também foi capaz de representar a diferença nas flutuações.

No caso do sistema 1AQT (Figura 19) - um dímero com uma molécula de ATP ligada - as correlações foram baixas em todos os modelos, ao contrário do sistema 2E5Y; um monômero da mesma proteína presente no sistema 1AQT. A proteína de código 2E5Y apresentou fortes correlações entre os valores ajustados e os dados experimentais nos modelos *Estrutural + NMA, Estrutural + ENM e Estrutura + BFAC.* É possível que essa discrepância entre as duas formas tenha sido gerada por determinantes de troca do sistema 1AQT que não foram incluídos nos modelos.

4.3. Modelagem de MS-HX em resolução de resíduo.

Após serem criados os modelos para as proteínas que possuem dados em resolução de peptídeo, foi estudado o caso da proteína SNase, que possui dados de MS-HX em resolução de resíduo.

Foram criados os modelos *Estrutural*, *Estrutural + NMA*, *Estrutural + ENM* e *Estrutural + BFAC* (Figura 21). No caso da SNase, a utilização do modelo *Estrutural* já foi suficiente para descrever os dados experimentais com fortes correlações (R=0,73) e baixos valores de RMSE, no entanto, a adição das variáveis dinâmicas neste modelo aumenta as correlações e diminui o AIC, contudo, novamente a utilização da NMA maiores correlações (R=0,78) e apresenta maiores reduções nos valores de AIC e RMSE, além disso apresenta também a maior significância quanto às diferenças quando se compara o modelo *Estrutura + dinâmico* com o modelo *Estrutural.*

É interessante notar a diferença na melhora dos modelos ao adicionar as variáveis dinâmicas nos casos de MS-HX em resolução de peptídeo e em resolução de resíduo. É possível que ao adicionar as variáveis dinâmicas no modelo criado para explicar os dados a nível de peptídeo, essas tenham sido mais representativas pelo fato de as flutuações de NMA e ENM, assim como os fatores B terem sido medidas médias dos peptídeos e não medidas de individuais de cada resíduo. Esta observação indica que talvez as flutuações de grandes segmentos sejam mais influentes na modelagem dos dados de troca do que a flutuação individual de cada resíduo.

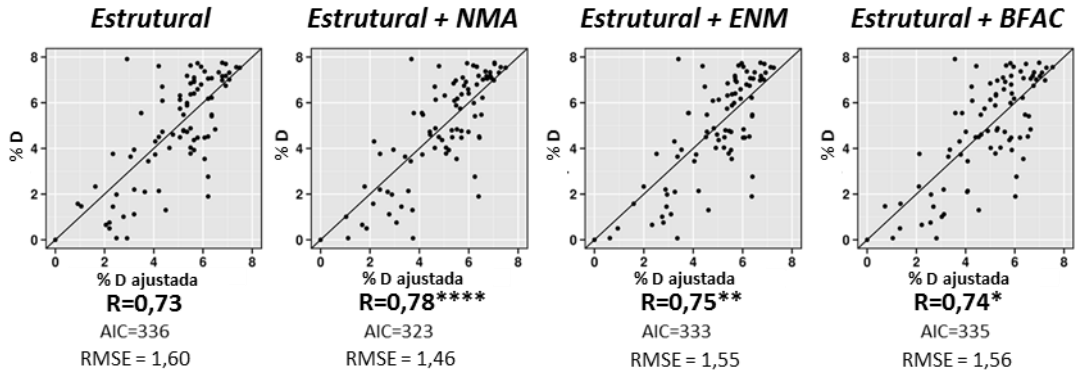


Figura 21 - Modelagem dos dados da proteína SNase.
Modelos ajustados aos dados da proteína SNase, utilizando apenas parâmetros estruturais ou utilizando parâmetros estruturais junto aos parâmetros dinâmicos, são mostrados os coeficientes de correlação de Pearson entre os valores ajustados e os dados experimentais, assim como o AIC e o RMSE de cada modelo, os asteriscos indicam a significância da diferença entre cada um dos modelos que utilizam variáveis estruturais e dinâmicas e o modelo que utiliza apenas variáveis estruturais.

Assim como foi feito com os dados a nível de resíduo, os dados da SNase também foram representados nas estruturas das proteínas para fins de comparação visual (Figura 22), sendo possível notar a clara semelhança entre os valores ajustados e os dados experimentais.

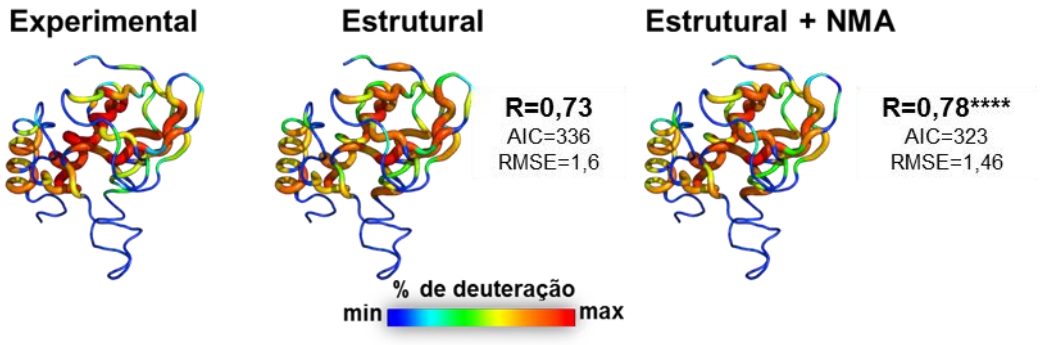


Figura 22 - Representação dos dados experimentais e teóricos da SNase em sua estrutura
Dados experimentais e dos modelos Estrutural e Estrutural + NMA representados na estrutura da proteína SNase utilizando uma escala de cores. Os valores dos coeficientes de correlação de Pearson, AIC e RMSE são mostrados ao lado da estrutural, os asteriscos indicam diferença significativa entre o modelo *Estrutural* e o modelo *Estrutural + NMA*.

CAPÍTULO V. MODELAGEM DOS DADOS DE NMR-HX

Neste capítulo serão tratados os resultados relacionados à modelagem estatística dos dados de HX obtidos pela técnica de NMR. Os dados experimentais (expressos em logPF) são disponíveis para cada resíduo da proteína, permitindo não somente a criação de modelos preditivos, mas possibilitando uma análise mais robusta que levou a proposição de dois novos modelos preditivos, que serão discutidos em detalhes. Um artigo com a proposição e análises destes modelos preditivos está em fase de elaboração.

5. RESULTADOS E DISCUSSÃO

5.1 Dataset

O *dataset* de NMR-HX foi construído com 14 proteínas, como pode ser visto na Tabela 4, a tabela mostra também informações como código no PDB, nome, função, número de resíduos da proteína, condições experimentais e as respectivas referências dos experimentos de HX. A análise de alinhamento múltiplo utilizando o servidor MUSCLE mostra que com exceção das duas lisozimas presentes no *dataset* (2LZT e 2EQL – que possuem identidade de 50%), as identidades entre as proteínas não passam de 38,6%, sendo ainda menos semelhantes entre si do que as proteínas contidas no *dataset* de MS-HX.

Tabela 4 - *Dataset* de proteínas para modelagem de NMR-HX

PDB ID	Nome	Função	# de res.	Condições Experimentais	Ref.
1EY0 (137)	SNase	hidrolase	149	pH: 5.5; T= 310,15 K	(150)
1MBC (151)	Mioglobina	transporte	153	pH:3,5; T= 308,15	(152)
1UBQ (153)	Ubiquitina	cromossômica	76	pH: 3,5; T=295,16 K	(154)
1A4V (155)	α-Lactoalbumina	sintetase	123	pH: 6,3; T= 289,15 K	(156)
5PTI (157)	BPTI	inibidor de protease	58	pH: 3,5; T= 303,15 K	(158)
1G68 (159)	PSE-4 Carbenicilinase	hidrolase	271	pH: 6,6; T= 304 K	(160)
2EQL (161)	Lisozima Equina	hidrolase	129	pH: 4,5; T= 298,15 K	(162)
2LZT (163)	Lisozima – *Gallus gallus*	hidrolase	129	pH: 7,5; T= 303,15 K	(164)
1BNR (165)	Barnase	ribonuclease	110	pH: 6,8; T= 310,15	(166)
1FCL (167)	proteína G estreptocócica	Ligação	56	pH: 5,3; T= 298,15 K	(168)
1LUD (169)	Diidrofolato redutase	Oxidorredutase	162	pH: 6,5; T= 288 K	(170)
2L52 (171)	SAMP1	Ligação	99	pH: 6,8; T= 298 K	(171)
1OZI (172)	Domínio PDZ2 da PTB-BL	Hidrolase	99	pH: 3,5; T=281,5 K	(172)
1MZK (173)	Domínio FAH de interação com cinase	hidrolase	139	pH: 6,3; T=298 K	(174)

As colunas mostram respectivamente: O código do PDB de cada proteína e a respectiva referência, o nome da proteína, a classificação funcional segundo o PDB, o número de resíduos da estrutura, as condições experimentais do experimento de HX e por último as referências bibliográficas de onde foram obtidos os dados experimentais de HX.

5.2. Critérios para o cálculo do N_c

Para realizar a otimização dos modelos estatísticos para a predição de NMR-HX, foram investigados critérios ótimos para o cálculo da variável N_c. Para tal, foi utilizado o modelo *Estrutural + NMA*, ou seja, aquele que apresentou os melhores resultados para os dados de MS-HX. Os *heatmaps* da Figura 23 mostram os 4 critérios utilizados para o cálculo do N_c e as variações das correlações entre os valores ajustados de cada proteína e os dados de HX em função da variação dos valores de R_c. Analisando os dados das correlações em cada um dos critérios, foi possível concluir que os que geraram as maiores correlações para a maioria das proteínas foram: considerando todos os átomos

do sistema, contabilizando o total de átomos em contato e adotando um raio de corte de 8.5 Å. É possível que ao considerar todos os átomos do sistema no cálculo do N_c, o grau de exposição de um determinado resíduo seja representado de forma mais acurada, diferente dos modelos simplificados que muitas vezes são utilizados para representar proteínas apenas pelos átomos da cadeia principal, desta forma parece razoável que a consideração de todos os átomos em contato num modelo que considera também as cadeias laterais tenha apresentado as maiores correlações com os dados experimentais. Os critérios aqui adotados para o cálculo dos contatos diferem de outros modelos já publicados. Chama atenção a diferença dos raios de corte considerados em algumas das tentativas de modelar os dados de HX na literatura, uma vez que os critérios normalmente diferem entre si (62, 104).

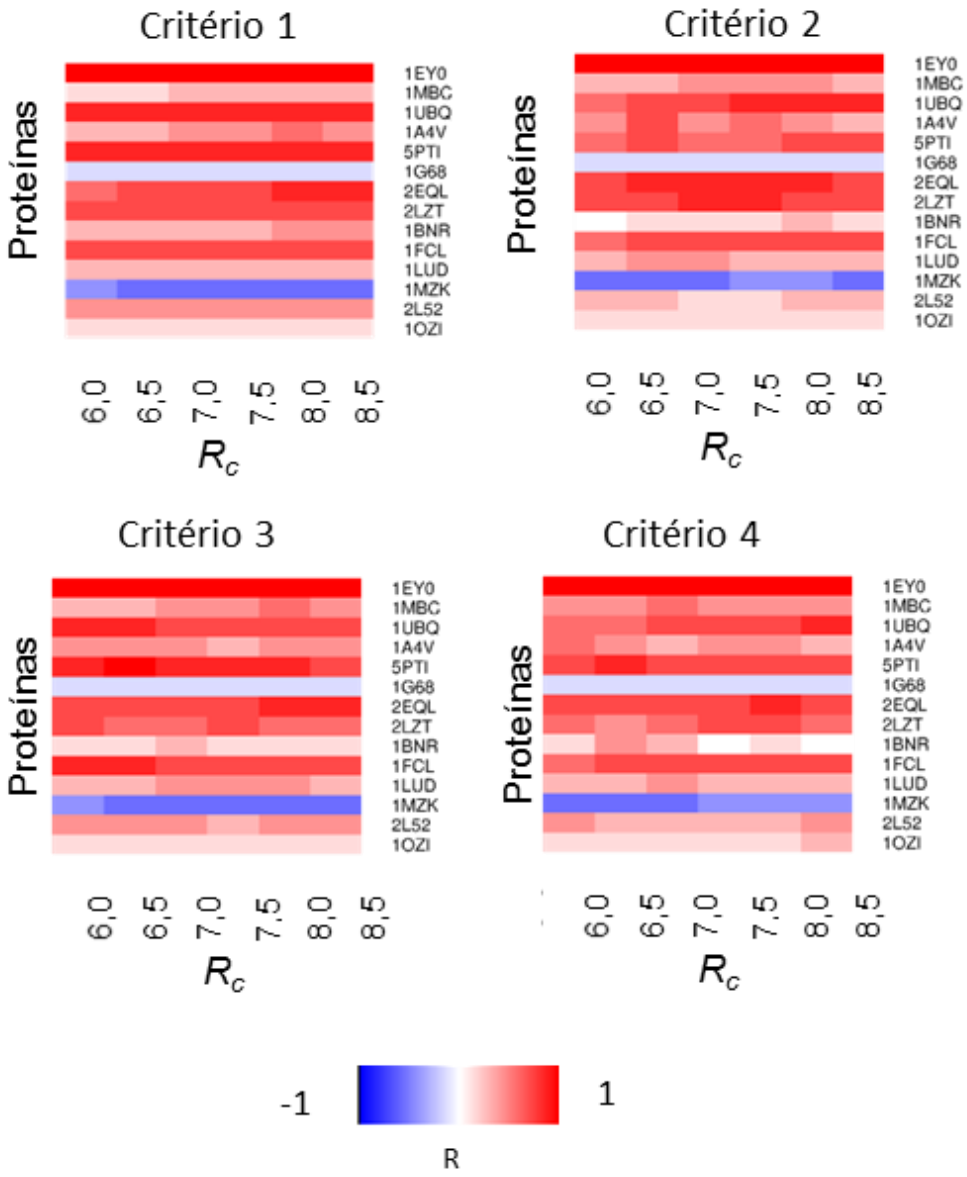


Figura 23 – *Influência do R_c* para o cálculo do N_c
Heatmaps representando as correlações entre os valores ajustados do modelo *Estrutural+NMA* e os dados experimentais empregando cada um dos critérios geométricos para o cálculo de N_c. Critério 1: Todos os átomos da proteína, contabilização de todos os átomos em contato; Critério 2: Todos os átomos da proteína, contabilização dos resíduos em contato; Critério 3: Apenas átomos da cadeia principal, contabilização de todos os átomos em contato; Critério 4: Apenas átomos da cadeia principal, contabilização dos resíduos em contato.

Tabela 5 - Modelos testados para o *dataset* reduzido

Modelo	Equação
contatos	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + e_i$
hbond	$\mathrm{HX}_i = \beta_0 + \beta_{hb} N_{hb_i} + e_i$
Estrutural	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{hb} N_{hb_i} + e_i$
SASA	$\mathrm{HX}_i = \beta_0 + \beta_{SASA} SASA + e_i$
Estrutural + NMA	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{N_{HB}} N_{HB_i} + \beta_{NMA} RMSF_{NMA_i} + \beta_G G + \beta_H H + \beta_I I + \beta_E E + \beta_S S + \beta_T T + \beta_c C + e_i$
ES (estrutura secundária)	$\mathrm{HX}_i = \beta_0 + \beta_G G + \beta_H H + \beta_I I + \beta_E E + \beta_S S + \beta_T T + \beta_c C + e_i$
Estrutural + ES +NMA	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{N_{HB}} N_{HB_i} + \beta_{NMA} RMSF_{NMA_i} + \beta_G G + \beta_H H + \beta_I I + \beta_E E + \beta_S S + \beta_T T + \beta_c C + e_i$
Estrutural + ES + NMA + pH	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{N_{HB}} N_{HB_i} + \beta_{NMA} RMSF_{NMA_i} + \beta_G G + \beta_H H + \beta_I I + \beta_E E + \beta_S S + \beta_T T + \beta_c C + \beta_{pH} pH + e_i$
Estrutural + ES + NMA + pH + temperatura	$\mathrm{HX}_i = \beta_0 + \beta_C N_{C_i} + \beta_{N_{HB}} N_{HB_i} + \beta_{NMA} RMSF_{NMA_i} + \beta_G G + \beta_H H + \beta_I I + \beta_E E + \beta_S S + \beta_T T + \beta_c C + \beta_{pH} pH + \beta_{temp} Temp + e_i$

São representados os modelos e suas respectivas estruturas: intercepto, coeficientes e variáveis associadas e erro. Onde G (Hélice-3), H (α-hélice), I (Hélice-5), E (Folha β), S (*Bends*), T (*Turns*) e C (*Random coil*) representam respectivamente os elementos de estrutura secundária preditos com o algoritmo DSSP.

Ao utilizar as variáveis individualmente ficou claro que todas as correlações entre os valores ajustados e os dados experimentais foram fracas, e mesmo quando foi utilizado o modelo *Estrutural* as correlações chegaram apenas a 0,47 (Figura 24).

Chama atenção o fato de que ao utilizar apenas a variável SASA como preditor existe um valor de acessibilidade a partir do qual o modelo não é mais capaz de realizar o ajuste correto, a partir desse ponto o modelo responde com valores ajustados muito próximos uns dos outros para diferentes valores da variável utilizada. No modelo *ES* – onde foi utilizada apenas a estrutura

secundária - é possível verificar grupos de observações referentes a cada uma das classes de estruturas secundárias em que os resíduos são classificados pelo *dssp*. É importante ressaltar também o comportamento do modelo *hbond*, este apresenta apenas dois valores possíveis, visto que N_{HB} é binário (presença ou não de ligação hidrogênio).

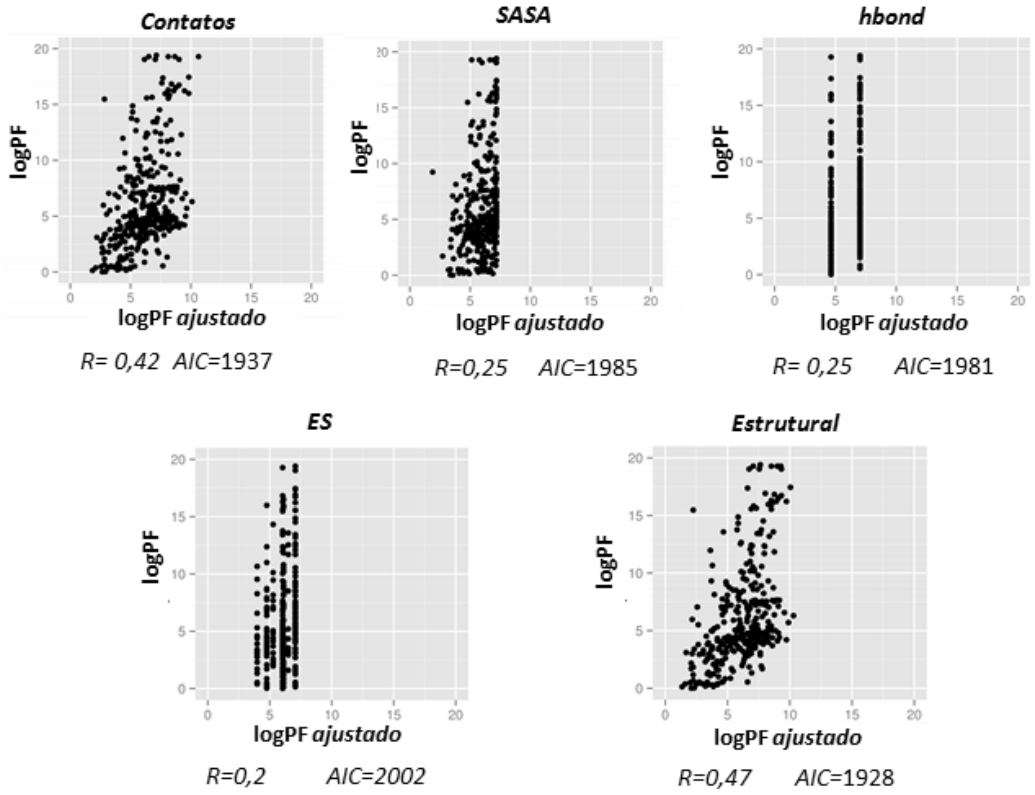


Figura 24 - Modelos ajustados aos dados de NMR-HX
Análise dos modelos *contatos*, SASA, $N_{hb,}$ ES, *Estrutural*. São mostradas as correlações entre os dados ajustados e dados experimentais, assim como o AIC.

Em seguida, foi construído o modelo *Estrutural* + *NMA*, onde as correlações também não aumentaram de forma significativa. É notável o fato de que a adição de NMA nos modelos criados para explicar os dados de MS resultou em um aumento na correlação muito maior do que nos modelos criados para os dados de NMR.

Ao adicionar as variáveis estrutura secundária e pH, as correlações para todo o *dataset* não aumentaram de forma significativa, porém foi com a adição da temperatura que a correlação chegou a aproximadamente 0,8. A adição do pH e temperatura também levam a uma diminuição do AIC, como pode ser observado na Figura 25, deixando evidente a importância da utilização dos

dados experimentais nesse tipo de modelo. É importante ressaltar que pH e temperatura não foram utilizados em nenhum dos modelos anteriores encontrados na literatura, e nos dados aqui mostrados eles parecem operar de forma significativa nos modelos, embora os valores de ambas as variáveis em cada um dos experimentos não sejam muito discrepantes.

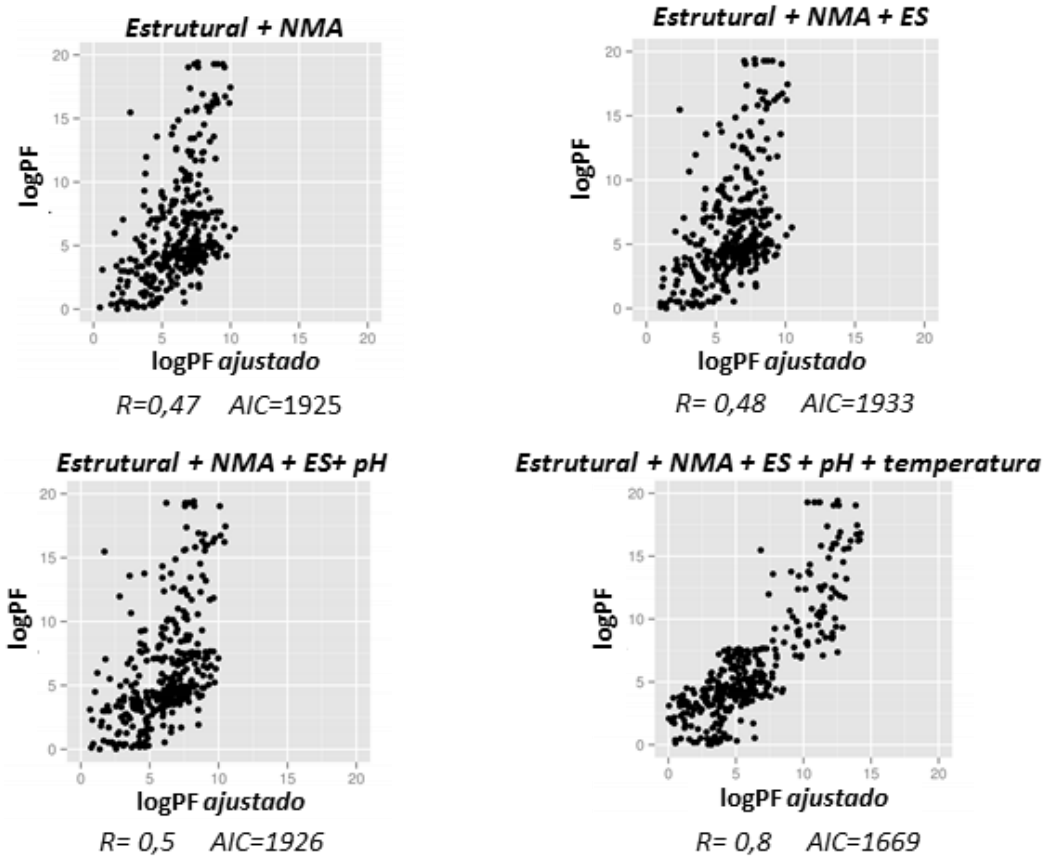


Figura 25 - Modelos ajustados aos dados de NMR-HX
Análise do modelo *Estrutural + NMA* e suas variações com adição da variável estrutura secundária, pH e temperatura, são mostradas as correlações entre os dados ajustados e dados experimentais, assim como o AIC.

Em seguida um modelo utilizando RF (empregando o modelo *Estrutural + NMA + ES + pH + temperatura*) (Figura 26), foi ajustado ao *dataset* de 6 proteínas, e mostrou correlações maiores do que o modelo linear mostrado na Figura 25, que utilizou as mesmas variáveis. O melhor desempenho do RF em relação ao modelo linear pode ser devido a relações não lineares entre as variáveis, que podem ser exploradas através das árvores de regressão.

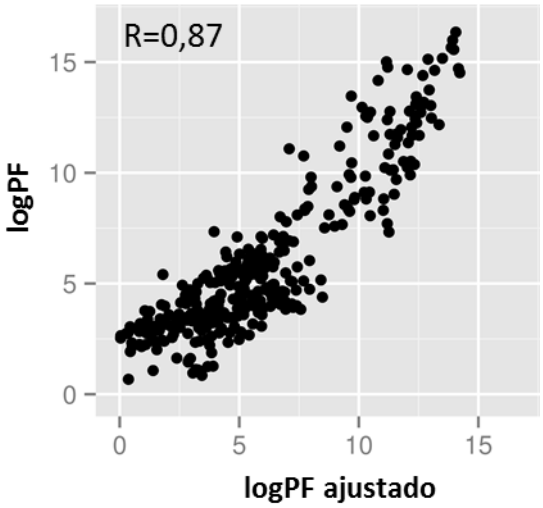


Figura 26 – Modelo de *Random Forest* treinado com o dataset reduzido
Correlação entre os valores ajustados de um modelo de RF e os dados experimentais para o *dataset* reduzido.

5.3. Modelos preditivos e validação cruzada

Em seguida, os modelos de RF e modelos lineares (ambos utilizando *Estrutural + NMA + ES + pH + temperatura*) foram empregados para realizar predições para as proteínas do *dataset* reduzido. A avaliação dos modelos foi realizada através de validação cruzada por *leave-one-out.* Para fins de comparação visual, os valores experimentais e preditos foram representados nas estruturas (Figura 27), os valores de RMSE e as correlações foram avaliados para cada predição utilizando cada um dos dois modelos e também constam na Figura 27.

É possível notar que para a maioria das proteínas o método de RF foi mais efetivo nas predições, alcançando tanto erros menores quanto correlações maiores. Ao observar a semelhança entre os dados preditos tanto por RF quanto pelo modelo linear, é possível notar que diversas regiões apenas foram preditas corretamente com o método de RF. Além disso, as correlações entre os dados de NMR-HX e os dados preditos são maiores ao usar RF na maioria dos casos, assim como os valores de RMSE para a maioria das proteínas são menores ao usar RF, levantando a hipótese de que as relações entre as variáveis e os dados experimentais podem ser não lineares.

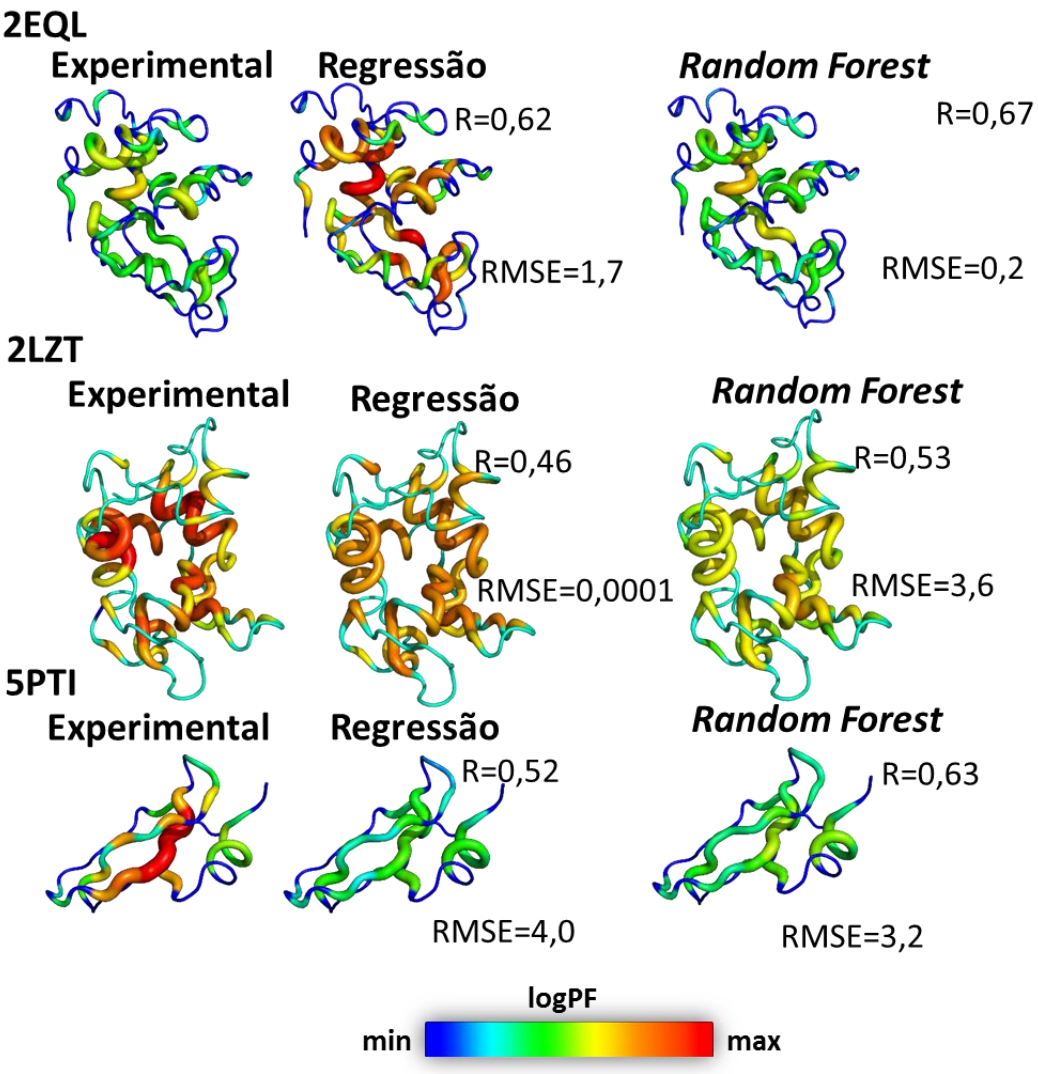
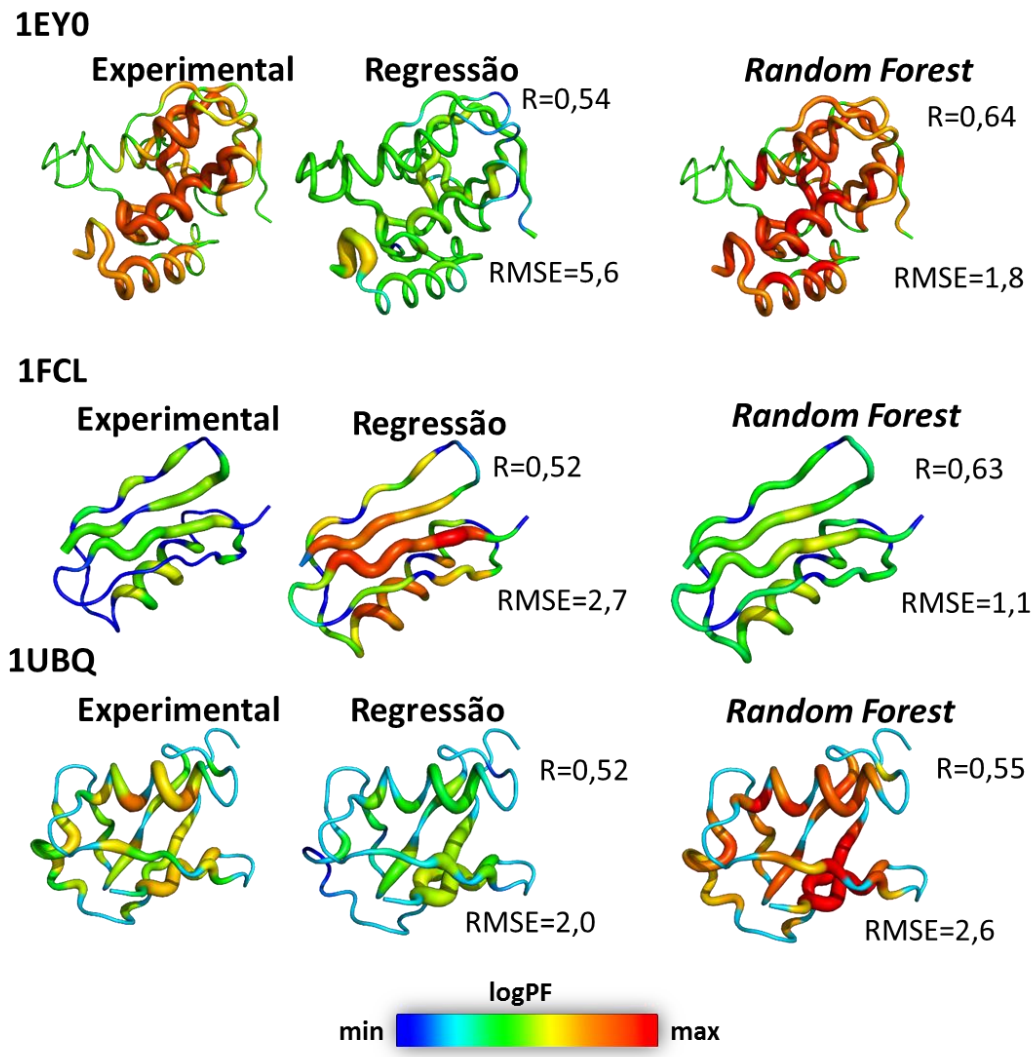


Figura 27 - Dados preditos e experimentais representados nas estruturas das proteínas
São mostrados os dados preditos na validação cruzada utilizando modelos de RF e LM, também são mostrados os dados experimentais para fins de comparação. As correlações e RMSE de cada modelo também são mostrados.



Continuação da Figura 27

É possível que embora o *dataset* apresente proteínas com valores de identidade iguais ou menores que 50 %, não possua diversidade suficiente que os modelos expliquem os dados de HX de alguns casos específicos. Utilizando *datasets* de 2000 estruturas e a abordagem de redes neurais, Tartaglia *et al.* obtiveram correlações entre 0,5-0,7, o que nos leva a crer que um aumento do *dataset* aqui utilizado possa levar a predições ainda melhores utilizando tanto modelos lineares quanto RF, visto que o conjunto de proteínas utilizado aqui foi pequeno e heterogêneo.

É importante ressaltar aqui que embora tenhamos utilizado as flutuações de obtidas a partir de NMA nos modelos tanto de NMR-HX quanto MS-HX, as variações no número de contatos e ligações hidrogênio que são causadas pelos

movimentos dos domínios não foram levadas em consideração, ou seja, os modelos aqui construídos são baseados em estruturas únicas, o que leva à hipótese de que a utilização de múltiplas estruturas obtidas por métodos de amostragem melhorada ou por NMR possa vir a melhorar a qualidade das predições.

6. CONCLUSÕES E PERSPECTIVAS

A dificuldade na predição de dados de HX tem sido um problema recorrente na literatura desde a década de 50 quando a técnica foi desenvolvida. Nos dias de hoje, mesmo com todo a evolução dos computadores para o estudo de estrutura e dinâmica de proteínas, as conclusões sobre o papel de cada fator determinantes de HX continuam sendo pouco sólidas. Neste trabalho foram apresentados métodos que empregam tanto dados estruturais quando dados que representam a flexibilidade da proteína calculada por métodos aproximados como NMA, ENM e fator B.

É possível observar que o chamado “modelo fenomenológico” na literatura, onde a troca é determinada pelo número de contatos e ligações hidrogênio, mostrou correlações fracas com os dados experimentais tanto nos *datasets* de MS-HX (com algumas exceções) quanto no de NMR-HX, sendo as correlações bem mais fracas neste último. Mostrando que, apenas a informação estrutural contida nos cristais não é suficiente para descrever o dado de HX, visto que o mesmo é dependente das diversas conformações exploradas pelas proteínas, assim como afirmado por Vendruscolo *et al.* (104).

É possível concluir também que a adição de preditores que representem a flexibilidade no modelo estrutural pode explicar grandes partes da troca ocorrida em uma proteína – como pode ser observado na maioria dos casos de MS-HX em que se aplica o modelo *Estrutural + NMA*. Outro resultado importante é a superioridade do modelo RF sobre o modelo linear na modelagem de NMR-HX, levantando à hipótese de que modelos não lineares possam explicar melhor o fenômeno de troca.

Como dito anteriormente, os modelos que incluem NMA não incorporam as variações no número de contatos e ligações hidrogênio que os movimentos dos domínios acarretam, o que mostra que o uso da estrutura em forma de uma “fotografia” embora tenha a informação de quão flexível é cada parte, não representa completamente as variações estruturais do sistema. Como solução para este problema, uma das perspectivas seria a aplicação de um novo método desenvolvido pelo nosso grupo, o MDeNM (175), dinâmica molecular excitada por modos normais. Esta metodologia consiste em uma simulação de dinâmica molecular, onde as velocidades iniciais são atribuídas obedecendo combinações

lineares de vetores de modos normais. Assim é possível acelerar a amostragem de movimentos de domínios em simulações de MD. É possível que modelos calibrados por *ensembles* contendo apenas os parâmetros estruturais, contenham informação suficiente para descrever a flexibilidade da proteína e consequentemente, os dados de troca.

Quanto às predições, é possível que o aumento do *dataset* permita melhor treinamento do modelo de RF, uma vez que já foi demonstrado por Tartaglia *et al* que com um *dataset* de 2000 proteínas e treinando uma rede neural para ajustar o modelo Estrutural aos dados das proteínas foi possível realizar predições com correlações entre 0,5 e 0,7. É relevante também que estudos futuros visem criar grupos de proteínas com elementos representativos de diversas categorias estruturais de proteínas.

O grupo também visa realizar estudos utilizando modelos que levem em consideração as vizinhanças de cada resíduo, para que estes não sejam tratados de forma independente na estrutura do modelo, já que existe uma óbvia dependência entre os resíduos e uma clara relação entre os dados de HX de resíduos próximos, assim como também normalmente existem correlações entre os movimentos destes vizinhos.

O presente trabalho esclareceu o papel das flutuações calculadas a partir da NMA na predição do fenômeno de HX, partindo da hipótese levantada pela primeira vez por Bahar *et al.* (115) que propôs qualitativamente uma relação entre as flutuações provenientes de uma modelo de redes Gaussianas e o PF. Aqui as flutuações de NMA foram utilizadas em conjunto com o modelo fenomenológico proposto na literatura para descrever a troca do hidrogênio, demonstrando sua utilidade ao explicar dados de MS ou predizer dados de NMR. Também foram mostradas as importâncias de outros fatores estruturais e pela primeira vez as condições experimentais foram inseridas no modelo.

Embora diversos fatores estruturais tenham sido explorados aqui, outros determinantes também são discutidos na literatura – como efeitos eletrostáticos e ligações hidrogênio com o próprio solvente -, a exploração sistemática destes fatores em estudos futuros deve ser realizada para desvendar as bases do fenômeno de HX, para melhor interpretação de resultados e melhor aplicação

dos mesmos em estudos computacionais que se utilizem dos dados experimentais como ponto de partida.

7. REFERÊNCIAS

1. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, et al. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. Nature. 1960;185(4711):422-7.
2. Williamson MP, Havel TF, Wüthrich K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1 H nuclear magnetic resonance and distance geometry. Journal of molecular biology. 1985;182(2):295-315.
3. Wuthrich K. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. Science. 1989;243(4887):45-50.
4. Whitford D. Proteins: structure and function: John Wiley & Sons; 2013.
5. Linderstrøm-Lang KU. Lane Medical Lectures: proteins and enzymes: Stanford University Press; 1952.
6. Branden CI. Introduction to protein structure: Garland Science; 1999.
7. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A. 1951;37(4):205-11.
8. Ponting CP, Russell RR. The natural history of protein domains. Annual review of biophysics and biomolecular structure. 2002;31(1):45-71.
9. Health BUSoP. [cited 2015 23 de Outubro]. Available from: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH/PH709_BasicCellBiology/PH709_BasicCellBiology26.html

10. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. Journal of molecular biology. 2001;307(4):1113-43.

11. Berendsen HJ, Hayward S. Collective protein dynamics in relation to function. Current opinion in structural biology. 2000;10(2):165-9.

12. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data Bank. Acta crystallographica Section D, Biological crystallography. 2002;58(Pt 6 No 1):899-907.
13. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402:C47-C52.
14. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. Journal of molecular biology. 1999;288(1):147-64.
15. Friedrich W, Knipping P, Laue M. Interferenzerscheinungen bei roentgenstrahlen. Annalen der Physik. 1913;346(10):971-88.
16. Watson JD, Crick FH. Molecular structure of nucleic acids. Nature. 1953;171(4356):737-8.
17. Jones TA, Zou J-Y, Cowan St, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallographica Section A: Foundations of Crystallography. 1991;47(2):110-9.
18. Lattman EE, Loll PJ. Protein crystallography: a concise guide: JHU Press; 2008.
19. Sakabe N. X-ray diffraction data collection system for modern protein crystallography with a Weissenberg camera and an imaging plate using synchrotron

radiation. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 1991;303(3):448-63.
20. McRee DE. Practical protein crystallography: Academic press; 1999.
21. Dale GE, Oefner C, D'Arcy A. The protein as a variable in protein crystallization. Journal of structural biology. 2003;142(1):88-97.
22. Hinsen K. Structural flexibility in proteins: impact of the crystal environment. Bioinformatics. 2008;24(4):521-8.
23. Van Gunsteren W, Karplus M. Effect of constraints, solvent and crystal environment on protein dynamics. 1981.
24. Sousa R. Use of glycerol, polyols and other protein structure stabilizing agents in protein crystallization. Acta Crystallographica Section D: Biological Crystallography. 1995;51(3):271-7.
25. Halle B. Flexibility and packing in proteins. Proceedings of the National Academy of Sciences. 2002;99(3):1274-9.
26. Rabi I, Millman S, Kusch P, Zacharias J. The Magnetic Moments of Li 6 3, Li 7 3 and F 19 9. Physical Review. 1938;53(6):495.
27. Kellogg J, Rabi I, Ramsey Jr N, Zacharias J. The Magnetic Moments of the Proton and the Deuteron. The Radiofrequency Spectrum of H 2 in Various Magnetic Fields. Physical Review. 1939;56(8):728.
28. Rabi I, Zacharias J, Millman S, Kusch P. Milestones in magnetic resonance:'a new method of measuring nuclear magnetic moment'. 1938. Journal of magnetic resonance imaging: JMRI. 1991;2(2):131-3.
29. Wüthrich K. The way to NMR structures of proteins. Nature Structural & Molecular Biology. 2001;8(11):923-5.
30. Wüthrich K. Protein structure determination in solution by NMR spectroscopy. Journal of Biological Chemistry. 1990;265(36):22059-62.
31. Rance M, Sørensen O, Bodenhausen G, Wagner G, Ernst R, Wüthrich K. Improved spectral resolution in COSY 1H NMR spectra of proteins via double quantum filtering. Biochemical and biophysical research communications. 1983;117(2):479-85.
32. Aue W, Bartholdi E, Ernst RR. Two-dimensional spectroscopy. Application to nuclear magnetic resonance. The Journal of Chemical Physics. 1976;64(5):2229-46.
33. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. Journal of magnetic resonance. 2003;160(1):65-73.
34. Güntert P. Automated NMR structure calculation with CYANA. Protein NMR Techniques. 2004:353-78.
35. Vriend G. WHAT IF: a molecular modeling and drug design program. Journal of molecular graphics. 1990;8(1):52-6.
36. Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. Journal of biomolecular NMR. 1996;8(4):477-86.
37. Marion D, Driscoll PC, Kay LE, Wingfield PT, Bax A, Gronenborn AM, et al. Overcoming the overlap problem in the assignment of proton NMR spectra of larger proteins by use of three-dimensional heteronuclear proton-nitrogen-15 Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1. beta. Biochemistry. 1989;28(15):6150-6.
38. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, et al. NMR structure determination for larger proteins using backbone-only data. Science. 2010;327(5968):1014-8.

39. Al-Amoudi A, Norlen LP, Dubochet J. Cryo-electron microscopy of vitreous sections of native biological cells and tissues. Journal of structural biology. 2004;148(1):131-5.
40. Watson ML. Staining of tissue sections for electron microscopy with heavy metals. The Journal of Biophysical and Biochemical Cytology. 1958;4(4):475-8.
41. Brenner S, Horne R. A negative staining method for high resolution electron microscopy of viruses. Biochimica et biophysica acta. 1959;34:103-10.
42. Frank J. Single-particle imaging of macromolecules by cryo-electron microscopy. Annual review of biophysics and biomolecular structure. 2002;31(1):303-19.
43. Carpenter EP, Beis K, Cameron AD, Iwata S. Overcoming the challenges of membrane protein crystallography. Current opinion in structural biology. 2008;18(5):581-6.
44. Milne JL, Borgnia MJ, Bartesaghi A, Tran EE, Earl LA, Schauder DM, et al. Cryo-electron microscopy–a primer for the non-microscopist. FEBS Journal. 2013;280(1):28-45.
45. Anfinsen CB, Haber E, Sela M, White Jr F. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proceedings of the National Academy of Sciences of the United States of America. 1961;47(9):1309.
46. Hartl FU, Hayer-Hartl M. Molecular chaperones in the cytosol: from nascent chain to folded protein. Science. 2002;295(5561):1852-8.
47. Levinthal C. How to fold graciously. Mossbauer spectroscopy in biological systems. 1969:22-4.
48. Socci N, Onuchic JN, Wolynes PG. Diffusive dynamics of the reaction coordinate for protein folding funnels. The Journal of chemical physics. 1996;104(15):5860-8.
49. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins: Structure, Function, and Bioinformatics. 1995;21(3):167-95.
50. Freddolino PL, Liu F, Gruebele M, Schulten K. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. Biophysical journal. 2008;94(10):L75-L7.
51. Takano M, Yamato T, Higo J, Suyama A, Nagayama K. Molecular dynamics of a 15-residue poly (L-alanine) in water: helix formation and energetics. Journal of the American Chemical Society. 1999;121(4):605-12.
52. Levitt M, Warshel A. Computer simulation of protein folding. Nature. 1975;253(5494):694-8.
53. Unger R, Moult J. Genetic algorithms for protein folding simulations. Journal of molecular biology. 1993;231(1):75-81.
54. Sánchez R, Šali A. Comparative protein structure modeling: introduction and practical examples with modeller. Protein Structure Prediction: Springer; 2000. p. 97-129.
55. Ginalski K. Comparative modeling for protein structure prediction. Current opinion in structural biology. 2006;16(2):172-7.
56. Smith TF, LO CONTE L, BIENKOWSKA J, Gaitatzes C, ROGERS Jr RG, LATHROP R. Current limitations to protein threading approaches. Journal of Computational Biology. 1997;4(3):217-25.
57. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins: Structure, Function, and Bioinformatics. 1999;35(4):408-14.

58. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annual review of biophysics and biomolecular structure. 2001;30(1):173-89.
59. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Structure, Function, and Bioinformatics. 2012;80(7):1715-35.
60. Ishima R, Torchia DA. Protein dynamics from NMR. Nature Structural & Molecular Biology. 2000;7(9):740-3.
61. Vivian JT, Callis PR. Mechanisms of tryptophan fluorescence shifts in proteins. Biophysical journal. 2001;80(5):2093-109.
62. Best RB, Vendruscolo M. Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. Structure. 2006;14(1):97-106.
63. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nature Structural & Molecular Biology. 2002;9(9):646-52.
64. Kay LE. NMR studies of protein structure and dynamics. Journal of Magnetic Resonance. 2005;173(2):193-207.
65. Wagner G. NMR relaxation and protein mobility. Current opinion in structural biology. 1993;3(5):748-54.
66. Kay LE, Torchia DA, Bax A. Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. Biochemistry. 1989;28(23):8972-9.
67. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature. 2005;433(7022):128-32.
68. Karas M, Bachmann D, Hillenkamp F. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Analytical Chemistry. 1985;57(14):2935-9.
69. Dole M, Mack L, Hines R, Mobley R, Ferguson L, Alice Md. Molecular beams of macroions. The Journal of Chemical Physics. 1968;49(5):2240-9.
70. Trauger SA, Webb W, Siuzdak G. Peptide and protein analysis with mass spectrometry. Journal of Spectroscopy. 2002;16(1):15-28.
71. Sinz A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. Mass spectrometry reviews. 2006;25(4):663-82.
72. Dass C. Fundamentals of contemporary mass spectrometry: John Wiley & Sons; 2007.
73. Resource TsIGE. 2016.
74. Kelleher NL. Peer reviewed: Top-down proteomics. Analytical chemistry. 2004;76(11):196 A-203 A.
75. Petrotchenko EV, Borchers CH. Crosslinking combined with mass spectrometry for structural proteomics. Mass spectrometry reviews. 2010;29(6):862-76.
76. Pan Y, Brown L, Konermann L. Mapping the structure of an integral membrane protein under semi-denaturing conditions by laser-induced oxidative labeling and mass spectrometry. Journal of molecular biology. 2009;394(5):968-81.
77. Englander JJ, Del Mar C, Li W, Englander SW, Kim JS, Stranz DD, et al. Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(12):7057-62.
78. Benson EE, Linderstrom-Lang K. Deuterium exchange between myoglobin and water. Biochimica et biophysica acta. 1959;32:579-81.

79. Englander S, Mayne L, Bai Y, Sosnick T. Hydrogen exchange: The modern legacy of Linderstrøm-Lang. Protein science. 1997;6(5):1101-9.
80. Englander SW, Sosnick TR, Englander JJ, Mayne L. Mechanisms and uses of hydrogen exchange. Current opinion in structural biology. 1996;6(1):18-23.
81. Lewis H, Wang C, Zhao X, Hamuro Y, Conners K, Kearins M, et al. Structure and dynamics of NBD1 from CFTR characterized using crystallography and hydrogen/deuterium exchange mass spectrometry. Journal of molecular biology. 2010;396(2):406-30.
82. Katta V, Chait BT. Hydrogen/deuterium exchange electrospray ionization mass spectrometry: a method for probing protein conformational changes in solution. Journal of the American Chemical Society. 1993;115(14):6317-21.
83. Engen JR. Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. Analytical chemistry. 2009;81(19):7870-5.
84. Hoofnagle AN, Resing KA, Ahn NG. Protein analysis by hydrogen exchange mass spectrometry. Annual review of biophysics and biomolecular structure. 2003;32(1):1-25.
85. Dempsey CE. Hydrogen exchange in peptides and proteins using NMR spectroscopy. Progress in Nuclear Magnetic Resonance Spectroscopy. 2001;39(2):135-70.
86. Zhang Y-Z. Protein and peptide structure and interactions studied by hydrogen exchange and NMR. 1995.
87. Bai Y, Milne JS, Mayne L, Englander SW. Primary structure effects on peptide group hydrogen exchange. Proteins. 1993;17(1):75.
88. Alder BJ, Wainwright T. Studies in molecular dynamics. I. General method. The Journal of Chemical Physics. 1959;31(2):459-66.
89. De Wette F, Allen R, Hughes D, Rahman A. Crystallization with a Lennard-Jones potential: A computer experiment. Physics Letters A. 1969;29(9):548-9.
90. Rahman A. Correlations in the motion of atoms in liquid argon. Physical Review. 1964;136(2A):A405.
91. van Gunsteren WF, Daura X, Mark AE. GROMOS force field. Encyclopedia of computational chemistry. 1998.
92. MacKerell AD, Banavali N, Foloppe N. Development and current status of the CHARMM force field for nucleic acids. Biopolymers. 2000;56(4):257-65.
93. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. Journal of computational chemistry. 2004;25(9):1157-74.
94. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977;267(5612):585-90.
95. Born M, Oppenheimer R. Zur quantentheorie der molekeln. Annalen der Physik. 1927;389(20):457-84.
96. Krivov SV, Karplus M. Hidden complexity of free energy surfaces for peptide (protein) folding. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(41):14766-70.
97. Ma J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure. 2005;13(3):373-80.
98. Brooks B, Karplus M. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. Proceedings of the National Academy of Sciences. 1985;82(15):4995-9.
99. Bahar I, Rader A. Coarse-grained normal mode analysis in structural biology. Current opinion in structural biology. 2005;15(5):586-92.

100. Ricardo Batista P. Estudo da flexibilidade da protease do HIV-1 por Modelagem e dinâmica molecular: análise dos modos normais e dos modos consensus: Paris 7; 2009.
101. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proceedings of the National Academy of Sciences. 1987;84(19):6611-5.
102. Hansmann UH, Okamoto Y. New Monte Carlo algorithms for protein folding. Current opinion in structural biology. 1999;9(2):177-83.
103. Derreumaux P. Generating ensemble averages for small proteins from extended conformations by Monte Carlo simulations. Physical review letters. 2000;85(1):206.
104. Vendruscolo M, Paci E, Dobson CM, Karplus M. Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. Journal of the American Chemical Society. 2003;125(51):15686-7.
105. Goffe WL, Ferrier GD, Rogers J. Global optimization of statistical functions with simulated annealing. Journal of Econometrics. 1994;60(1):65-99.
106. Bohachevsky IO, Johnson ME, Stein ML. Generalized simulated annealing for function optimization. Technometrics. 1986;28(3):209-17.
107. Nilges M, Gronenborn AM, Brünger AT, Clore GM. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. Protein Engineering. 1988;2(1):27-38.
108. Suvorina MY, Surin A, Dovidchenko N, Lobanov MY, Galzitskaya O. Comparison of experimental and theoretical data on hydrogen-deuterium exchange for ten globular proteins. Biochemistry (Moscow). 2012;77(6):616-23.
109. Lobanov MY, Suvorina MY, Dovidchenko NV, Sokolovskiy IV, Surin AK, Galzitskaya OV. A novel web server predicts amino acid residue protection against hydrogen-deuterium exchange. Bioinformatics. 2013;29(11):1375-81.
110. Dovidchenko NV, Lobanov MY, Garbuzynskiy SO, Galzitskaya OV. Prediction of amino acid residues protected from hydrogen-deuterium exchange in a protein chain. Biochemistry Biokhimiia. 2009;74(8):888-97.
111. Tartaglia GG, Cavalli A, Vendruscolo M. Prediction of local structural stabilities of proteins from their amino acid sequences. Structure. 2007;15(2):139-43.
112. Shrake A, Rupley J. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. Journal of molecular biology. 1973;79(2):351-71.
113. Truhlar SM, Croy CH, Torpey JW, Koeppe JR, Komives EA. Solvent accessibility of protein surfaces by amide H/2H exchange MALDI-TOF mass spectrometry. Journal of the American Society for Mass Spectrometry. 2006;17(11):1490-7.
114. Park I-H, Venable JD, Steckler C, Cellitti SE, Lesley SA, Spraggon G, et al. Estimation of Hydrogen-Exchange Protection Factors from MD Simulation Based on Amide Hydrogen Bonding Analysis. Journal of chemical information and modeling. 2015.
115. Bahar I, Wallqvist A, Covell D, Jernigan R. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. Biochemistry. 1998;37(4):1067-75.
116. Skinner JJ, Lim WK, Bedard S, Black BE, Englander SW. Protein hydrogen exchange: testing current models. Protein science : a publication of the Protein Society. 2012;21(7):987-95.

117. Hilser VJ, Freire E. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. Journal of molecular biology. 1996;262(5):756-72.
118. Kan Z-Y, Walters BT, Mayne L, Englander SW. Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. Proceedings of the National Academy of Sciences. 2013;110(41):16438-43.
119. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792-7.
120. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. Nucleic acids research. 2004;32(suppl 2):W665-W7.
121. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins: Structure, Function, and Bioinformatics. 2005;61(4):704-21.
122. MacKerell Jr AD, Bashford D, Bellott M, Dunbrack Jr RL, Evanseck JD, Field MJ, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins†. The journal of physical chemistry B. 1998;102(18):3586-616.
123. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. Journal of computational chemistry. 2010;31(4):671-90.
124. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.
125. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 2006;22(21):2695-6.
126. Mouawad L, Perahia D. Diagonalization in a mixed basis: A method to compute low-frequency normal modes for large macromolecules. Biopolymers. 1993;33(4):599-611.
127. Tama F, Gadea FX, Marques O, Sanejouand YH. Building-block approach for determining low-frequency normal modes of macromolecules. Proteins: Structure, Function, and Bioinformatics. 2000;41(1):1-7.
128. Kurkcuoglu O, Jernigan RL, Doruker P. Collective Dynamics of Large Proteins from Mixed Coarse-Grained Elastic Network Model. Qsar & Combinatorial Science. 2005;24(4):443-8.
129. Atilgan A, Durell S, Jernigan R, Demirel M, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysical journal. 2001;80(1):505-15.
130. Hinsen K, Petrescu A-J, Dellerue S, Bellissent-Funel M-C, Kneller GR. Harmonicity in slow protein dynamics. Chemical Physics. 2000;261(1):25-37.
131. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
132. Akaike H. A new look at the statistical model identification. Automatic Control, IEEE Transactions on. 1974;19(6):716-23.
133. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of behavioral medicine. 2003;26(3):172-81.
134. Uhlin U, Cox GB, Guss JM. Crystal structure of the ϵ subunit of the proton-translocating ATP synthase from Escherichia coli. Structure. 1997;5(9):1219-30.

135. Rodriguez AD, Dunn SD, Konermann L. ATP-Induced Dimerization of the F0F1 ε Subunit from Bacillus PS3: A Hydrogen Exchange–Mass Spectrometry Study. Biochemistry. 2014;53(24):4072-80.
136. Yagi H, Kajiwara N, Tanaka H, Tsukihara T, Kato-Yamada Y, Yoshida M, et al. Structures of the thermophilic F1-ATPase ε subunit suggesting ATP-regulated arm motion of its C-terminal domain in F1. Proceedings of the National Academy of Sciences. 2007;104(27):11233-8.
137. Chen J, Lu Z, Sakon J, Stites WE. Increasing the thermostability of staphylococcal nuclease: implications for the origin of protein thermostability. Journal of molecular biology. 2000;303(2):125-30.
138. Milano SK, Pace HC, Kim Y-M, Brenner C, Benovic JL. Scaffolding functions of arrestin-2 revealed by crystal structure and mutagenesis. Biochemistry. 2002;41(10):3321-8.
139. Yun Y, Kim DK, Seo M-D, Kim K-M, Chung KY. Different conformational dynamics of β-arrestin1 and β-arrestin2 analyzed by hydrogen/deuterium exchange mass spectrometry. Biochemical and biophysical research communications. 2015;457(1):50-7.
140. Jacobs MD, Harrison SC. Structure of an IκBα/NF-κB complex. Cell. 1998;95(6):749-58.
141. Truhlar SM, Torpey JW, Komives EA. Regions of IκBα that are critical for its inhibition of NF-κB· DNA interaction fold upon binding to NF-κB. Proceedings of the National Academy of Sciences. 2006;103(50):18951-6.
142. DiDonato M, Craig L, Huff ME, Thayer MM, Cardoso RM, Kassmann CJ, et al. ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization. Journal of molecular biology. 2003;332(3):601-15.
143. Molnar KS, Karabacak NM, Johnson JL, Wang Q, Tiwari A, Hayward LJ, et al. A Common Property of Amyotrophic Lateral Sclerosis-associated Variants DESTABILIZATION OF THE COPPER/ZINC SUPEROXIDE DISMUTASE ELECTROSTATIC LOOP. Journal of Biological Chemistry. 2009;284(45):30965-73.
144. Borrego-Diaz E, Kerff F, Lee SH, Ferron F, Li Y, Dominguez R. Crystal structure of the actin-binding domain of α-actinin 1: Evaluating two competing actin-binding models. Journal of structural biology. 2006;155(2):230-8.
145. Full SJ, Deinzer ML, Ho PS, Greenwood JA. Phosphoinositide binding regulates α-actinin CH2 domain structure: Analysis by hydrogen/deuterium exchange mass spectrometry. Protein Science. 2007;16(12):2597-604.
146. Lieberman RL, Wustman BA, Huertas P, Powe AC, Pine CW, Khanna R, et al. Structure of acid β-glucosidase with pharmacological chaperone provides insight into Gaucher disease. Nature chemical biology. 2007;3(2):101-7.
147. Kornhaber GJ, Tropak MB, Maegawa GH, Tuske SJ, Coales SJ, Mahuran DJ, et al. Isofagomine induced stabilization of glucocerebrosidase. ChemBioChem. 2008;9(16):2643-9.
148. Aranda R, Cai H, Worley CE, Levin EJ, Li R, Olson JS, et al. Structural analysis of fish versus mammalian hemoglobins: effect of the heme pocket environment on autooxidation and hemin loss. Proteins: Structure, Function, and Bioinformatics. 2009;75(1):217-30.
149. Sowole MA, Konermann L. Comparative Analysis of Oxy-Hemoglobin and Aquomet-Hemoglobin by Hydrogen/Deuterium Exchange Mass Spectrometry. Journal of the American Society for Mass Spectrometry. 2013;24(7):997-1005.
150. Loh SN, Prehoda KE, Wang J, Markley JL. Hydrogen exchange in unligated and ligated staphylococcal nuclease. Biochemistry. 1993;32(41):11022-8.

151. Kuriyan J, Wilz S, Karplus M, Petsko GA. X-ray structure and refinement of carbon-monoxy (Fe II)-myoglobin at 1.5 Å resolution. Journal of molecular biology. 1986;192(1):133-54.
152. Cavagnero S, Thériault Y, Narula SS, Dyson HJ, Wright PE. Amide proton hydrogen exchange rates for sperm whale myoglobin obtained from 15N-1H NMR spectra. Protein Science. 2000;9(1):186-93.
153. Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 Åresolution. Journal of molecular biology. 1987;194(3):531-44.
154. Pan Y, Briggs MS. Hydrogen exchange in native and alcohol forms of ubiquitin. Biochemistry. 1992;31(46):11405-12.
155. Chandra N, Brew K, Acharya KR. Structural evidence for the presence of a secondary calcium binding site in human α-lactalbumin. Biochemistry. 1998;37(14):4767-72.
156. Schulman BA, Redfield C, Peng Z-y, Dobson CM, Kim PS. Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human α-lactalbumin. Journal of molecular biology. 1995;253(5):651-7.
157. Wlodawer A, Walter J, Huber R, Sjölin L. Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and x-ray refinement of crystal form ii. Journal of molecular biology. 1984;180(2):301-29.
158. Kim KS, Fuchs JA, Woodward CK. Hydrogen exchange identifies native-state motional domains important in protein folding. Biochemistry. 1993;32(37):9600-8.
159. Lim D, Sanschagrin F, Passmore L, De Castro L, Levesque RC, Strynadka NC. Insights into the molecular basis for the carbenicillinase activity of PSE-4 β-lactamase from crystallographic and kinetic studies. Biochemistry. 2001;40(2):395-402.
160. Morin S, Gagné SM. NMR Dynamics of PSE-4 β-Lactamase: An Interplay of ps-ns Order and μs-ms Motions in the Active Site. Biophysical journal. 2009;96(11):4681-91.
161. Tsuge H, Ago H, Noma M, Nitta K, Sugai S, Miyano M. Crystallographic Studies of a Calcium Binding Lysozyme from Equine Milk at 2.5 ÅResolution. Journal of biochemistry. 1992;111(2):141-3.
162. Morozova-Roche LA, Arico-Muendel CC, Haynie DT, Emelyanenko VI, Van Dael H, Dobson CM. Structural characterisation and comparison of the native and A-states of equine lysozyme. Journal of molecular biology. 1997;268(5):903-21.
163. Ramanadham M, Sieker L, Jensen L. Refinement of triclinic lysozyme: II. The method of stereochemically restrained least squares. Acta Crystallographica Section B: Structural Science. 1990;46(1):63-9.
164. Radford SE, Buck M, Topping KD, Dobson CM, Evans PA. Hydrogen exchange in native and denatured states of hen egg-white lysozyme. Proteins: Structure, Function, and Bioinformatics. 1992;14(2):237-48.
165. Bycroft M, Ludvigsen S, Fersht AR, Poulsen FM. Determination of the three-dimensional solution structure of barnase using nuclear magnetic resonance spectroscopy. Biochemistry. 1991;30(35):8697-701.
166. Clarke J, Fersht AR. An evaluation of the use of hydrogen exchange at equilibrium to probe intermediates on the protein folding pathway. Folding and Design. 1996;1(4):243-54.
167. Ross SA, Sarisky CA, Su A, Mayo SL. Designed protein G core variants fold to native-like structures: Sequence selection by ORBIT tolerates variation in backbone specification. Protein Science. 2001;10(2):450-4.

168. Soss SE, Flynn PF. Functional implications for a prototypical K-turn binding protein from structural and dynamical studies of 15.5 K. Biochemistry. 2007;46(51):14979-86.
169. Gargaro AR, Soteriou A, Frenkiel TA, Bauer CJ, Birdsall B, Polshakov VI, et al. The solution structure of the complex of Lactobacillus casei dihydrofolate reductase with methotrexate. Journal of molecular biology. 1998;277(1):119-34.
170. Feeney J, Birdsall B, Kovalevskaya NV, Smurnyy YD, Navarro Peran EM, Polshakov VI. NMR structures of Apo L. casei dihydrofolate reductase and its complexes with trimethoprim and NADPH: Contributions to positive cooperative binding from ligand-induced refolding, conformational changes, and interligand hydrophobic interactions. Biochemistry. 2011;50(18):3609-20.
171. Ranjan N, Damberger FF, Sutter M, Allain FH-T, Weber-Ban E. Solution structure and activation mechanism of ubiquitin-like small archaeal modifier proteins. Journal of molecular biology. 2011;405(4):1040-55.
172. Walma T, Aelen J, Nabuurs SB, Oostendorp M, Van Den Berk L, Hendriks W, et al. A closed binding pocket and global destabilization modify the binding properties of an alternatively spliced form of the second PDZ domain of PTP-BL. Structure. 2004;12(1):11-20.
173. Lee G-i, Ding Z, Walker JC, Van Doren SR. NMR structure of the forkhead-associated domain from the Arabidopsis receptor kinase-associated protein phosphatase. Proceedings of the National Academy of Sciences. 2003;100(20):11261-6.
174. Lee G-i, Li J, Walker JC, Van Doren SR. Letter to the Editor: 1 H, 13 C and 15 N resonance assignments of the kinase-interacting FHA domain of Arabidopsis thaliana kinase-associated protein phosphatase. Journal of biomolecular NMR. 2003;25(3):253-4.
175. Costa MGS, Batista PR, Bisch PM, Perahia D. Exploring Free Energy Landscapes of Large Conformational Changes: Molecular Dynamics with Excited Normal Modes. J Chem Theory Comput. 2015;11(6):2755-67.