Universidade de Brasília

Programa de Pós-graduação em Biologia Molecular

Simulação do enovelamento de proteínas com potenciais de enterramentos atômicos dependentes da sequência

por

Marx Gomes Van der Linden

Orientador:

Prof. Antônio F. Pereira de Araújo

Brasília - DF

Dezembro de 2013

A Luciana e a meus pais.

Agradecimentos

A Luciana, por ter surgido em minha vida e a transformado para sempre.

A meus pais, pelo suporte e incentivo constantes, sem os quais nada disso teria sido possível.

Ao professor Antônio Francisco Pereira de Araújo, pela confiança e apoio na execução deste projeto.

Aos companheiros de laboratório, Diogo, Juliana, Leandro, Lindomar e Victor, pela companhia inestimável nesses últimos quatro anos de jornada.

Ao Programa de Pós-Graduação em Biologia Molecular, à Universidade de Brasília e a seus funcionários.

Ao CNPq, pelo financiamento.

Resumo da Tese apresentada ao Departamento de Biologia Molecular da Universidade de Brasília como parte dos requisitos necessários para a obtenção do grau de Doutor em Biologia Molecular.

Simulação do enovelamento de proteínas com potenciais de enterramentos atômicos dependentes da sequência

Marx Gomes van der Linden

Dezembro de 2013

Orientador: Prof. Dr. Antônio Francisco Pereira de Araújo.

Há muito tempo se sabe que estruturas tridimensionais de proteínas são determinadas por suas respectivas sequências de aminoácidos, entretanto as possíveis regras que associam sequências a estruturas continuam em larga medida desconhecidos. A construção de um algoritmo geral para predição *ab initio* de estruturas proteicas, isto é, uma metodologia que determine computacionalmente a estrutura nativa de qualquer proteína com base em sua sequência de aminoácidos, é possivelmente o maior desafio teórico atual da Biofísica computacional.

Esta tese parte da hipótese inovadora de que a única informação dependente de sequência necessária para se determinar a conformação nativa de uma proteína é a distância de cada um de seus átomos até o centro geométrico da estrutura, uma medida que denominamos enterramento atômico. O objetivo principal do trabalho é avaliar a aplicabilidade dessa hipótese através da construção da primeira versão de um método computacional para predição *ab initio* que utilize os enterramentos como único intermediário informacional entre sequência e estrutura proteica.

Nossa metodologia está dividida em duas partes principais: a primeira descreve a construção de um método computacional capaz de realizar predições de enterramentos atômicos a partir de sequências de aminoácidos; a segunda trata de simulações computacionais do enovelamento que utilizam as predições obtidas na primeira parte para obter estruturas terciárias próximas à nativa para algumas proteínas selecionadas.

O método computacional que foi desenvolvido neste trabalho para prever os enterramentos atômicos de proteínas a partir de suas sequências de aminoácidos é um algoritmo de aprendizado supervisionado baseado em um modelo oculto de Markov (*Hidden Markov Model* - HMM). Análises informacionais realizadas sobre os resultados alcançados pelo HMM implementado revelaram que suas predições são ótimas, no sentido de que elas são capazes de extrair praticamente toda a informação sobre enterramentos disponível na sequência, de acordo com o modelo de dados adotado.

Na segunda parte do trabalho, um método de dinâmica molecular com um potencial simplificado, que não inclui nenhuma informação derivada da estrutura nativa e utiliza as predições de enterramentos atômicos como base para o único termo dependente de sequência do potencial, foi aplicado a três proteínas pertencentes a diferentes classes estruturais, selecionadas entre os melhores resultados de predição alcançados na etapa anterior. Nos três casos, o algoritmo foi capaz de obter e distinguir a conformação nativa correta em simulações do enovelamento que partiram de conformações completamente estendidas.

Os resultados obtidos demonstram que, ao menos para algumas classes de proteínas, os enterramentos atômicos podem de fato atuar como os únicos intermediários informacionais entre sequência e estrutura, fornecendo um novo arcabouço conceitual que esperamos contribuir para a compreensão dos fundamentos do enovelamento proteico e para a investigação do problema da predição estrutural de proteínas.

Abstract of Thesis presented to the Department of Molecular Biology of the University of Brasília as a partial fulfillment of the requirements for the degree of Doctor of Sciences in Molecular Biology.

Protein folding simulation with
sequence-dependent atomic burial potentials

Marx Gomes van der Linden

December, 2013

Advisor: Prof. Dr. Antônio Francisco Pereira de Araújo.

It has been long known that the three-dimensional structures of proteins are determined by their respective amino acid sequences. The possible rules, however, that associate protein sequences to structures remain at large elusive. The development of an *ab initio* general algorithm for protein structure prediction, that is, a computational methodology that should be able to determine the native structure of any protein from its amino acid sequence, is possibly the greatest theoretical challenge of current computational biophysics.

The premise of this work is based on the innovative hypothesis that the only sequence-dependent information needed to determine the native conformation of a protein is the distance of each of its atoms to the geometric center of the structure, a measure we call atomic burial. Our main objective is to evaluate the applicability of this hypothesis through the development of the first version of a computational method for *ab initio* prediction that employs burials as the only informational intermediate between protein sequence and structure.

Out methodology is divided in two main parts: the first part describes the construction of a computational method to produce atomic burial predictions from amino acid sequences; the second part refers to folding simulations that employ the previously obtained predictions to obtain tertiary structures that are close to the native configuration for selected proteins.

The computational method that was developed in this work for atomic burial prediction from amino acid sequences is a supervised learning algorithm based in a Hidden Markov Model (HMM). Informational analyses performed on the HMM results revealed that its predictions are optimal, in the sense that they are capable of extracting almost the totality of the information about burials that is available in amino acid sequences, according to the data model that was adopted.

In the second part of this work, a molecular dynamics method with a simplified potential, which does not include any information derived from the native structure and employs atomic burial predictions as its only sequence-dependent term, was applied to three proteins belonging to different structural classes, selected among the best prediction results achieved in the previous step. In all three cases, the algorithm was capable of obtaining and distinguishing the native conformation in folding simulations that started from fully extended conformations.

The results achieved in this work demonstrate that, at least for some protein classes, atomic burials are in fact able to act as the sole informational intermediates between sequence and structure, providing a new conceptual framework that we expect to be able to contribute for our knowledge of the fundamentals of protein folding and to the problem of protein structure prediction.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

- $A = \{a_{ij}\}$: Matriz de probabilidades de transição entre os estados i e j (HMM)
- $B = \{b_j(k)\}$: Matriz das probabilidades de emissão do símbolo k quando o sistema se encontra no estado j (HMM)
- C_α: Carbono-α
- C_β: Carbono-β
- CHARMM: *Chemistry at HARvard Molecular Mechanics*
- CsBc: Proteína de choque frio Bc-Csp de *Bacillus caldolyticus*
- CASP: *Critical Assessment of Techniques for Protein Structure Prediction* – Avaliação Crítica de Técnicas para Predição de Estruturas de Proteínas
- DSSP: *Define Secondary Structure of Proteins* – Definir a Estrutura Secundária de Proteínas
- f: Tamanho do fragmento usado na construção dos estados ocultos no programa HmmPred
- $H(X)$: Entropia de X
- $h(X)$: Densidade de entropia de X
- $H(X|Y)$: Entropia condicional de X dado Y
- $h(X|Y)$: Densidade de entropia condicional de X dado Y
- HMM: *Hidden Markov Model* – Modelo Oculto de Markov
- HP: Hidrofóbico-Polar (alfabeto de estrutura primária)
- HPN: Hidrofóbico-Neuto-Polar (alfabeto de estrutura primária)
- $I(X;Y)$: Transinformação entre X e Y

- $i(X;Y)$: Densidade de transinformação entre X e Y
- PDB: *Protein Data Bank*
- ProtGSsp: Proteína G de estreptococos
- Q: Uma sequência de aminoácidos
- $\mathcal{Q}$: Alfabeto de símbolos observáveis no programa HmmPred
- RePc: Proteína efetora RxLR de *Phytophthora capsici*
- RMSD: *Root Mean Square Deviation* – Desvio Médio Quadrático
- Y: Uma sequência de enterramentos
- $\mathcal{Y}$: Alfabeto de variáveis ocultas no programa HmmPred
- $\gamma_t(i)$: A probabiliade de que o sistema esteja no estado i quando o símbolo observado na posição t foi emitido (HMM)

Capítulo 1

Introdução

1.1 Problema a ser tratado

Esta tese de doutorado tem como tema principal a proposição de uma nova abordagem para o estudo de alguns aspectos do chamado problema do enovelamento proteico. De acordo com uma revisão recente publicada por Dill e MacCallum[2], esse problema consiste principalmente na busca à solução de algumas questões básicas, entre as quais se incluem:

1. Qual é o código físico através do qual sequências de aminoácidos determinam a estrutura nativa de proteínas?

2. É possível desenvolver um algoritmo computacional que seja capaz de prever estruturas de proteínas a partir de suas sequências?

O ponto de partida deste trabalho é a proposição de uma possível resposta à primeira pergunta, com a apresentação da hipótese de que a linguagem do enovelamento consiste em uma codificação, na sequência, da distância física de cada átomo da proteína ao centro geométrico de sua estrutura nativa, uma medida que denominamos enterramento atômico. Essa hipótese será validada através da construção de uma metodologia computacional que tem como objetivo final prever estruturas proteicas a partir de sequências de aminoácidos, configurando uma abordagem inovadora que acreditamos poder acrescentar uma contribuição importante à eventual solução do segundo problema.

A metodologia geral do trabalho é dividida em duas partes (figura 1.1): a primeira consiste na construção de um algoritmo para predição de enterramentos atômicos a partir de sequências; a segunda, no desenvolvimento de um método que seja capaz de empregar essas predições para obter e distinguir estruturas nativas de proteínas em simulações computacionais do enovelamento.

Esta introdução parte de uma breve revisão sobre as forças que estabilizam estruturas de proteínas e em seguida discute o conhecimento atual sobre o código do enovelamento. A definição precisa de

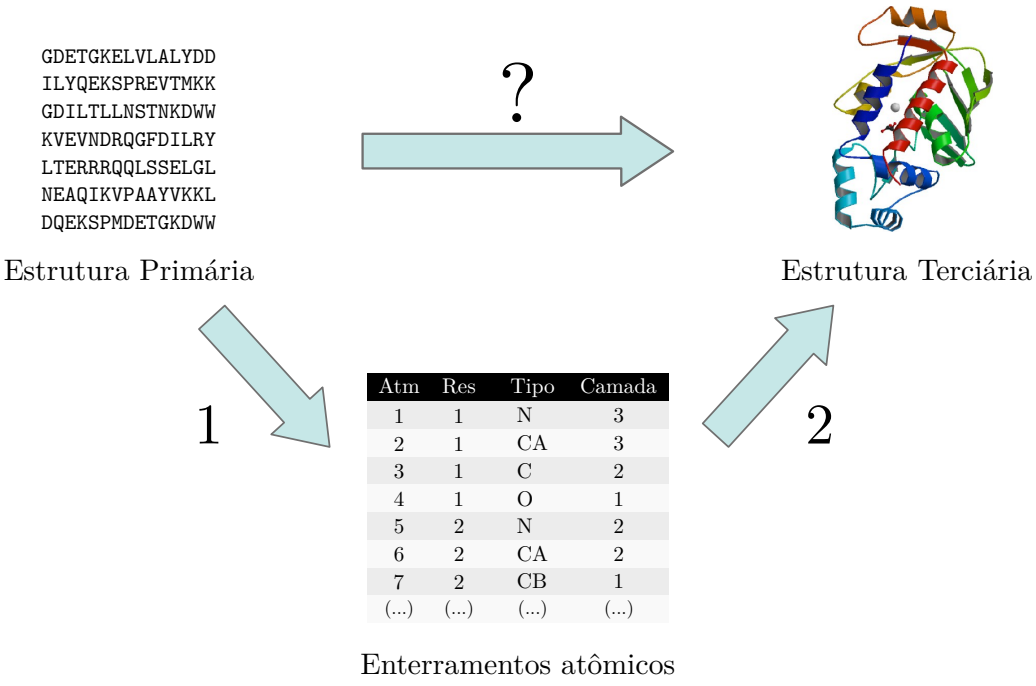


Figura 1.1: Plano geral da metodologia proposta neste trabalho para predição de estruturas proteicas a partir de sequências.

enterramentos atômicos, tal como será utilizada no restante do trabalho, será apresentada e justificada no contexto desses fundamentos. O capítulo conclui com uma discussão do método de predição de estruturas proposto em comparação a outras abordagens conhecidas, seguida pela delineação dos objetivos específicos do trabalho.

O restante da tese se divide em duas partes principais, refletindo a divisão de tarefas empregada pela metodologia. Os capítulos 2 e 3 tratam do problema de prever os enterramentos atômicos a partir de sequências de aminoácidos; os capítulos 4 e 5 descrevem o procedimento de simulação computacional que será utilizado para tentar determinar a estrutura nativa com o auxílio dos enterramentos preditos. Por fim, o capítulo 6 apresenta as conclusões gerais do trabalho e as perspectivas para o futuro.

1.2 Estruturas tridimensionais de proteínas

A primeira proteína a ter sua estrutura tridimensional determinada experimentalmente foi a mioglobina, em 1958[3]. O modelo revelado por este trabalho, embora de baixa resolução, foi suficiente para que se percebesse que o enovelamento proteico produzia uma estrutura complexa, assimétrica e muito mais irregular que aquelas preditas pelas teorias então vigentes. Hoje se sabe que a conformação tridimensional específica no espaço de todos os átomos de uma proteína, conhecida como estrutura terciária ou estrutura nativa, é o fator característico e determinante de sua função biológica, uma observação que desencadeou o início dos estudos em biologia estrutural de proteínas.

O conhecido experimento de Anfinsen e colaboradores em 1963[4] demonstrou que pequenas proteí-

nas, após desnaturadas, são capazes de retornar espontaneamente ao seu estado nativo sem o auxílio de quaisquer catalisadores. Este resultado estabeleceu que a estrutura nativa de uma proteína está no mínimo global da energia livre acessível da molécula e é determinada apenas por sua sequência de aminoácidos. A elucidação de como exatamente ocorre a codificação da estrutura terciária de uma proteína em sua sequência ainda é um dos principais problemas não resolvidos da biofísica molecular.

Em 1969, considerando todos os graus de liberdade das ligações covalentes de um estrutura proteica, Cyrus Levinthal calculou que uma proteína pequena pode ser capaz de assumir uma quantidade teórica de aproximadamente 10^{300} conformações diferentes[5]. Se a proteína precisasse percorrer cada uma dessas conformações aleatoriamente até encontrar a estrutura nativa, o processo de enovelamento demoraria mais que a idade do universo. Uma vez que proteínas podem se enovelar em escalas de mili a microssegundos, conclui-se que o processo de busca da estrutura nativa deve ser mais parecido com uma busca direcionada que com uma série de flutuações aleatórias. Esta observação ficou conhecida como o "paradoxo de Levinthal".

O sistema formado por uma cadeia polipeptídica desnaturada interagindo com o solvente por meio de ligações de hidrogênio tem, a princípio, um número de graus de liberdade muito maior que o de uma proteína enovelada, caracterizando uma alta entropia conformacional que precisa ser compensada pelo processo de enovelamento. A energia livre da proteína enovelada é apenas marginalmente mais favorável que a do estado desnaturado[6] e emerge como resultado de um balanço sutil entre forças poderosas com efeitos opostos. A contribuição entálpica para a estabilização do estado nativo vem de interações internas entre átomos não ligados covalentemente, como pontes salinas e interações de van der Waals. A contribuição entrópica vem principalmente do chamado efeito hidrofóbico.

1.2.1 Interações não-ligadas

Átomos pertencentes a cadeias laterais carregadas positiva ou negativamente podem atrair ou repelir um ao outro durante o processo de enovelamento proteico, formando ligações eletrostáticas estáveis conhecidas como pontes salinas.

Íons livres em solução são altamente solvatados, e a dessolvatação dos grupos polares de cadeias laterais acarreta em uma penalidade entrópica que não é compensada pela formação de pontes salinas. Por esse motivo, acredita-se que o efeito principal da formação dessas interações seja o de conferir especificidade à estrutura[7], eliminando do espaço conformacional configurações estruturais próximas à nativa que resultem em grupos polares livres em um ambiente apolar.

Em geral, pontes salinas localizadas em posições mais enterradas na proteína são mais bem conservadas na evolução que aquelas mais expostas, mas em todos os casos, os resíduos envolvidos fre-

quentemente são substituídos na evolução por cadeias laterais hidrofóbicas[8].

Força de van der Waals são interações não-covalentes que ocorrem entre moléculas eletricamente neutras que formam dipolos permanentes ou induzidos. A alta compactação da estrutura de proteínas globulares faz com que esses contatos ocorram em grandes números no interior da molécula. Por esse motivo, embora individualmente fracas, as interações de van der Walls têm um efeito conjunto significativo na contribuição energética para a estabilidade estrutural de proteínas[9].

1.2.2 O efeito hidrofóbico

A rede de ligações de hidrogênio que normalmente é formada entre as moléculas de água no estado líquido é subitamente interrompida quando uma molécula apolar, incapaz de formar ligações de hidrogênio, é adicionada ao meio. Essa interrupção força as moléculas de água a se organizem em um envoltório regular ao redor do soluto, o qual pode se estender por várias camadas de moléculas, em um arranjo que diminui desfavoravelmente a entropia total do sistema. Em proteínas, esse efeito é suavizado quanto menor for o número de resíduos hidrofóbicos expostos ao solvente. O resultado é a tendência universal em estruturas proteicas globulares de distribuir resíduos com cadeias polares ao longo da superfície, em contato com o solvente, enquanto aminoácidos com cadeias laterais hidrofóbicas se concentram no núcleo da proteína. Esse fenômeno é conhecido como o **efeito hidrofóbico**.

A importância do efeito hidrofóbico no processo de enovelamento foi sugerida primeiramente por Kauzmann em um trabalho publicado em 1959[10]. Neste artigo, Kauzmann observou que cada interação hidrofóbica no núcleo da proteína implica na formação de uma nova ligação de hidrogênio entre um átomo da superfície proteica e outro do solvente e que a importância da formação dessa nova ligação é necessariamente superior, por mais de uma ordem de magnitude, a qualquer possível mudança nas forças das ligações de hidrogênio existentes. O autor também mencionou como evidências da importância do efeito hidrofóbico as observações de que proteínas são desnaturadas por solventes apolares e de que a estabilidade da estrutura decai não só com o aumento, mas também com a queda da temperatura ambiente, de maneira consistente com o conhecimento de que solutos apolares se tornam mais solúveis em água em baixas temperaturas[10;11].

1.3 O código do enovelamento

Todas as estruturas de proteínas têm atributos conformacionais comuns, como a conformação da ligação peptídica, a tendência à formação de ligações de hidrogênio e as restrições estereoquímicas características da topologia do polipeptídeo. Esses atributos independem da sequência de aminoácidos e funcionam como restrições ao espaço conformacional disponível a estruturas proteicas em geral. De

alguma forma, entretanto, a sequência determina a estrutura específica em que cada diferente proteína se enovela. A expressão **código do enovelamento** se refere ao mecanismo através do qual isso ocorre.

Até meados da década de 1980, acreditava-se que a principal informação estrutural codificada na sequência de aminoácidos seria o padrão de interações na cadeia principal da proteína, isto é os ângulos ϕ e ψ e as ligações de hidrogênio que originavam as estruturas secundárias[12]. De fato, existem padrões conhecidos que associam os diferentes aminoácidos a preferências específicas de formação de estruturas secundárias[13], e é possível prever essas estruturas a partir da sequência com razoável grau de precisão[14]. Entretanto, mesmo um conhecimento perfeito da estrutura secundária é insuficiente para que se determine a estrutura nativa de uma proteína[15], o que torna implausível que essa informação constitua o cerne do código do enovelamento. Outro problema com hipóteses para o código do enovelamento que enfatizam as interações responsáveis por estruturas secundárias é que elas são incapazes de explicar como os ângulos ϕ e ψ podem assumir valores distintos para a mesma sequência de aminoácidos de maneira dependente do solvente ou da temperatura, nos estados nativo e desnaturado[12].

O efeito hidrofóbico, embora tivesse um papel conhecido no enovelamento proteico desde os trabalhos de Kauzmann em 1959, foi visto durante muito tempo como uma força que apenas auxiliava no colapso da macromolécula, mas não tinha um papel mais importante na determinação de estruturas terciárias específicas[12].

Nos anos 1980, com o desenvolvimento de modelos simplificados de estruturas proteicas para fins de estudos de mecânica estatística, foi possível começar a investigar de maneira quantitativa questões que modelos atômicos detalhados não conseguiam responder. Destes modelos simplificados, o mais conhecido é o modelo HP, no qual aminoácidos são representados por esferas pertencentes a apenas um de dois tipos (polar ou hidrofóbico), conectadas por ligações de comprimento fixo e restritos a posições discretas em uma matriz de duas ou três dimensões.

Experimentos computacionais com modelos HP normalmente tentam encontrar estruturas que maximizam o número de contatos entre resíduos hidrofóbicos para uma dada sequência. Essas simulações são capazes de produzir enumerações completas do espaço conformacional, o que permite que a termodinâmica do processo de enovelamento seja totalmente investigada[16]. Os resultados destes experimentos trouxeram novas ideias sobre a formação de estruturas proteicas, sendo a mais importante delas a de que a hidrofobicidade é a força dominante no processo de enovelamento.

Nessa nova visão, a necessidade de se enterrarem os resíduos hidrofóbicos na posição central da proteína é o principal fator determinante da conformação da estrutura nativa. O motivo pelo qual apenas uma estrutura está codificada na sequência vem simplesmente do fato de que há pouquíssimas

maneiras de se enovelar uma proteína no espaço de modo a maximizar o número de contatos hidrofóbicos formados em seu núcleo[11]. Então, ao invés de haver uma relação mais ou menos direta entre aminoácidos e ângulos diedrais, as estruturas secundárias nesse cenário seriam formadas simplesmente como uma consequência do colapso determinado pelo padrão de resíduos polares e hidrofóbicos. Uma das maneiras mais fundamentais como isso ocorre pode ser observada nos frequentes exemplos de α-hélices e folhas β anfipáticas, nas quais uma das faces da estrutura se encontra exposta ao solvente e a outra, protegida dele[7].

Uma importante evidência do protagonismo das interações hidrofóbicas no processo de enovelamento vem da observação de que proteínas normalmente conseguem manter sua estrutura após mutações aleatórias que envolvem substituições de um resíduo hidrofóbico por outro[17]. Além disso, experimentos mostram que sequências artificiais que retêm o mesmo padrão de aminoácidos hidrofóbicos e polares tendem a preservar um enovelamento comum[18–20].

1.4 Enterramentos atômicos

Este trabalho apresenta uma medida relacionada ao efeito hidrofóbico que denominamos **enterramento atômico**, definido simplesmente como a distância de cada átomo da proteína ao centro geométrico da estrutura.

Em um estudo anterior, Pereira de Araújo e colaboradores utilizaram o conceito de enterramentos atômicos para construir um potencial de mecânica molecular para estruturas de proteínas cujo mínimo de energia se situa nas conformações em que os átomos se encontram em seus enterramentos nativos[21]. Utilizando esse potencial em combinação com restrições estruturais gerais, sem o uso de qualquer outra informação derivada da estrutura determinada experimentalmente, foi possível determinar a conformação nativa correta para pequenas proteínas globulares em simulações que partiram de topologias completamente estendidas. Esses resultados demonstraram pela primeira vez que, ao menos para as classes de proteínas analisadas, enterramentos atômicos contêm informação suficiente para a definição da estrutura nativa.

Em um trabalho subsequente[22], proteínas foram divididas em números discretos de camadas concêntricas, de modo que cada camada contivesse o mesmo número de átomos (figura 1.2). Essa divisão em camadas permite uma definição menos precisa de enterramento atômico, que consiste na simples descrição de em qual camada cada átomo se encontra. Utilizando argumentos da teoria da informação de Shannon, a quantidade de informação sobre os enterramentos necessária para categorizar todos os átomos da proteína dessa maneira foi estimada como sendo suficientemente pequena, a ponto de ser codificável na sequência. Foi demonstrado, então, que os enterramentos assim definidos ainda são

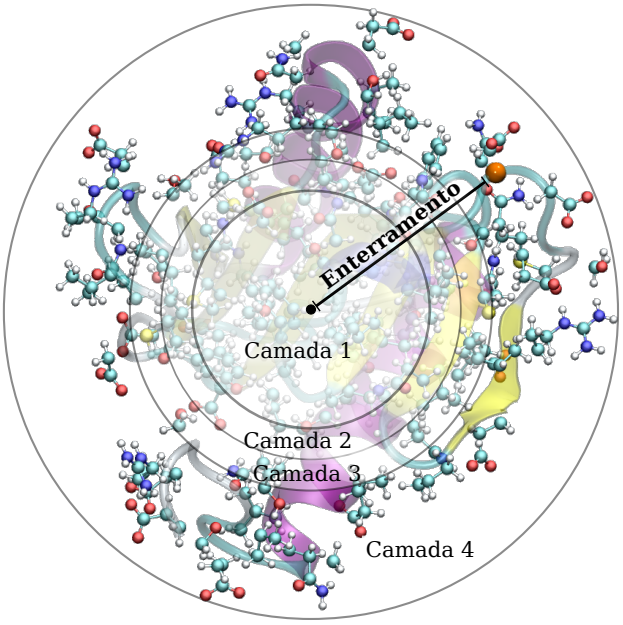


Figura 1.2: Uma proteína globular dividida em quatro camadas. A espessura das camadas é desigual, porque a divisão é projetada de modo que cada camada contenha o mesmo número de átomos.

suficientes para permitir a predição da estrutura nativa completa em simulações computacionais.

Níveis de enterramentos atômicos, portanto, configuram um descritor de estrutura terciária que reúne as propriedades de ser potencialmente previsível a partir da sequência de aminoácidos e de poder ser usado, mesmo quando obtido de maneira imprecisa, para levar a simulação computacional de uma proteína à estrutura nativa. Tomados em conjunto, estes resultados indicam que o enterramento atômico é um excelente candidato a atuar como intermediário em um processo de predição de estruturas de proteínas (figura 1.1).

Em uma analogia com a comunicação humana, os enterramentos podem ser vistos como correspondentes à linguagem na qual as estruturas terciárias estão codificadas na escrita dos aminoácidos. Nessa perspectiva, a decodificação da mensagem requer primeiro que os enterramentos sejam lidos a partir da sequência, de maneira análoga a uma leitura de fonemas a partir do texto escrito, e então que essa informação seja combinada com restrições estruturais independentes de sequência, análogas às regras gramaticais das linguagens humanas, para se chegar à estrutura nativa[22].

1.5 Modelagem computacional de estruturas

Aproximadamente metade das sequências de proteínas conhecidas tem uma relação detectável com a sequência de alguma estrutura que já foi determinada experimentalmente[23]. Quando este é o caso para uma nova proteína cuja estrutura se deseja conhecer, as estruturas homólogas podem ser utilizadas como referência em um processo computacional de predição. Métodos que seguem esse princípio são

conhecidos como técnicas de modelagem por homologia.

A primeira etapa de um um processo de modelagem por homologia é a busca de estruturas conhecidas (chamadas *moldes*) relacionadas à proteína-alvo. A sequência da proteína-alvo é então alinhada às dos moldes, de modo a encontrar as regiões-chave de onde serão extraídas as informações estruturais utilizadas para construir o modelo. A precisão de uma estrutura gerada por métodos de homologia está largamente relacionada à porcentagem de identidade de sequência que ela tem com os moldes usados. Com o uso de sequências-molde com identidade de ao menos 50% em relação à sequência-alvo, já é possível obter modelos com precisão de até 1Å [24]. O software de modelagem por homologia mais comumente utilizado atualmente é o MODELLER [25], cuja metodologia se baseia na busca pela satisfação de restrições espaciais derivadas do alinhamento da estrutura-alvo com os moldes dados.

Quando não há estruturas homólogas conhecidas para a sequência que se deseja prever, a predição de estruturas precisa partir apenas da sequência de aminoácidos dada e do conhecimento acumulado sobre estruturas de proteínas em geral, para se chegar o mais próximo possível da estrutura nativa. Este é o problema conhecido como predição *ab initio* de estruturas de proteínas, um dos temas principais deste trabalho, e um dos maiores desafios atuais da biofísica computacional.

Métodos computacionais para predição *ab initio* de estruturas podem ser classificados em duas categorias principais: métodos baseados em física, que tentam simular da maneira mais precisa possível as forças que atuam em sistemas biológicos durante o processo de enovelamento, e métodos baseados em conhecimento (*knowledge-based*), que utilizam informações obtidas a partir de bancos de dados de estruturas conhecidas para construir algoritmos ajustados por meio de parâmetros derivados estatisticamente.

Desde 1994, o principal indicador global de progresso na área da predição *ab initio* tem sido o CASP [23] (*Critical Assessment of Techniques for Protein Structure Prediction* – Avaliação Crítica de Técnicas para Predição de Estruturas de Proteínas), um evento bienal descrito pelos organizadores como um experimento que visa avaliar a qualidade da performance dos métodos de predição propostos pela comunidade científica. A cada edição do CASP, um conjunto de proteínas é selecionado, cujas estruturas já foram determinadas experimentalmente, mas ainda não divulgadas. Grupos participantes empregam seus métodos para tentar prever as estruturas nativas a partir das sequências de aminoácidos correspondentes, e os resultados destas predições são posteriormente comparados com os dados experimentais. As publicações resultantes dos experimentos CASP procuram destacar tanto os melhores avanços nas metodologias empregadas, quanto os principais gargalos que impedem o progresso na área, de modo a orientar o foco do desenvolvimento de futuros estudos.

No primeiro experimento CASP, dominado por abordagens baseadas em física, nenhum dos mé-

todos avaliados foi capaz de gerar modelos próximos às estruturas nativas para qualquer proteína que não tinha estruturas homólogas conhecidas[26]. Os eventos seguintes testemunharam melhorias progressivas nos resultados, até que, na sexta edição, em 2004, a maioria das proteínas com menos de 100 resíduos já tinham entre os resultados apresentados modelos que pareciam bastante próximos às estruturas experimentais em uma inspeção visual[27].

1.5.1 Rosetta

O avanço mais significativo na qualidade das predições em eventos CASP se deu na terceira edição, em 1998, com a introdução do programa Rosetta[28]. Empregando uma abordagem inovadora baseada em conhecimento, ou seja, que faz amplo uso de informações estatisticamente derivadas de bancos de dados de estruturas conhecidas, o Rosetta foi o primeiro método de predição proposto a ser capaz de resolver a estrutura de pequenas proteínas com enovelamentos completamente novos[29].

O Rosetta representa estruturas de proteínas utilizando um modelo simplificado que inclui apenas os átomos pesados da cadeia principal e o C_β da cadeia lateral. Os comprimentos e ângulos formados pelas ligações covalentes são fixos, e os únicos graus de liberdade da estrutura são os ângulos diedrais da cadeia principal.

O primeiro passo do algoritmo de predição implementado pelo Rosetta é a geração de uma biblioteca de fragmentos de ângulos diedrais da cadeia principal[30]. Esses fragmentos, com comprimento de três ou nove resíduos, são obtidos a partir de um banco de dados de estruturas conhecidas, utilizando parâmetros ajustados de acordo com predições de estruturas secundárias realizadas em sequências homólogas à modelada[29]. O algoritmo de busca consiste em uma simulação de Monte Carlo. Iniciando a partir de uma conformação estendida, a cada passo é selecionado aleatoriamente um fragmento local da cadeia, cujos ângulos torcionais podem ou não ser substituídos pelos ângulos correspondentes em um fragmento da biblioteca com sequência local similar.

A função potencial usada pelo Rosetta inclui termos considerados dependentes de sequência, como os que modelam o efeito hidrofóbico, pontes eletrostáticas, pontes dissulfídicas e estruturas de sequências locais, assim como termos independentes de sequência, que modelam a formação e o empacotamento de estruturas secundárias, a orientação de componentes das folhas β e a repulsão estérica[31].

Versões mais recentes do Rosetta incluem uma etapa posterior de refinamento das estruturas, utilizando um modelo que inclui todos os átomos da proteína. As operações empregadas durante essa etapa consistem principalmente em ajustes finos e otimizações locais nos ângulos da cadeia principal e das cadeias laterais[32].

1.5.2 I-TASSER

Mais recentemente, um segundo método de predição baseado em conhecimento denominado I-TASSER[33;34] tem rivalizado com, e em muitos casos superado[35], os resultados obtidos pelo Rosetta. Assim como ocorre com este, a primeira etapa do processo de predição empregado pelo I-TASSER envolve uma busca em um banco de dados de estruturas conhecidas. Ao invés de buscar fragmentos locais, entretanto, o I-TASSER procura nesta etapa identificar proteínas que possam ter um enovelamento global similar ao da proteína sendo predita. Essas proteínas têm as suas estruturas alinhadas com a sequência-alvo utilizando informações derivadas de um alinhamento de sequências combinado a medidas de similaridade de fragmentos locais e padrões de resíduos hidrofóbicos, assim como valores preditos para estruturas secundárias, ângulos de torção e acessibilidade ao solvente[36].

No modelo simplificado de estrutura proteica adotado pelo I-TASSER, cada resíduo é representado apenas pelo C_α e pelo centro de massa de sua cadeia lateral. A cadeia proteica é dividida em regiões alinhadas e não-alinhadas aos moldes encontrados, sendo que as últimas são modeladas em uma grade cúbica, enquanto as primeiras ocupam posições livres no espaço[37].

A etapa seguinte do processo de predição consiste na geração de modelos completos da cadeia, através da construção de caminhos aleatórios entre os átomos C_α que ligam os fragmentos contínuos de estruturas secundárias detectados para cada alinhamento. Esses modelos são submetidos a simulações de Monte Carlo com um campo de forças que inclui termos derivados de predições de estruturas secundárias e superfície acessível ao solvente, assim como potenciais estatísticos que modelam correlações entre contatos de curta e longa distância[37].

As trajetórias resultantes desta primeira rodada de simulações são agrupadas em *clusters*, de acordo com um critério de similaridade estrutural[38]. Uma segunda rodada, então, é aplicada, partindo apenas das conformações pertencentes ao *cluster* mais populado. Essa segunda rodada tem como objetivo principal remover colisões entre átomos C_α próximos no espaço e não modifica significativamente a topologia das estruturas. Ao final, a estrutura de menor energia é selecionada, e os átomos restantes da cadeia principal e das cadeias laterais são adicionados ao modelo.

1.5.3 Modelagem fisicamente realista

Métodos de predição baseados em simulações fisicamente realistas normalmente são bem menos eficientes, sob o ponto de vista computacional, que abordagens baseadas em conhecimento, como as descritas acima. A adoção de potenciais físicos na simulação do enovelamento tem a vantagem conceitual de, a princípio, possibilitar não apenas a determinação da estrutura nativa final de uma proteína, mas também uma reconstrução precisa da trajetória adotada pelo sistema molecular durante todo o

processo de enovelamento, algo que os métodos baseados em conhecimento não são capazes de obter.

Até recentemente, simulações fisicamente realistas de sistemas proteicos só eram computacionalmente viáveis para escalas tempo inferiores mesmo às requeridas pelas proteínas de enovelamento mais rápido. Em 2008, entretanto, o desenvolvimento de um novo tipo de *hardware* pelo grupo de pesquisa D. E. Shaw Research, com processadores especialmente otimizados para a execução de simulações de dinâmica molecular, permitiu que, pela primeira vez, simulações fisicamente realistas do processo do enovelamento pudessem ser realizadas em escalas de tempo de milissegundos, ou uma ordem de magnitude mais rápida que o anteriormente possível[39]. Utilizando uma versão modificada do campo de força clássico CHARMM[40], o grupo pôde usar essa tecnologia para, pela primeira vez, simular com sucesso o enovelamento de doze proteínas de enovelamento rápido, chegando a estruturas idênticas às nativas[41].

Embora os resultados obtidos por esse estudo não possam competir com as predições de proteínas novas e mais complexas, como as apresentadas pelos métodos que participam do CASP, sua principal contribuição científica é a possibilidade de permitirem o estudo, pela primeira vez em nível de detalhe atomístico, dos detalhes dos mecanismos físicos envolvidos no processo de enovelamento proteico[42].

1.5.4 Método proposto neste trabalho

O algoritmo de predição *ab initio* proposto neste trabalho se encaixa na categoria dos métodos baseados em conhecimento. No entanto, diferentemente das outras abordagens apresentadas, nosso método se fundamenta em uma hipótese muito específica sobre a natureza da linguagem na qual estruturas terciárias são codificadas em sequências, fazendo uma separação metodológica clara entre a informação derivada deste código e as restrições estruturais gerais de proteínas.

Algoritmos como o Rosetta e o I-TASSER são projetados de modo a tentar extrair o máximo possível da informação estatística disponível em bancos de dados de estruturas experimentais, com uma metodologia que combina informações derivadas de vários sub-métodos diferentes para construir potenciais complexos, ajustados automaticamente para cada sequência individual de aminoácidos a ser predita, ao longo de diferentes etapas, com um foco pragmático na obtenção de resultados finais. Por causa dessa complexidade metodológica e da grande dependência de correlações estatísticas empíricas na obtenção do resultado, a relação final entre a sequência de aminoácidos e a estrutura nativa continua obscurecida, mesmo quando esses métodos são bem sucedidos em obter predições com alto grau de precisão.

Nossa proposta, por contraste, parte de uma hipótese simples a respeito da natureza do código do enovelamento, a de que a única informação estrutural codificada nas sequências de aminoácidos são

os enterramentos atômicos. Consequentemente, a relação entre sequências e enterramentos é a única informação estatística obtida a partir de bancos de dados de estruturas conhecidas. Ao invés de usar um potencial com um grande número de parâmetros ajustáveis, de modo a modelar o maior número possível de correlações entre sequência e estrutura, nosso algoritmo emprega um potencial o mais simples possível, cujo único termo dependente de sequência consiste em uma modelagem da tendência de cada átomo a permanecer em sua camada de enterramento predita. Conforme será discutido no capítulo 4, todos os outros termos do potencial independem de qualquer informação derivada estatisticamente e têm como objetivo modelar o menor número possível de interações independentes de sequência requeridas para manter a topologia de uma estrutura nativa.

O grau de sucesso obtido por um método construído de acordo com esses princípios poderá ser traçado diretamente à hipótese que o fundamenta. Este trabalho representa o primeiro passo na construção do que esperamos poder vir a se tornar futuramente uma metodologia completa para a predição *ab initio* de estruturas de proteínas. As predições estruturais resultantes desse método, se bem sucedido, representarão não apenas a obtenção de resultados finais pragmáticos, mas a validação de hipóteses fundamentais a respeito da natureza do mecanismo de codificação de estruturas nativas em sequências de aminoácidos.

1.6 Objetivos

O principal objetivo deste trabalho é estudar a hipótese de que enterramentos atômicos codificados na sequência podem ser a única informação dependente de sequência necessária para determinar a estrutura nativa de proteínas. Para esse fim, será construído um método *ab initio* de predição de estruturas proteicas em que os enterramentos formam o único intermediário informacional entre sequência e estrutura.

Embora a construção de um algoritmo completamente generalizável para a obtenção de resultados de predição esteja além do escopo deste trabalho, esperamos que a metodologia aqui proposta possa servir como a primeira validação de uma abordagem inovadora para o problema.

A contribuição mais imediata pretendida por este trabalho é a de propor e validar um novo arcabouço conceitual para o problema do código do enovelamento. Em longo prazo, esperamos também que a metodologia proposta possa servir como o primeiro passo para a construção de um futuro método de predição que seja competitivo com as abordagens baseadas em conhecimento existentes.

A estrutura deste projeto se divide em duas partes principais, com os seguintes objetivos:

1. Predição de enterramentos atômicos a partir da sequência de aminoácidos.

    1. Desenvolver um algoritmo de predição de enterramentos atômicos a partir de sequências de aminoácidos.
    2. Avaliar a qualidade do algoritmo desenvolvido de acordo com parâmetros da Teoria da Informação.
    3. Construir um banco de dados de predições de enterramento, o qual possa ser usado como base para a seleção de proteínas para simulação.

2. Simulação do enovelamento proteico com potenciais dependentes de sequência derivados da predição de enterramentos atômicos.

    1. Obter um potencial de enterramento atômico derivado de predições realizadas a partir da sequência e estudar a melhor maneira de utilizá-lo em conjunto com potenciais independentes de sequência.
    2. Realizar simulações *ab initio* de modelos geometricamente realistas de proteínas, utilizando os potenciais desenvolvidos.
    3. Avaliar o comportamento das simulações e dos resultados obtidos.

Capítulo 2

Predições de enterramento e análise informacional

Este capítulo inicia a primeira parte deste trabalho, que tem como objetivo principal o desenvolvimento de um algoritmo computacional capaz de prever enterramentos atômicos a partir de sequências de aminoácidos. A maior parte da metodologia e dos resultados descritos a seguir foram publicados pela primeira vez no ano de 2012 em um artigo na revista *Bioinformatics*[43], o qual é parte integrante desta tese.

O algoritmo de predição de enterramentos implementado neste trabalho pertence à categoria conhecida como algoritmos de aprendizado supervisionado, assim denominados porque são capazes de realizar predições em novos conjuntos de dados após passarem por uma chamada etapa de treinamento, na qual um modelo estatístico é construído com base em associações previamente conhecidas entre os dados de entrada e o tipo de dado que se deseja prever[44].

No nosso caso, a base de aprendizado do algoritmo é um banco de dados não-redundante de cadeias proteicas com estruturas determinadas experimentalmente e, portanto, com enterramentos atômicos conhecidos, e o modelo estatístico construído é do tipo conhecido como Modelo Oculto de Markov (HMM - *Hidden Markov Model*). A teoria geral que fundamenta os HMMs é descrita no apêndice 2.2 na página seguinte.

O capítulo inicia com uma breve descrição dos procedimentos usados na construção dos conjuntos de proteínas empregados para treinamento e avaliação do HMM e segue com uma revisão geral do conceito de Modelos Ocultos de Markov. Em seguida, o método de predição implementado é apresentado e os resultados obtidos são discutidos por meio de estudos comparativos entre os valores informados pelo algoritmo e medidas informacionais estimadas para a relação observada entre sequências e enterramentos em estruturas determinadas experimentalmente. A principal conclusão que pode ser

obtida a partir destas análises é que o método de predição desenvolvido é capaz de recuperar quase a totalidade da informação sobre enterramentos atômicos disponível em sequências de aminoácidos.

2.1 Conjuntos de treinamento e avaliação

A base de dados usada neste trabalho partiu da lista de cadeias proteicas compilada pelo PDBSelect [45] na versão de novembro de 2009. O PDBSelect é um seleção de estruturas representativas extraídas do *Protein Data Bank* (PDB) [46], selecionadas de acordo com um critério que minimiza a ocorrência de pares de sequências com mais de 25% de identidade entre si.

Utilizamos quatro critérios adicionais para filtrar a lista fornecida pelo PDBSelect: foram excluídas todas as estruturas que tivessem resolução pior ou igual a 2.5Å, que tivessem sido determinadas por ressonância magnética nuclear (ao invés de difração de cristais em raios-X), que fossem descritas como proteínas de membrana, ou que não atendessem a um critério de globularidade, conforme descrito por Gomes e colaboradores [47]. A lista de proteínas resultante está disposta no apêndice B na página 75.

Metade das proteínas desse banco foi aleatoriamente atribuída a um conjunto de treinamento, utilizado para construir os modelos estatísticos do HMM; a outra metade constituiu o conjunto de avaliação, para o qual todas as predições foram realizadas.

2.2 Processos de Markov e o Modelo Oculto de Markov (HMM)

Em vários campos das ciências e das engenharias, são encontradas variações do problema de se tentar extrair, através de métodos computacionais, a informação oculta por trás de uma sequência de dados observados. Exemplos incluem a determinação das sequências de fonemas que melhor explicam a emissão de determinados padrões sonoros (reconhecimento de voz), ou a de quais letras têm a sequência de traços com a maior probabilidade de gerar uma imagem analisada (reconhecimento de caligrafia). Situações como estas vêm sendo modeladas e resolvidas com elevado grau de sucesso através de um mecanismo conhecido como Modelo Oculto de Markov (*Hidden Markov Model*, ou HMM) [48].

Na Biologia Estrutural e na Bioinformática, HMMs têm encontrado aplicações em áreas como a busca de padrões de estruturas secundárias em proteínas, a construção de alinhamentos múltiplos de sequências e a identificação de famílias proteicas. A teoria que fundamenta os HMMs se baseia no conceito de processos de Markov discretos. Um processo de Markov discreto é um sistema que, a cada momento, pode estar em um determinado estado entre um conjunto preestabelecido de N estados possíveis, movendo-se de um estado para outro a cada intervalo fixo de tempo. A característica distintiva de um processo de Markov é que o sistema sempre decide o próximo estado a ser visitado

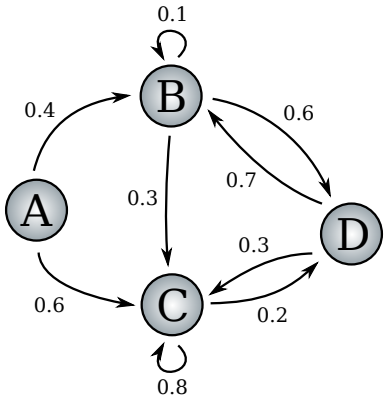


Figura 2.1: Exemplo de um processo de Markov com 4 estados. Cada círculo representa um estado, as setas representam as transições entre os estados, e os números, as probabilidades de cada transição. Por exemplo, uma vez que o sistema se encontra no estado B, ele tem uma probabilidade de 0.3 de se mover ao estado C, uma probabilidade de 0.6 de se mover ao estado D, e uma probabilidade de 0.1 de se manter onde está.

de acordo com um uma distribuição de probabilidades que depende apenas do estado em que ele se encontra no momento, independentemente do caminho anteriormente percorrido. A Figura 2.1 mostra, como exemplo, um HMM simples com 4 estados.

As probabilidades de transição que compõem um processo de Markov podem ser representadas por uma matriz $A = \{a_{ij}\}$ com $i,j \leq N$, sendo que a_{ij} representa a probabilidade de o sistema fazer uma transição do estado i ao estado j. A Equação 2.1 ilustra a matriz A correspondente ao sistema disposto na Figura 2.1.

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.4 & 0.1 & 0 & 0.7 \\ 0.6 & 0.3 & 0.8 & 0.3 \\ 0 & 0.6 & 0.2 & 0 \end{bmatrix} \tag{2.1}$$

Na modelagem de sistemas utilizando Modelos Ocultos de Markov (HMMs), supõe-se que existe um processo de Markov como o descrito acima atuando por trás do sistema observado, com a característica adicional de que, cada vez que o sistema efetua uma transição entre estados, um *símbolo* visível é emitido, de um alfabeto predeterminado de M possíveis símbolos. A probabilidade de o sistema emitir o símbolo k quando o sistema se encontra no estado j é dada pela distribuição de probabilidades $B = \{b_j(k)\}$, com $1 \leq k \leq M$ e $1 \leq j \leq N$. Deve ser definida também uma distribuição de probabilidades de estados iniciais $\pi = \{\pi_i\}$, com $1 \leq i \leq N$, onde π_i indica a probabilidade do processo de Markov iniciar no estado i.

Em sistemas modelados por HMMs, tudo o que podemos observar diretamente é a sequência de símbolos emitidos pelo sistema; as transições entre estados são opacas e normalmente correspondem à informação oculta que objetivamos extrair. Por exemplo, em sistemas de reconhecimento de voz,

que são uma das principais aplicações conhecidas de HMMs, os símbolos emitidos correspondem aos sons registrados (informação observada), e os estados ocultos, às letras que representam os fonemas correspondentes (informação que se deseja extrair).

No sistema implementado neste trabalho, os símbolos emitidos são as identidades de aminoácidos, e os estados ocultos são sequências locais de enterramentos atômicos (figura 2.2 na página 19).

2.2.1 O Algoritmo *Forward-Backward*

Um dos problema fundamentais que podemos querer resolver com a aplicação de HMMs pode ser formulado da seguinte maneira: dado o HMM definido por $\lambda = (A,B,\pi)$ e a sequência de dados observados $O = O_1O_2 \ldots O_T$, qual é a sequência de estados $Q = q_1q_2 \ldots q_T$ em que cada estado q_t tem a máxima probabilidade de ter sido visitado no tempo t, durante a geração da sequência O? Esta pergunta pode ser resolvida através da aplicação do procedimento conhecido como *forward-backward*[48]:

Considere a variável "forward" $\alpha_t(i)$, que descreve a probabilidade de o sistema se encontrar no estado S_i (sendo $1 \leq i \leq N$) quando o t-ésimo símbolo observado tiver sido emitido:

$$\alpha_t(i) = P(O_1O_2 \ldots O_t, q_t = S_i|\lambda) \tag{2.2}$$

ou seja, $\alpha_t(i)$ é a probabilidade de se observar a sequência de símbolos $O_1O_2 \ldots O_t$, sendo que, no momento em que o último símbolo é emitido, o sistema se encontra no estado S_i.

Para $t = 1$ é fácil obter o valor dessa probabilidade pela simples multiplicação da probabilidade inicial de cada estado S_i pela probabilidade conhecida de o símbolo observado ter sido emitido naquele estado:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \leq i \leq N \tag{2.3}$$

Para todos símbolos seguintes, então, basta considerar, por indução:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1}), \qquad 1 \leq t \leq T-1, 1 \leq j \leq N \tag{2.4}$$

A variável "backward" funciona de maneira análoga: $\beta_t(i)$ representa a probabilidade de todo o restante da sequência ser emitido a partir da posição t, uma vez que, nesse momento, o sistema se encontra no estado S_i:

$$\beta_t(i) = P(O_{t+1}O_{t+2} \ldots O_T|q_t = S_i,\lambda) \tag{2.5}$$

Para resolver esse caso, definimos que esta probabilidade é sempre 1 para o último símbolo:

$$\beta_T(i) = 1, \qquad 1 \leq i \leq N \tag{2.6}$$

Então, os símbolos anteriores podem ser resolvidos por indução:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j), \qquad t = T-1, T-2, \ldots, 1,\ 1 \leq j \leq N \tag{2.7}$$

Para qualquer valor t, então, a variável $\alpha_t(i)$ nos dá a probabilidade de, partindo do estado inicial, o sistema se encontrar no estado S_i quando o t-ésimo símbolo for observado e $\beta_t(i)$ nos dá a probabilidade de, uma vez que o sistema está nesse ponto, observarmos o restante da sequência, até o último símbolo. Multiplicando os dois valores e normalizando as probabilidades, temos $\gamma_t(i)$, que é a probabilidade o sistema estar em S_i quando o t-ésimo símbolo for observado, considerando toda a sequência:

$$\gamma_t(i) = \frac{\alpha_t(i)\, \beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\, \beta_t(i)} \tag{2.8}$$

Como normalmente estamos interessados apenas em saber que valor de i maximiza $\gamma_t(i)$ para cada t, o denominador da equação 2.8 costuma ser ignorado nos cálculos computacionais.

2.2.2 Metodologia de predição

Foi desenvolvido para este trabalho um programa de predição de enterramentos atômicos em camadas a partir de sequências de aminoácidos, denominado **HmmPred**. O algoritmo implementado pelo HmmPred é modelado a partir da generalização do método de predição de estruturas secundárias descrito por Crooks e Brenner em 2004[14]. O programa foi escrito em linguagem C++, e o código-fonte está disponível sob uma licença de *software* livre na página do Laboratório de Biofísica Teórica e Computacional*.

Duas observações estatísticas sobre códigos do enovelamento formam a base do modelo de dados empregado pelo HmmPred. Primeiramente, sabe-se que a correlação entre a identidade de aminoácidos vizinhos em sequências proteicas é praticamente desprezível, a ponto de, na prática, sua distribuição poder ser considerada estatisticamente independente[14]. Por outro lado, é fácil perceber que os enterramentos de aminoácidos pertencentes à mesma sequência local são fortemente correlacionados entre si[43;49]: dados dois aminoácidos conectados por uma ligação peptídica, pode-se esperar com considerável probabilidade que eles estejam enterrados na mesma camada ou em camadas vizinhas; a

*`http://www.lbtc.unb.br/software/`

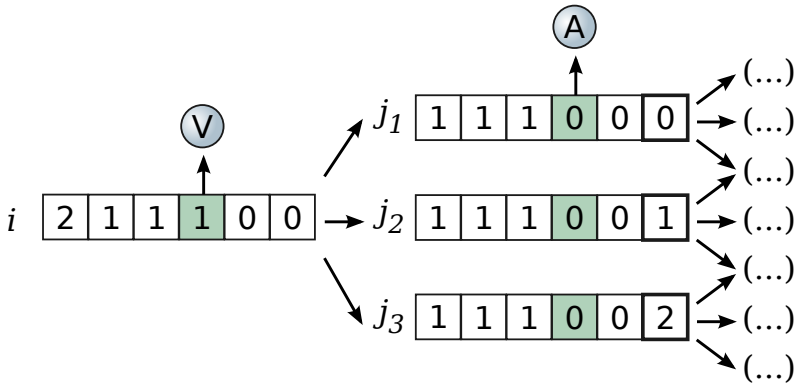


Figura 2.2: No HmmPred, cada estado corresponde a uma sequência de $f-1$ símbolos, representando o padrão de enterramentos discretos de um fragmento de aminoácidos. Cada símbolo é implementado como um número entre 0 e $L_Y - 1$, onde L_Y é o número de camadas em que a proteína foi dividida. Cada estado tem apenas L_Y sucessores possíveis. No exemplo desta figura, $f = 7$ e $L_Y = 3$.

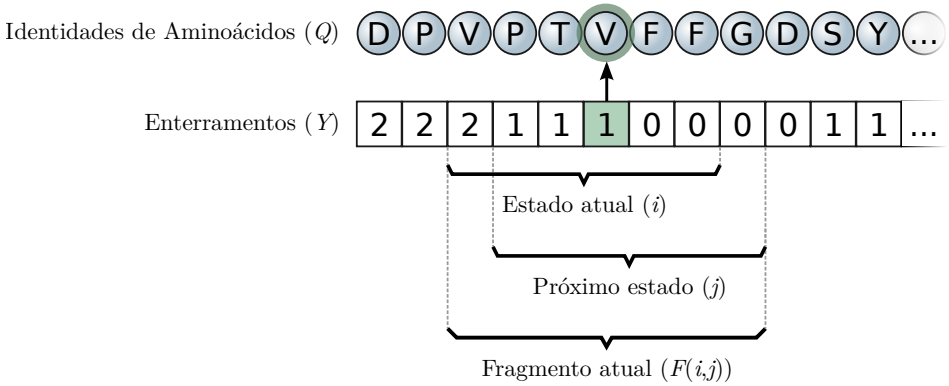


Figura 2.3: A probabilidade de transição de um estado i para um estado j está relacionada à probabilidade de ocorrência do fragmento $F_{i,j}$, que inclui ambos os estados (equação 2.9).

probabilidade de que eles estejam separados por duas ou mais camadas é progressivamente menor.

No texto que segue, a palavra símbolo se refere a um elemento discreto pertencente a um conjunto finito chamado alfabeto. O alfabeto de símbolos observáveis que podem ser emitidos pelo HMM implementado corresponde ao conjunto $\mathcal{Q} = \{\chi_1, \cdots, \chi_{L_\mathcal{Q}}\}$, composto pelos 20 aminoácidos comuns ($L_\mathcal{Q} = 20$), ou por alfabetos reduzidos que serão vistos adiante. Adicionalmente, define-se um alfabeto de variáveis ocultas $\mathcal{Y} = \{y_1, \cdots, y_{L_\mathcal{Y}}\}$; no caso mais simples, $L_\mathcal{Y}$ é igual ao número de camadas de enterramento em que a proteína é dividida, e cada y_n representa a n-ésima camada. Outros possíveis alfabetos para as variáveis ocultas serão discutidos adiante. Uma das entradas do algoritmo é f, que especifica o tamanho do fragmento de resíduos considerado no modelo (normalmente definido entre 3 e 9).

Um estado do HMM é formado por uma sequência de símbolos de enterramento. Cada estado corresponde a uma sequência de $f-1$, símbolos, para um total de $T_S = (L_\mathcal{Y})^{f-1}$ possíveis estados. A probabilidade de transição entre dois estados reflete a frequência em que as duas sequências correspondentes se sobrepõem no conjunto de treinamento. Por exemplo, para $L_\mathcal{Y} = 3$, $f = 7$, um estado

que corresponde à sequência de enterramentos $i \rightarrow [211100]$ tem apenas três possíveis sucessores: $j_1 \rightarrow [111000]$, $j_2 \rightarrow [111001]$ e $j_3 \rightarrow [111002]$ (Figura 2.2). Quando os estados i e j se sobrepõem, definimos $F_{i,j}$ como a sequência de f símbolos que abarca i e j (Figura 2.3). Neste caso, os valores da matriz de probabilidades de transição $A = \{a_{ij}\}$ são dados por:

$$a_{ij} = \frac{p\left(F_{i,j}\right)}{p(i)p(j)} \tag{2.9}$$

onde $p\left(F_{i,j}\right)$, $p(i)$ e $p(j)$ são estimados por uma simples contagem de frequência das sequências correspondentes no conjunto de treinamento. Para todos os pares i,j que não se sobrepõem, $a_{ij} = 0$. As probabilidades são normalizadas de modo que, para todos os valores de i, $\sum_j a_{ij} = 1$.

As probabilidades de emissão de observáveis são uma função das transições entre estados (e não apenas do estado em que o sistema se encontra, como ocorre na definição tradicional de um HMM). A distribuição de probabilidades de emissão é definida como $B = \{b_{ij}(k)\}$, onde $b_{ij}(k)$ é a probabilidade de que o símbolo k seja emitido quando o sistema se move do estado i para o estado j. Sendo Q e Y, respectivamente, as sequências de símbolos de estrutura primária e de enterramentos e $h = (f-1)/2$, temos então:

$$b_{ij}(k) \;=\; \sum_{i=1}^{Ts}\sum_{j=1}^{Ts} \underbrace{p(Q_t = k|Y_{[t-h,t+h]} = F_{i,j})}_{\#1} + \underbrace{p(Q_t = k|Y_t = F_{i,j(h+1)}) \cdot Ps}_{\#2}, \forall t \tag{2.10}$$

em que a o termo #1 se refere à probabilidade de que o símbolo k seja encontrado na sequência primária, desde que o fragmento de enterramentos $F_{i,j}$ esteja centralizado em sua posição, e o termo #2, à probabilidade de que k seja encontrado, dado apenas o símbolo de enterramento correspondente à posição central do mesmo fragmento. Ambos os termos são estimados a partir de contagens de frequências no conjunto de treinamento. $Ps = 20$ é uma pseudocontagem. A inclusão do termo #2 tem o efeito de permitir que $b_{ij}(k)$ seja estimado apenas a partir do símbolo central de $F_{i,j}$, quando não existem dados suficientes na amostra do conjunto de treinamento para que o valor seja estimado a partir do fragmento completo (termo #1), algo que pode ocorrer principalmente como decorrência da saturação de dados estatísticos no banco de treinamento, por exemplo, para tamanhos de fragmento muito grandes ou para alfabetos mais complexos. As probabilidades são normalizadas de modo que $\sum_k b_{ij}(k) = 1$ para todas as combinações de i e j.

A topologia descrita acima para o HMM e ilustrada nas figuras 2.2 e 2.3 é uma consequência direta dos dois pressupostos anteriormente adotados: o fato de aminoácidos em proteínas serem distribuídos de maneira estatisticamente independente justifica a escolha de os símbolos observáveis corresponde-

rem a identidades de aminoácidos consideradas isoladamente; o fato de enterramentos serem correlacionados entre si justifica a escolha de estados do HMM corresponderem a fragmentos de símbolos de enterramento vizinhos na sequência.

A etapa de treinamento do HmmPred consiste em utilizar as frequências encontradas no conjunto de treinamento para definir todos os valores das distribuições $A = \{a_{ij}\}$ e $B = \{b_{ij}(k)\}$, conforme descrito acima. Em seguida, na etapa de predição, esses valores são utilizados como entrada no algoritmo *forward-backward*, descrito anteriormente, para cada sequência de aminoácidos cujos enterramentos se desejam prever.

A saída do *forward-backward* é uma série de valores $\gamma_t(i)$, que são as probabilidades de que o sistema esteja no estado i quando o símbolo observado na posição t foi emitido, para todos os valores válidos de t e i. É importante ressaltar que $\gamma_t(i)$ corresponde apenas ao estado individualmente mais provável para a emissão de cada símbolo, não sendo, por si só, informativo da sequência de estados que mais provavelmente gerou toda a cadeia de símbolos observados. Essa sequência poderia ser obtida a partir de todos os valores $\gamma_t(i)$ por meio de um procedimento recursivo conhecido como algoritmo de Viterbi[48]. Neste trabalho, como estamos interessados em maximizar o número de resíduos preditos para as camadas corretas em cada posição, o que não necessariamente corresponde à sequência de camadas mais provável globalmente, o algoritmo de Viterbi não é utilizado.

Sendo que cada estado corresponde a uma sequência de enterramentos, e estamos interessados apenas no enterramento mais provável para a posição central de cada estado, o passo final, é calcular $p_t(y)$, a probabilidade de encontrar o símbolo de enterramento y na posição central de um fragmento de enterramentos que possa corresponder à posição t. Este valor é encontrado através do somatório das probabilidades de todos os valores de $\gamma_t(i)$ em que a posição central do fragmento i corresponde a y:

$$p_t(y) = \sum_{i=1}^{Ts} (\gamma_t(i) \mid [i]_{h+1} = y) \tag{2.11}$$

onde $[i]$ é a sequência de símbolos de enterramento correspondente ao estado i.

Enfim, para cada posição t da sequência dada, o algoritmo prevê a correspondência de um símbolo de enterramento y para o qual $p_t(y)$ tem o valor mais alto.

2.2.3 Alfabetos

O alfabeto de variáveis ocultas $\mathcal{Y}$ é usado para construir os fragmentos que compõem os estados ocultos do HMM e, portanto, define os tipos de dados que são reconhecidos pelo modelo na etapa de treinamento e previstos a partir da sequência na etapa de predição. Este alfabeto pode ser construído

de três maneiras diferentes:

1. Alfabetos numéricos simples, em que cada dígito representa uma camada de enterramento, começando em 0 (para duas camadas, $\mathcal{Y} = \{0,1\}$; para três, $\mathcal{Y} = \{0,1,2\}$, etc...). Esse é o alfabeto utilizado no exemplo das figuras 2.2 e 2.3.

2. Alfabetos que representam o enterramento do C_α em camadas, mais o enterramento relativo do C_β (ou o oposto). Esta é uma maneira de representar a orientação relativa da cadeia lateral. Por exemplo, para duas camadas, podemos ter $\mathcal{Y} = \{\alpha_0\beta_\uparrow, \alpha_0\beta_\downarrow, \alpha_1\beta_\uparrow, \alpha_1\beta_\downarrow\}$, em que $\alpha_0\beta_\uparrow$ indica um resíduo em que o C_α está na camada 0 (mais interna) e o C_β está menos enterrado que o C_α.

3. Alfabetos que combinam o enterramento com informação sobre a estrutura secundária. Por exemplo, para duas camadas, temos $\mathcal{Y} = \{0^H, 0^E, 0^L, 1^H, 1^E, 1^L\}$, em 1^H que representa um resíduo que se encontra na camada 1 (mais externa) e faz parte de uma α-hélice ($H \rightarrow \alpha$-hélice; $E \rightarrow$ folha β; $L \rightarrow$ região de loop).

É importante ressaltar que, mesmo quando se usam os alfabetos dos tipos (2) e (3) para o treinamento do modelo, nenhuma informação sobre a estrutura é fornecida ao algoritmo durante a fase de predição. Assim, por exemplo, no alfabeto mencionado no tipo (3), a saída do algoritmo pode ser lida não apenas como uma predição dos níveis de enterramento de cada resíduo, mas também como uma predição de estruturas secundárias, ambas obtidas apenas a partir da sequência.

Para o alfabeto de estrutura primária, além da versão completa com 20 letras, podem ser usadas também duas simplificações: HP, em que os resíduos são classificados em apenas hidrofóbicos (H) ou polares (P) , sendo H={A, C, F, G, I, L, M, V, W, Y} e P={D, E, H, K, N, P, Q, R, S, T}; ou HPN, em que resíduos são classificados como hidrofóbicos (H), polares (P) ou neutros (N), sendo N={A, G, H, S, T}.

2.3 Avaliação dos resultados

A qualidade das predições de enterramento realizadas por um método como o HmmPred pode ser avaliada primeiramente através do cálculo da proporção de resíduos corretamente classificados (n_c), em relação ao número total de resíduos avaliados (n_t):

$$A = \frac{n_c}{n_t} \tag{2.12}$$

Essa medida, que denominamos **acurácia**, é o parâmetro mais simples e direto que pode ser

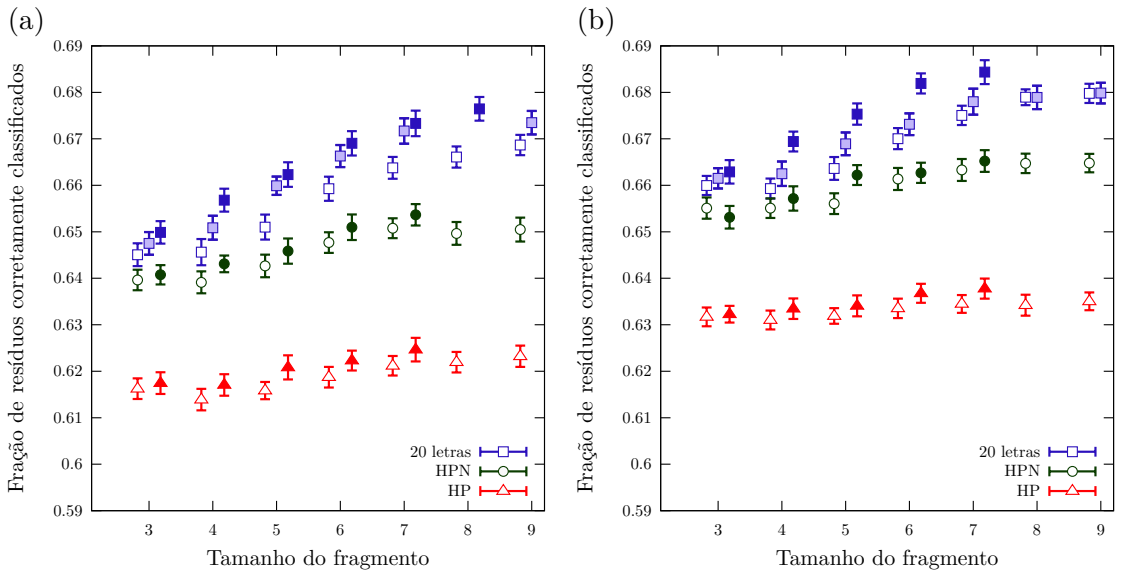


Figura 2.4: Proporção de resíduos corretamente classificados para a predição em duas camadas. (a) Predição para C_α; (b) Predição para C_β. Pontos vazios, semipreenchidos e preenchidos representam, respectivamente, os alfabetos dos tipos 1, 2 e 3, descritos na seção 2.2.3. Formatos diferentes (□,○,△) representam diferentes alfabetos de estrutura primária, conforme legenda.

calculado para a avaliação de precisões discretas, podendo funcionar como medida de comparação entre diferentes predições realizadas com representações similares de entrada e saída.

Para cálculo dos erros estatísticos deste e dos outros parâmetros de avaliação que serão apresentados adiante, o HmmPred suporta que as predições sejam realizadas repetidas vezes através de uma operação conhecida em estatística como *bootstrapping*[50]. Na nossa implementação, são realizadas 50 rodadas de predições, cada uma com um banco de treinamento reconstruído de modo que cada proteína tenha um peso aleatório atribuído de acordo com uma distribuição de Poisson, com $\lambda = 1$. O desvio médio entre os resultados dessas repetições é usado como estimativa do desvio da amostra que compõe o conjunto de treinamento em relação à população total de sequências da qual ele é derivado.

Os resultados obtidos pelo HmmPred foram inicialmente avaliados a partir de sua performance na predição de enterramentos em duas camadas, pois este é o número para o qual foi possível obter os valores mais precisos para as estimativas informacionais que serão usadas como parâmetros de comparação. Mais adiante, a discussão será estendida para predições realizadas com números maiores de camadas.

A figura 2.4 representa a acurácia dos resultados obtidos para a predição de C_α (2.4-a) e C_β (2.4-b), em função do tamanho de fragmento f para valores entre 3 e 9 (eixos horizontais). Os valores de f são números inteiros; alguns pontos foram ligeiramente deslocados no eixo horizontal para facilitar a visualização.

Diferentes formatos de símbolo (□, ○ ,△) foram usados nessa figura para representar os diversos alfabetos usados para codificar a estrutura primária das proteínas (20 letras, HPN e HP, respectiva-

mente). Em relação às variáveis ocultas usadas nos estados do HMM, símbolos vazios representam o alfabeto simples de enterramento (tipo 1); símbolos semipreenchidos, o alfabeto que inclui a direção da cadeia lateral (tipo 2); símbolos totalmente preenchidos, o alfabeto que inclui estruturas secundárias (tipo 3). Omissões no gráfico para algumas pontos com f entre 7 e 8 correspondem a configurações em que não foi possível realizar as respectivas predições, devido a limitações computacionais relacionadas à quantidade de memória requerida para execução.

O primeiro padrão claramente visível na figura 2.4 é o fato de que a acurácia de predição cresce com o aumento do tamanho do fragmento, pelo menos até valores aproximadamente iguais a 6 ou 7. Esse aumento é condizente com o pressuposto de que o contexto local dos enterramentos tem um papel relevante no código do enovelamento. O fato de a acurácia melhorar mais lentamente a partir do uso de fragmentos maiores que 7 pode significar que este é o limite aproximado da informação local útil sobre enterramentos disponível em sequências, ao menos quando se utiliza um banco de treinamento finito, cuja disponibilidade de informações estatísticas pode se tornar saturada. Os melhores resultados de predição alcançados ficaram em torno de 67.5% e 68.5% para a predição do enterramento de C_α e C_β, respectivamente.

Comparando os três alfabetos usados para representar a estrutura primária, observa-se, como esperado, que a representação em 20 letras produz a melhor acurácia, seguida pelas representações HPN e HP. É notável, entretanto, que a diferença de qualidade entre a representação com 20 letras e a que utiliza o alfabeto HPN é menor que entre a diferença observada entre os alfabetos HPN e HP.

Uma outra tendência observada em todos os casos dispostos na figura é que a qualidade de predição do enterramento de C_β é consistentemente melhor que a do enterramento de C_α. Possíveis razões para esse padrão serão exploradas na seção seguinte.

A comparação entre alfabetos utilizados para as variáveis ocultas revela que ambas as variações que incluem informações estruturais adicionais ao enterramento de um átomo (tipos 2 e 3) produziram valores de acurácia melhores que aqueles obtidos com apenas os enterramentos (tipo 1). O alfabeto tipo 2 foi usado apenas nas predições com 20 letras, e seus resultados são sempre intermediários aos alcançados pelos tipos 1 e 3.

2.3.1 Análise informacional

A medida de acurácia dos resultados, embora possa ser estudada em uma análise relativa da importância das diferentes dimensões de parâmetros do algoritmo, tem duas limitações importantes. Primeiro, ela não permite fazer nenhuma avaliação sobre a qualidade absoluta dos resultados, isto é, indicar o quão eficiente é o algoritmo, independentemente de seus parâmetros, em extrair a informação sobre

enterramentos que está disponível nas sequências. Em segundo lugar, a acurácia não pode ser comparada para predições realizadas com números diferentes de camadas. Por exemplo, para uma predição em duas camadas equiprováveis, $A = 50\%$ corresponde a um resultado não melhor que uma atribuição aleatória de valores; o mesmo valor, entretanto, pode representar um bom resultado em uma predição realizada com quatro ou mais camadas.

Um tipo de dado mais esclarecedor que a acurácia, então, é fornecido por análises baseadas na Teoria da Informação de Shannon[51], as quais podem ser comparadas com limites superiores de qualidade derivados das relações observadas entre enterramentos e sequências em estruturas determinadas experimentalmente.

O primeiro dos parâmetros de qualidade que podemos calcular dessa maneira é a transinformação entre enterramentos reais e preditos, $I(Y;Y(Q))$, medida em bits*. Esta é a diferença entre a entropia simples $H(Y)$ da sequência de enterramentos nativos Y e a entropia desta, condicional à sequência de enterramentos preditos a partir da sequência Q de aminoácidos, $H(Y|Y(Q))$:

$$I(Y;Y(Q)) = H(Y) - H(Y|Y(Q)), \tag{2.13}$$

Neste trabalho, chamaremos essa medida de **informação da predição**, pois ela pode ser interpretada como uma medida do quanto da informação a respeito dos enterramentos codificados na sequência o algoritmo é capaz de extrair. Uma vez que as camadas são equiprováveis, a entropia não-condicional $H(Y)$ é igual a simplesmente $\log_2 L$, onde L é o número de camadas. De acordo com a definição de entropia condicional, $H(Y|Y(Q))$ é dado por

$$H(Y|Y(Q)) = \sum_{n\in\mathcal{Y}} \sum_{r\in\mathcal{Y}} p(Y = n, Y(Q) = r)\ \log_2 p(Y = n|Y(Q) = r) \tag{2.14}$$

onde $p(Y = n|Y(Q) = r)$ representa a probabilidade condicional de que o enterramento real n seja encontrado na estrutura nativa, dada o enterramento predito r. Esse valor corresponde simplesmente a $p(y_n)$, ou seja, a probabilidade predita pelo algoritmo para o enterramento observado y_n. Sendo o fator $p(Y = n, Y(Q) = r)$ necessário na equação 2.14 apenas para normalizar as probabilidades, temos então que ela é equivalente a:

$$H(Y|Y(Q)) = -\frac{1}{M} \sum_{i=1}^{M} \log_2 p(y_n), \tag{2.15}$$

onde M é o número total de resíduos no banco de dados.

*Para uma breve revisão sobre as definições de Teoria da Informação usadas neste capítulo, ver o apêndice A na página 73.

Uma vez que o HMM trabalha com probabilidades para fragmentos de enterramentos, e não apenas para posições individuais, também se pode considerar uma aproximação para o cálculo da *densidade* de informação da predição, $i(Y;Y(Q))$, isto é, o quanto a informação da predição aumenta, em bits, para cada novo resíduo considerado, quando já se conhecem os enterramentos dos resíduos anteriores. Sendo Y^N uma sequência de N enterramentos:

$$\begin{aligned} i(Y;Y(Q)) &= \lim_{N\to\infty} \frac{I(Y^N;Y^N(Q))}{N} \\ &\approx \frac{I(Y^N;Y^N(Q)) - I(Y^1;Y^1(Q))}{N-1} \\ &= h_N(Y) - h_N(Y|Y(Q)) \end{aligned}$$

com

$$h_N(Y) = \frac{H(Y^N) - H(Y^1)}{N-1} \tag{2.16}$$

e

$$h_N(Y|Y(Q)) = \frac{H(Y^N|Y^N(Q)) - H(Y^1|Y^1(Q))}{N-1}. \tag{2.17}$$

Uma vez que o enterramento predito, $Y(Q)$, é derivado da sequência de aminoácidos Q, a desigualdade do processamento de dados[52] estabelece que a predição da informação $I(Y;Y(Q))$ não pode ser maior que a transinformação entre sequências e valores individuais de enterramento, $I(Y;Q)$, isto é, $I(Y;Y(Q)) \leq I(Y;Q)$. De forma análoga, para as respectivas densidades de transinformação, $i(Y;Y(Q)) \leq i(Y;Q)$.

As figuras 2.5 e 2.6, formatadas de acordo com as mesmas convenções da figura 2.4, indicam no eixo vertical, respectivamente, o valor da predição da informação e da densidade de informação da predição para os mesmos resultados apresentados anteriormente. Essas figuras indicam também, como linhas horizontais adicionais, o valor dos limites superiores correspondentes $I(Y;Q)$ e $i(Y;Q)$, de acordo com estimativas calculadas por Juliana Rocha[49] como parte de uma dissertação de mestrado desenvolvida paralelamente a esta tese.

De maneira geral, os mesmos padrões anteriormente observados na análise da acurácia são confirmados por essas figuras. Além disso, elas trazem novas informações que podem ser usadas para avaliar a qualidade geral dos resultados encontrados.

Primeiramente, analisaremos a figura 2.5, que mostra a informação da predição para os resultados obtidos, assim como os limites superiores correspondentes estabelecidos por $I(Y;Q)$.

O algoritmo de predição implementado neste trabalho parte de um modelo de dados que destaca

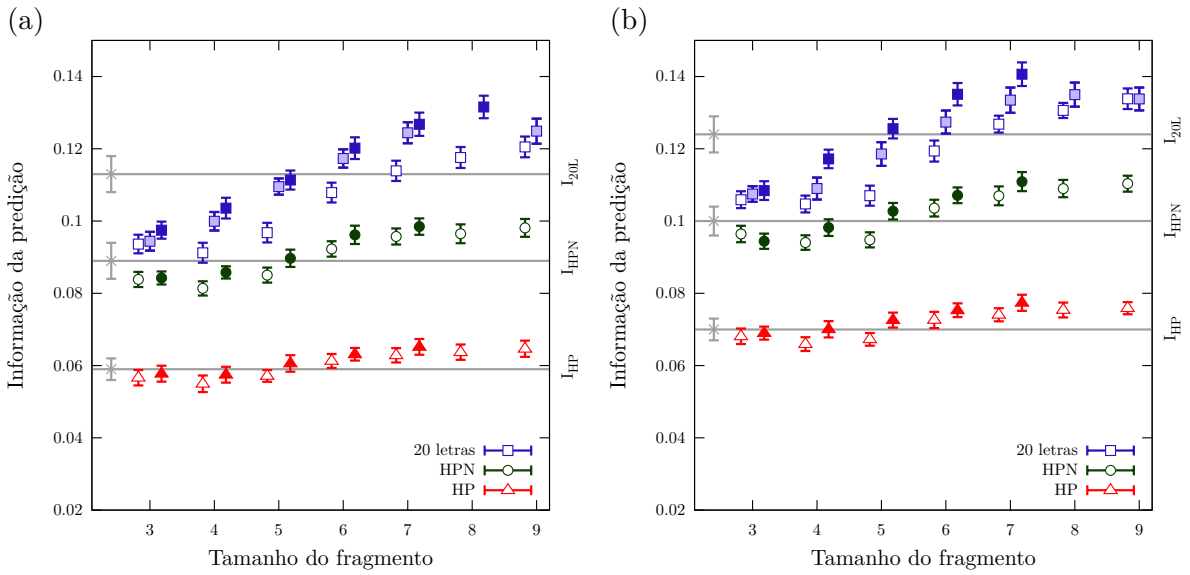


Figura 2.5: Informação da predição da para duas camadas de enterramento. (a) Predição para C_α; (b) Predição para C_β. As linhas horizontais indicam $I(Y;Q)$, o limite máximo da transinformação entre a sequência de aminoácidos e cada posição de enterramento considerada isoladamente, isto é, sem o contexto dos enterramentos vizinhos.

a importância das correlações entre enterramentos de aminoácidos vizinhos, modelados como estados ocultos compostos por sequências de símbolos de enterramento. As correlações entre identidades de resíduos vizinhos são ignoradas, por serem consideradas estatisticamente irrelevantes. Uma vez que o HmmPred utiliza correlações entre enterramentos vizinhos como parte fundamental de sua metodologia, podemos ver na figura 2.5 que ele é capaz de alcançar valores de informação da predição entre 0.13 e 0.14 bits mais altos que os valores estimados para $I(Y;Q)$, principalmente para as configurações que empregaram os alfabetos dos tipos 2 e 3 e para fragmentos de ao menos 6 ou 7 resíduos. A superação desses limites confirma a importância do modelo de dados adotado, pois ela implica que os resultados do algoritmo são superiores aos que poderiam ser obtidos por qualquer método de predição que se baseasse apenas em correlações individuais entre identidades de resíduos e enterramentos.

Um indicador adequado para os limites superiores de predição de um algoritmo como o HmmPred, então, é dado pela densidade de informação da predição, $i(Y;Y(Q))$, isto é, a quantidade de nova informação sobre os enterramentos que pode ser extraída da sequência a cada novo resíduo lido, considerando o conhecimento dos enterramentos de resíduos anteriores. De maneira análoga ao caso anterior, os possíveis valores de $i(Y;Y(Q))$ são limitados teoricamente pelos valores estimados para $i(Y;Q)$, que é a densidade de transinformação entre enterramentos e sequências.

Conforme mostrado na figura 2.6, os resultados do HmmPred se aproximam do valor estimado para essa medida, dentro dos limites de erro. Esses resultados demonstram que o algoritmo é capaz de extrair praticamente toda a informação sobre enterramentos que está disponível na sequência de aminoácidos, de acordo com o modelo adotado. Em outras palavras, nenhuma outra metodologia

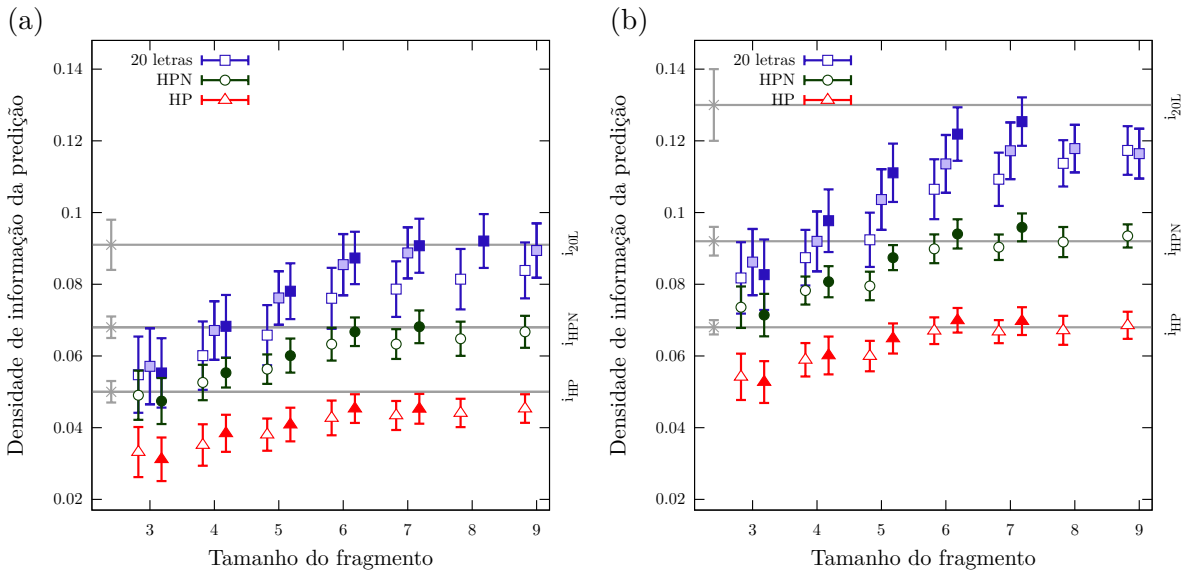


Figura 2.6: Densidade da informação da predição para duas camadas de enterramento. (a) Predição para C_α; (b) Predição para C_β. As linhas horizontais indicam $i(Q;Y)$, o limite máximo para a densidade de transinformação entre a sequência de aminoácidos e os enterramentos.

de predição de enterramentos que partisse dos mesmos pressupostos estatísticos considerados pelo HmmPred poderia, em princípio, obter resultados melhores que os apresentados, independentemente do algoritmo empregado. Nesse sentido, portanto, podemos dizer que o algoritmo de predição implementado pelo programa é ótimo.

A comparação entre os valores informacionais estimados e preditos para as predições de C_α e C_β também pode trazer algumas ideias sobre por que estes são preditos com acurácia consistentemente maior que aqueles. A incerteza relativa de enterramentos vizinhos, expressa pela densidade de entropia $h(Y)$ já é naturalmente mais alta para C_β que para C_α [43], provavelmente por consequência do fato de que carbonos β são unidos na estrutura terciária por um número maior de ligações covalentes. A predição do HmmPred ainda reduz essa incerteza de maneira mais acentuada para C_β que para C_α, resultando em uma maior densidade de predição da informação para estes átomos* (figura 2.6), o que se reflete também em resultados com melhor acurácia (figura 2.4).

O efeito hidrofóbico, que promove a compactação da estrutura proteica e fundamenta o conceito de enterramentos atômicos, é determinado por interações de átomos pertencentes a cadeias laterais. Por esse motivo, é razoável supor que estas tenham um papel importante no código do enovelamento subjacente. Os resultados apresentados para o enterramento de C_β corroboram essa hipótese.

Adicionalmente, a incorporação de um descritor simples da orientação de C_β em relação a C_α nas variáveis ocultas (alfabeto tipo 2), um procedimento que, como sub-produto, fornece também uma

*A densidade de informação da predição é calculada como $i(Y;Y(Q)) = h(Y) - h(Y|Y(Q))$, sendo que $h(Y)$ é estimado a partir de correlações observadas no banco de dados e $h(Y|Y(Q))$ é calculado a partir das predições.

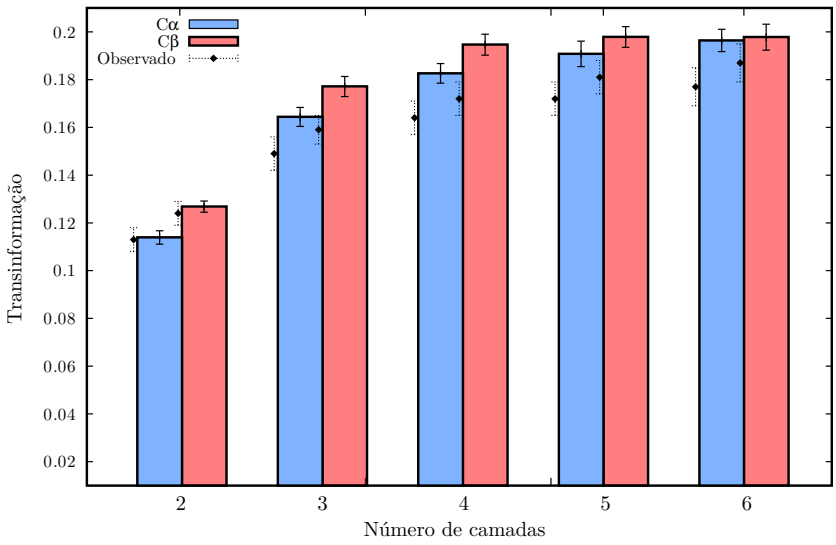


Figura 2.7: Predição da informação, conforme resultados obtidos usando o alfabeto do tipo 1 e fragmentos de tamanho $f = 7$ para os estados do HMM, em duas ou mais camadas. O ponto com as barras de erro na lateral de cada barra indica os valores estimados para a transinformação entre sequências e enterramentos correspondente.

predição dessas orientações, é capaz de melhorar a predição do enterramento de C_α quase tanto quanto a incorporação de estruturas secundárias (tipo 3) – utilizando, incidentalmente, apenas uma fração dos recursos computacionais.

Essas observações permitem concluir que não apenas o formato da cadeia principal, mas também, ao menos, a orientação relativa das cadeias laterais está codificada na sequência de aminoácidos, e que esta informação precisa ser levada em consideração por qualquer método proposto de predição de estruturas que pretenda utilizar enterramentos como intermediários informacionais.

No trabalho publicado como parte desta tese[43], é discutida a ideia de que a quantidade de informação dependente de sequência necessária para complementar as informações independentes de sequência em um processo de modelagem de estruturas pode ser menor que aquela dada pelas estimativas de densidade de transinformação entre sequências e enterramentos reais usadas na figura 2.6. Uma vez que o algoritmo desenvolvido e implementado neste trabalho é capaz de alcançar resultados muito próximos a esses valores, segue-se que a qualidade de seus resultados pode ser suficiente para, a princípio, fornecer a quantidade de informação dependente de sequência necessária para conduzir uma estrutura ao seu estado nativo em um procedimento de simulação computacional. Essa hipótese será testada no capítulo 4, utilizando também, em parte, os resultados que serão descritos nas próximas seções.

2.3.2 Número de camadas

As predições discutidas até este ponto foram realizadas com as proteínas divididas em apenas duas camadas. A figura 2.7 mostra a evolução dos resultados da informação da predição quando números maiores de camadas são utilizados. A informação da predição pode ser empregada nesta análise porque, ao contrário de medidas de acurácia, ela fornece valores comparáveis entre diferentes alfabetos de variáveis ocultas.

O fragmento utilizado para os estados do HMM nessa figura é o de tamanho 7, que é o maior valor para o qual foi computacionalmente viável fazer a predição em até 6 camadas, utilizando o alfabeto de enterramentos mais simples (tipo 1) e o alfabeto completo de 20 letras para a estrutura primária. Pode ser visto no gráfico que a qualidade da predição aumenta até quando são usadas aproximadamente 4 camadas. Depois disso, o aumento da qualidade com o número de camadas parece atingir um teto. A predição de C_β, conforme já havia sido observado para o caso de duas camadas, é consistentemente melhor que a de C_α.

É esperado que a divisão de uma proteína em um número grande de camadas discretas de enterramento deva fornecer resultados mais úteis para a determinação de estruturas que uma divisão em apenas duas ou três camadas. Esses resultados mostram que, além disso, o aumento no número de camadas resulta em uma melhora da própria qualidade da predição. Essa observação corrobora a hipótese inicial deste trabalho de que a informação codificada pela estrutura primária é um medida do nível do enterramento atômico, e não, por exemplo, uma simples bipartição entre resíduos expostos e enterrados.

2.4 Considerações finais e próximas etapas

Este capítulo descreveu o desenvolvimento da primeira versão do HmmPred e discutiu a análise informacional dos resultados de predição obtidos com o método. A principal conclusão dessa análise é a de que o algoritmo é capaz de recuperar praticamente toda a informação sobre enterramentos que está disponível em sequências de aminoácidos, de acordo com o modelo de dados adotado. Adicionalmente, foram comparados os resultados decorrentes de diversas dimensões de parâmetros de entrada do algoritmo (tamanho do fragmento, número de camadas e alfabetos de símbolos empregados), de modo a encontrar as combinações de valores capazes de gerar os melhores resultados de predição.

Todos esses resultados serão úteis na continuação desse estudo, que será discutida no capítulo seguinte e que tem o objetivo duplo de aprimorar alguns aspectos do algoritmo em relação a esta primeira versão e de utilizar o conhecimento adquirido até este ponto para construir um banco de

dados padronizado de predições de enterramento. Este banco padronizado será utilizado para a seleção das proteínas que serão empregadas na segunda parte desta tese, que terá como tema a realização de simulações do enovelamento proteico para a determinação *ab initio* de estruturas nativas.

Capítulo 3

Construção do banco de predições e seleção de proteínas

Uma vez construída a primeira versão do algoritmo de predição de enterramentos e estabelecida a sua qualidade por meio de análises informacionais, as próximas etapas da metodologia empregada na primeira parte desta tese têm como objetivos principais o aprimoramento de alguns aspectos do modelo de dados adotado pelo HMM e a construção de um banco de dados de predições de enterramento.

Este capítulo descreve, primeiramente, os novos ajustes realizados em nossa metodologia de predição após a publicação dos resultados descritos no capítulo anterior e como estes foram capazes de melhorar a qualidade ou aplicabilidade de nossos resultados de predição.

Em seguida, serão apresentadas e justificadas as escolhas de parâmetros que levaram à construção de um banco de dados padronizado de predições de enterramentos atômicos. A partir das predições pertencentes a esse banco, três proteínas, representativas de diferentes classes estruturais, foram selecionadas para a aplicação das simulações do enovelamento proteico que serão realizadas na segunda parte da tese.

3.1 Aprimoramentos ao método de predição

Três aperfeiçoamentos importantes foram realizadas em nossa metodologia de predição após a publicação dos resultados descritos no capítulo anterior. Os dois primeiros se referem a atualizações no nosso modelo de dados, com o objetivo de capturar mais precisamente durante a etapa de treinamento alguns aspectos da relação entre enterramentos e sequências que não foram modelados na versão anterior. O terceiro é o desenvolvimento de uma metodologia para expandir as predições realizadas pelo HmmPred, que, a princípio, se referem a apenas um ou dois átomos por resíduo, para todos os átomos

da proteína.

Todas as predições discutidas neste capítulo partiram de bancos de treinamento extraídos a partir de uma versão mais atualizada que a anterior (de março de 2012) da lista de proteínas compilada pelo PDBSelect[45], com as mesmas restrições anteriormente aplicadas para a filtragem de estruturas não-globulares ou de baixa qualidade.

Outra modificação importante na metodologia de predição aplicada a partir deste ponto é que, ao invés de dividir o conjunto de proteínas em dois subconjuntos, um para treinamento, e outro para predição, as predições passaram a ser realizadas para todas as proteínas do banco de dados, utilizando, em cada caso, todas as outras proteínas do banco como conjunto de treinamento, com exceção daquelas que tenham mais de 25% de identidade de sequência com a proteína sendo predita, conforme calculado pelo programa Clustal[53]. Esse passo adicional é importante porque, embora o PDBSelect seja construído de modo a minimizar a ocorrência de pares de proteínas com este grau de identidade, sua metodologia não exclui completamente a possibilidade de sua ocorrência ocasional[54].

3.1.1 Estados iniciais dependentes de posição

O modelo de dados usado para a construção do HmmPred parte do princípio de que enterramentos atômicos seguem uma distribuição estatística na qual a probabilidade de cada símbolo de enterramento ser encontrado em uma dada posição depende dos enterramentos observados em posições adjacentes. A primeira versão do HmmPred, descrita no capítulo anterior, assume implicitamente que não há dependência entre essas probabilidades e a posição absoluta do aminoácido correspondente na sequência. Consequentemente, os enterramentos de resíduos localizados nas extremidades da proteína são modelados como seguindo a mesma distribuição de probabilidades daqueles localizados em qualquer outra posição. Em termos das variáveis *forward-backward**, as probabilidades $\alpha_1(i)$ e $\beta_T(i)$ (em que T é o índice do último elemento) foram calculadas naquela versão a partir das distribuições de frequências dos estados em qualquer posição das cadeias do conjunto de treinamento.

Uma modificação importante nesse modelo foi efetuada a partir da observação de que, em proteínas reais, as extremidades da cadeia polipeptídica são encontradas mais comumente expostas ao solvente que enterradas na estrutura. Para que as correlações estatísticas relacionadas a esse fato pudessem ser aprendidas pelo nosso algoritmo de treinamento, foi adicionado ao HmmPred, através de um parâmetro configurável pelo usuário, a possibilidade de calcular $\alpha_1(i)$ e $\beta_T(i)$ a partir das distribuições de frequências encontradas no conjunto de treinamento apenas para as posições localizadas nas extremidades da cadeia.

*Apêndice 2.2, seção 2.2.1 (página 17).

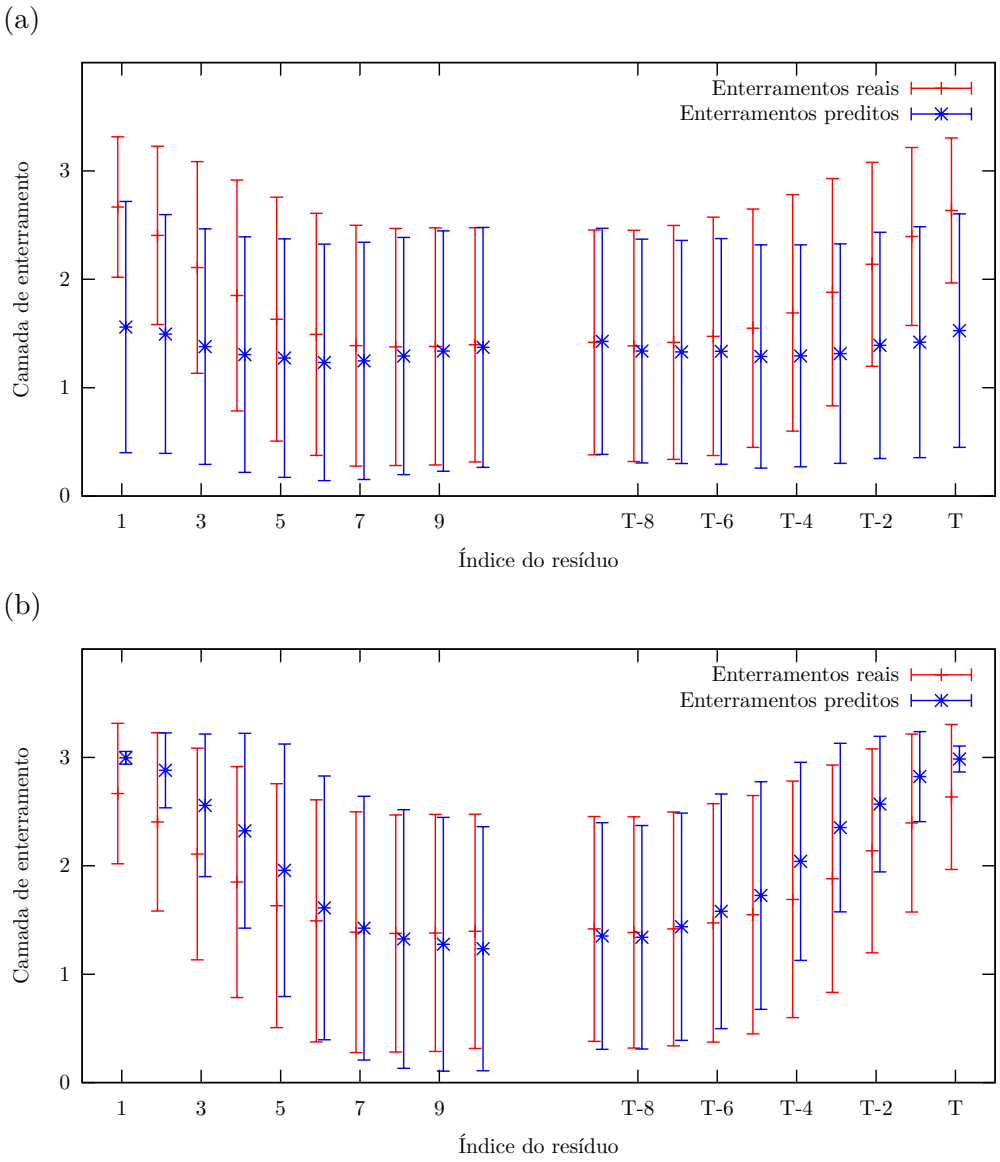


Figura 3.1: Comparação entre os enterramentos nativos de C_α e as predições realizadas com (a) a primeira versão do HmmPred, que não considera as posições absolutas dos resíduos durante a etapa de treinamento e (b) a segunda versão, em que os estados iniciais do HMM são dependentes da posição dos resíduos em relação às extremidades da cadeia. Em ambos os gráficos, a metade esquerda do eixo horizontal representa os 10 primeiros resíduos, e a metade direita, os 10 últimos (sendo T o número de resíduos de cada proteína). O eixo vertical representa a camada média de enterramento de cada posição em quatro camadas (0,1,2,3), com os desvios-padrão indicados por barras verticais. Todos os resultados foram obtidos com fragmentos de tamanho $f = 6$ e o alfabeto do tipo 2.

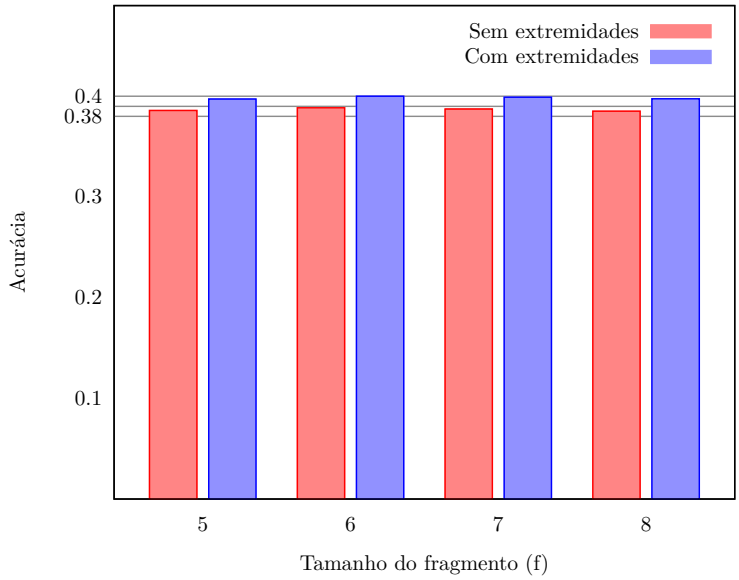


Figura 3.2: Fração de resíduos corretamente classificados para a predição de C_α com os dois modelos de dados discutidos, para tamanhos de fragmento f entre 5 e 7. Em todos os casos, é usando o alfabeto de tipo 2, em quatro camadas.

A figura 3.1 compara resultados obtidos com ambas as versões. A razão pelo qual a atualização do modelo é de fato importante fica evidente no contraste entre os padrões de enterramento observados e preditos com o modelo antigo para os resíduos localizados nas extremidades (figura 3.1-a): enquanto proteínas reais têm uma clara tendência de localizar esses resíduos nas camadas mais expostas ao solvente, as camadas preditas têm uma média entre as posições centrais da proteína e um desvio-padrão que abrange a maior parte do intervalo de valores possíveis, não fazendo, portanto, qualquer distinção significativa entre o enterramento dessas e de outras posições da cadeia. A segunda versão, por contraste, captura um padrão mais consistente de enterramentos, com o nível das camadas ligeiramente superestimado em relação ao nativo, mas bem mais próximo deste que na versão anterior (figura 3.1-b).

Conforme indicado na figura 3.2, a fração de resíduos corretamente classificados aumenta, em média, em 1% com a aplicação do novo modelo. Esse aumento ocorre de maneira consistente para tamanhos de fragmento f entre 5 e 7.

3.1.2 Treinamento dependente do comprimento de sequência

Predições realizadas usando todo o banco de dados como conjunto de treinamento também partem de um pressuposto implícito, o de que as associações estatísticas entre identidades de aminoácidos e as camadas preditas para seus enterramentos atômicos independem do tamanho de cada proteína. Considerando, entretanto, a natureza de nossa definição de camadas de enterramento, cujos limites são dependentes do número de resíduos da cadeia, é razoável supor que os modelos aprendidos pelo algoritmo podem se tornar mais precisos se cada processo de treinamento se restringir à análise de

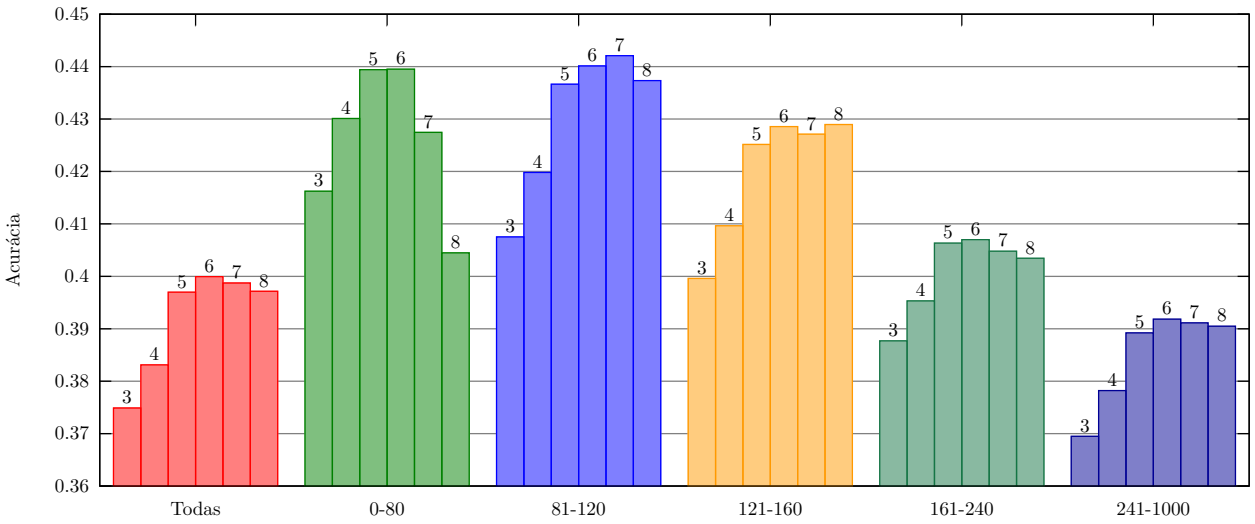


Figura 3.3: Fração média de resíduos corretamente classificados quando o banco completo é usado como conjunto de treinamento para cada proteína (primeiro grupo de barras, à esquerda) e quando são utilizados os sub-bancos classificados por comprimento da sequência de aminoácidos (demais grupos). O número acima de cada barra indica o tamanho do fragmento f usado na definição dos estados do HMM. Todas as predições foram realizadas em 4 camadas, com o alfabeto do tipo 2.

sequências com tamanhos próximos aos da proteína sendo predita.

Para estudar essa hipótese, foram produzidos cinco subconjuntos do nosso banco de dados, cada um contendo apenas proteínas cujo número total de resíduos está limitado a um dos seguintes intervalos: 1-80, 81-120, 121-160, 161-240 e 241-1000. A figura 3.3 mostra os resultados dessas predições, realizadas com o tamanho do fragmento f variando entre 3 e 8. O primeiro padrão que se destaca nessa figura é uma perceptível melhoria da fração de resíduos corretamente classificados pelo algoritmo para os sub-bancos restritos a tamanhos menores, principalmente para os bancos de tamanhos 1-80 e 81-120.

Também é possível perceber que a relação entre acurácia e tamanho do fragmento é dependente do tamanho do banco: como consequência da saturação de informações estatísticas disponíveis para combinações mais longas de símbolos em sequências de menor comprimento, o pico de qualidade tende a ser atingido com valores de f menores nos primeiros bancos que nos demais. Considerando todas as combinações testadas, os melhores índices médios de acurácia, em torno de 44%, foram obtidos para f entre 5 e 6 no banco 0-80 e para f entre 6 e 7 no banco 81-120.

3.1.3 Predição de todos os átomos

Conforme visto no capítulo anterior, a saída do programa HmmPred, é, para cada resíduo t da sequência de entrada, uma distribuição de probabilidades $p_t(y)$, que define a probabilidade de cada símbolo $y \in \mathcal{Y}$ estar associado àquela posição, onde $\mathcal{Y}$ é o alfabeto de variáveis ocultas. O símbolo predito para cada posição é aquele para o qual $p_t(y)$ tem o valor máximo.

O HmmPred, portanto, prevê apenas um unidade de informação por resíduo. Ainda que essa informação possa estar relacionada ao enterramento de mais de um átomo, como no caso do alfabeto do tipo 2, que considera os enterramentos relativos de C_β em relação a C_α, ela é, por si só, insuficiente para a determinação dos enterramentos dos outros átomos.

A predição do enterramento de todos os átomos pesados de uma proteína para N camadas pode ser efetuada através de uma extrapolação simples realizada a partir das distribuições de probabilidade dadas pelo HmmPred:

$$p_i^a(y') = \sum_y p_{t(i)}(y) p_a(y'|y) \tag{3.1}$$

onde $p_i^a(y')$ é a probabilidade do átomo i, de tipo atômico a, ser encontrado na camada $y' \in \mathcal{Y}' = \{y_1, \cdots, y_N\}$, de modo que $\sum_{\mathcal{Y}'} p_i^a(y') = 1$; $p_{t(i)}(y)$ é a probabilidade emitida pelo HmmPred de encontrar o símbolo $y \in \mathcal{Y}$ no resíduo $t(i)$, que inclui o átomo i; $p(y'|y)$ é a probabilidade de que o átomo a esteja na camada y', dado que seu resíduo correspondente está associado ao símbolo y. Este último conjunto de probabilidades é calculado diretamente a partir de contagens de frequências no mesmo banco de dados usado para o treinamento do HMM.

É importante ressaltar que os alfabetos $\mathcal{Y}'$ e $\mathcal{Y}$ são completamente independentes; enquanto o primeiro é necessariamente um alfabeto simples em que cada símbolo corresponde a uma camada, o segundo pode ser de qualquer dos tipos descritos na seção 2.2.3 (página 21), associado a qualquer número de camadas.

O script que implementa a equação 3.1 neste trabalho foi escrito por Diogo César Ferreira, como parte de seu projeto de Iniciação Científica no Laboratório de Biofísica Teórica e Computacional.

3.2 Preparação do banco de predições

Como parte dos preparativos para a segunda parte deste trabalho, é necessário construir um banco de dados de predições de enterramento, a partir do qual serão selecionadas estruturas para simulação do enovelamento.

Os parâmetros do algoritmo usados na construção desse banco precisam ser escolhidos de modo a simultaneamente maximizar a qualidade possível das predições e manter a viabilidade computacional de sua obtenção. Em termos de recursos computacionais, o principal fator limitador das execuções do HmmPred são os seus requisitos de memória, decorrentes da necessidade de armazenamento dos estados do HMM durante todo o processo de predição. Os dados referentes a esses estados são armazenados sob a forma de matrizes de tamanho L^f, onde L é o número de símbolos do alfabeto de variáveis ocultas, e f, o comprimento do fragmento utilizado na definição dos estados. Os resultados

discutidos no capítulo anterior podem ser interpretados de modo a encontrar a combinação desses parâmetros que permite maximizar a qualidade das predições, sem incorrer em custos de armazenamento em memória desnecessários ou impraticáveis.

Em relação ao alfabeto de variáveis ocultas, as figuras 2.4-2.6 (páginas 23, 27 e 28) mostram que os alfabetos de tipo 2 (que consideram a orientação relativa da cadeia lateral) e tipo 3 (que consideram estruturas secundárias) são consistentemente superiores aos alfabetos simples de tipo 1 (que consideram apenas o enterramento de um átomo do resíduo). Embora o tipo 3 seja superior ao tipo 2, a diferença de qualidade é pequena para justificar o custo computacional envolvido em seu cálculo, que se torna proibitivo em algumas configurações quando aplicado a um número de camadas superior a duas ou três. Por esse motivo, as predições usadas na construção do banco de dados foram todas realizadas com o alfabeto de tipo 2. Os resultados foram extrapolados para todos os átomos utilizando a metodologia descrita na seção anterior.

Foram utilizadas quatro camadas na definição dos enterramentos, pois este é o número a partir do qual as predições descritas na figura 2.7 (página 29) atingem o limite na melhoria de sua qualidade. Para as sequências de aminoácidos, foi empregado o alfabeto de 20 letras. Todas as predições foram realizadas com a nova versão do HmmPred, que considera as porções inicial e final das cadeias polipeptídicas de maneira distinta, com fragmentos de tamanho $f = 6$ para os estados do HMM e o alfabeto de variáveis ocultas do tipo 2.

Por razões planejadas de eficiência computacional das simulações e como resultado das análises discutidas anteriormente, foi decidido que o banco de dados que será utilizado para a seleção de proteínas para as simulações do enovelamento será o subconjunto de estruturas com tamanho igual ou inferior a 80 resíduos, o qual totaliza 278 proteínas.

Essa combinação de parâmetros se aproxima do limite computacional disponível em nosso laboratório e está suficientemente próxima à combinação ideal, de modo que possíveis melhorias obtidas com o aumento de seus valores provavelmente não alterariam de maneira substancial os resultados finais obtidos.

A figura 3.4 mostra a distribuição da qualidade das predições aplicadas para cada proteína do conjunto de predições resultante, de acordo com a medida da fração de resíduos atribuídos à camada correta (acurácia). Este valor varia entre aproximadamente 25%, que equivale a uma distribuição aleatória entre quatro camadas, até mais de 65%, com uma média de 44% e desvio-padrão de 7%.

Conforme também pode ser visto na figura 3.4, não há correlação entre a qualidade da predição de uma proteína e a distribuição de identidades de sequência entre ela e as demais proteínas do banco de dados. A figura 3.5 mostra que, além disso, também não há correlação entre a qualidade da predição

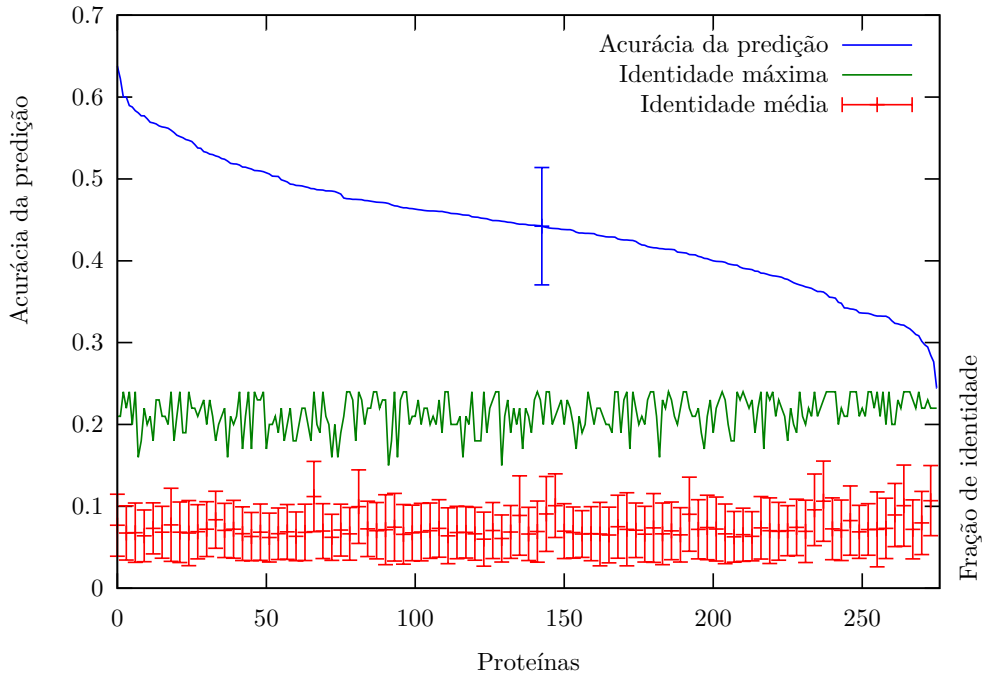


Figura 3.4: Distribuição da qualidade das predições realizadas em um conjunto não-redundante de 278 proteínas globulares com 80 resíduos ou menos. Cada ponto ao longo do eixo horizontal representa uma proteína. A curva na parte de cima do gráfico indica no eixo vertical a fração de átomos classificados nas camadas corretas pela predição, com a média e o desvio-padrão indicados no centro da curva. Na parte de baixo do gráfico, o eixo vertical representa a fração de identidade de sequência de todas as outras proteínas do banco de dados em relação à proteína correspondente. Estão indicados os valores médios, desvios-padrão e valores máximos encontrados.

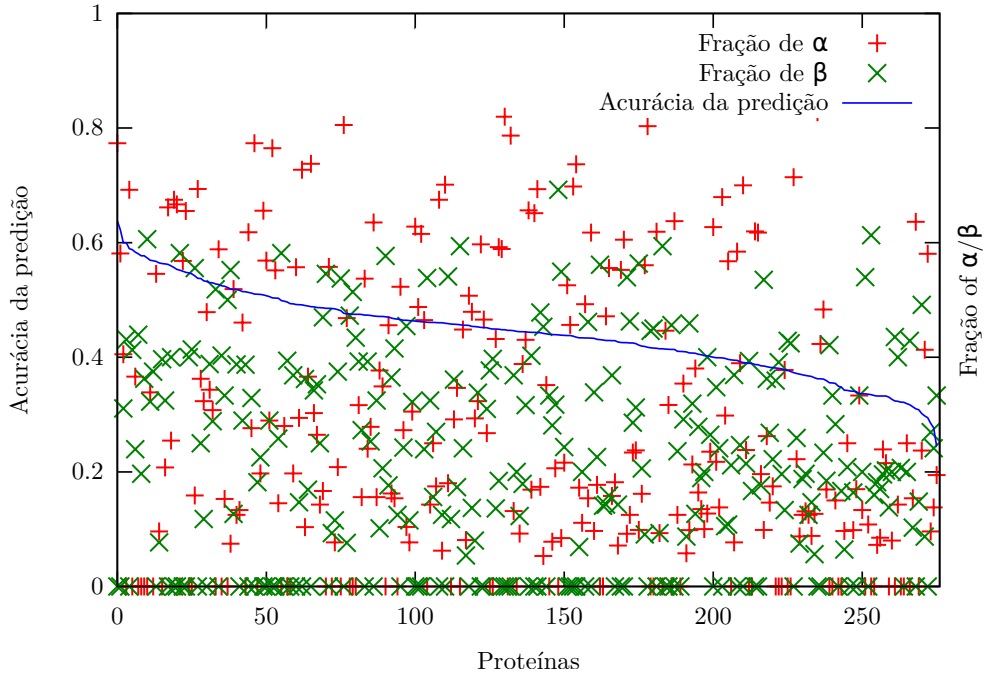


Figura 3.5: Fração de resíduos que adotam conformação em α-hélice ou folha β para cada uma das proteínas do mesmo banco de dados utilizado na figura 3.4.

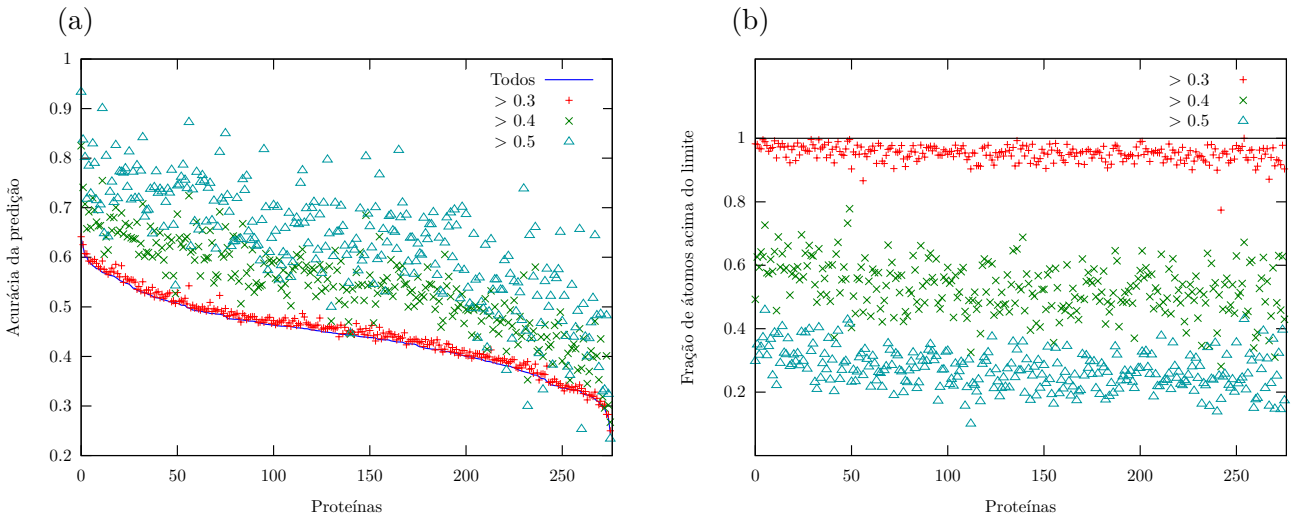


Figura 3.6: Variações nos resultados da predição considerando apenas subconjuntos de átomos com probabilidades maiores para as camadas preditas. (a) Valor da acurácia para cada conjunto de átomos cujo valor $p(y_n)$ está acima de um certo patamar mínimo; (b) Fração de átomos que atendem a cada um desses requisitos.

e os tipos de estruturas secundárias formadas pelas proteínas do banco.

Pode-se observar, entretanto, que a acurácia das predições melhora quando se consideram apenas as predições de enterramento que resultaram em probabilidades $p(y_n)$ para a camada predita maiores que certos valores-limite. Conforme disposto na figura 3.6-a, a acurácia da predição melhora um pouco para $p(y_n) > 0{,}3$, e mais expressivamente para $p(y_n) > 0.4$ e $p(y_n) > 0.5$. Uma desvantagem prática de se fazer essa filtragem, entretanto, é que cada um destes patamares diminuiu mais substancialmente o número de átomos disponível para predição (figura 3.6-b). Além disso, é importante ressaltar que essas variações médias na acurácia dos resultados não são suficientemente regulares para permitir a estimativa da qualidade de acurácia de proteínas individuais apenas com base nas probabilidades emitidas pelo algoritmo. Por esses motivos, nas predições descritas neste trabalho, todos os átomos são utilizados, independentemente das diferenças de probabilidades associadas.

3.2.1 Seleção das proteínas

Três proteínas de diferentes classes estruturais, destacadas na figura 3.7, foram selecionadas no banco de dados para um estudo das simulações do enovelamento: (1) a proteína efetora RxLR[55] de *Phytophthora capsici*, formada por quatro α-hélices (RePc); (2) a proteína G de estreptococos[56], formada por uma α-hélice e quatro folhas β (ProtGSsp); (3) a proteína de choque frio Bc-Csp[57], de *Bacillus caldolyticus* (CsBc), constituída por um barril β de topologia não trivial (tabela 3.1). A figura 3.8 mostra as estruturas nativas das três proteínas selecionadas, coloridas de acordo com as diferenças entre os enterramentos reais e preditos, em camadas. Nessa figura, fica evidente que a maior parte

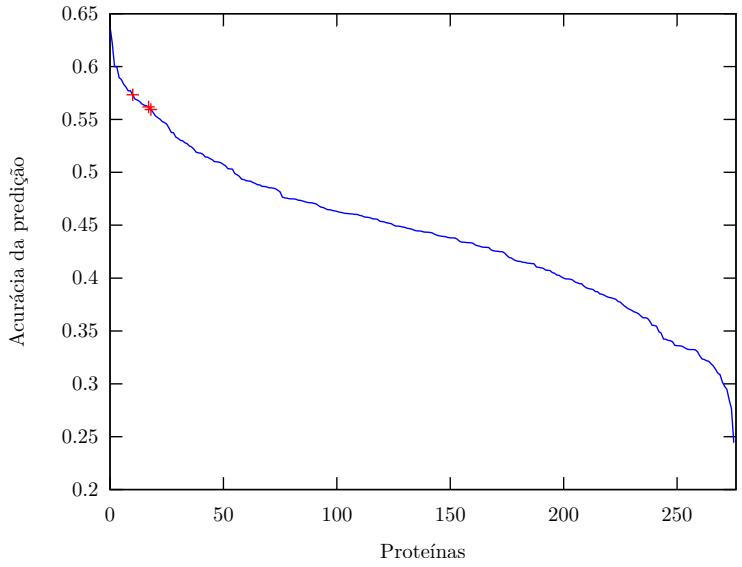


Figura 3.7: O banco de dados do qual foram extraídas as predições de enterramento que serão usadas nas simulações. Cada ponto ao longo do eixo horizontal representa uma proteína. O eixo vertical representa a fração de átomos atribuídos à camada correta, entre quatro camadas possíveis. As três proteínas selecionadas estão assinaladas em destaque.

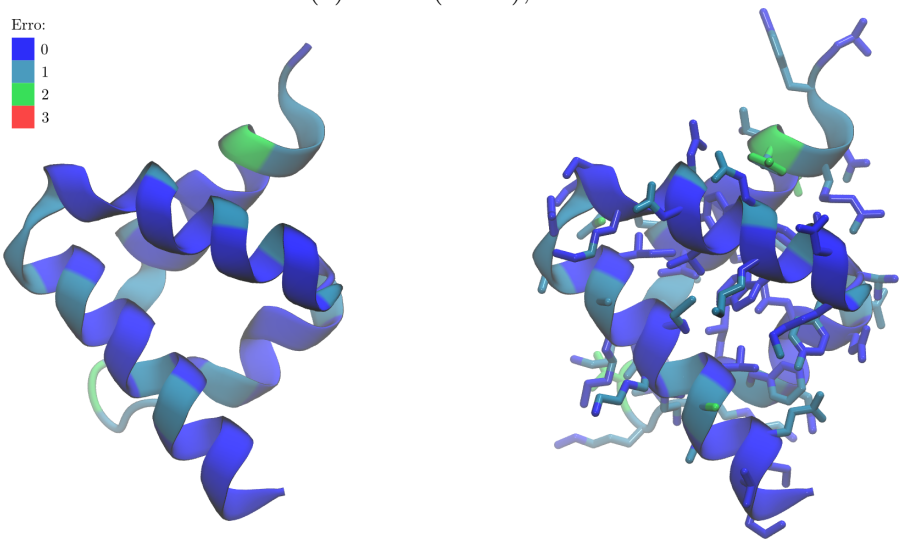
Proteína	Descrição	Topologia	Código PDB	Número de aminoácidos
RePc	Proteína efetora RxLR de *Phytophthora capsici.*	α	3zr8(X)	65
ProtGSsp	Proteína G de estreptococos.	β	3fil(A)	56
CsBc	Proteína de choque frio Bc-Csp de *Bacillus caldolyticus.*	α/β	1c9o(A)	66

Tabela 3.1: As três proteínas selecionadas para este estudo.

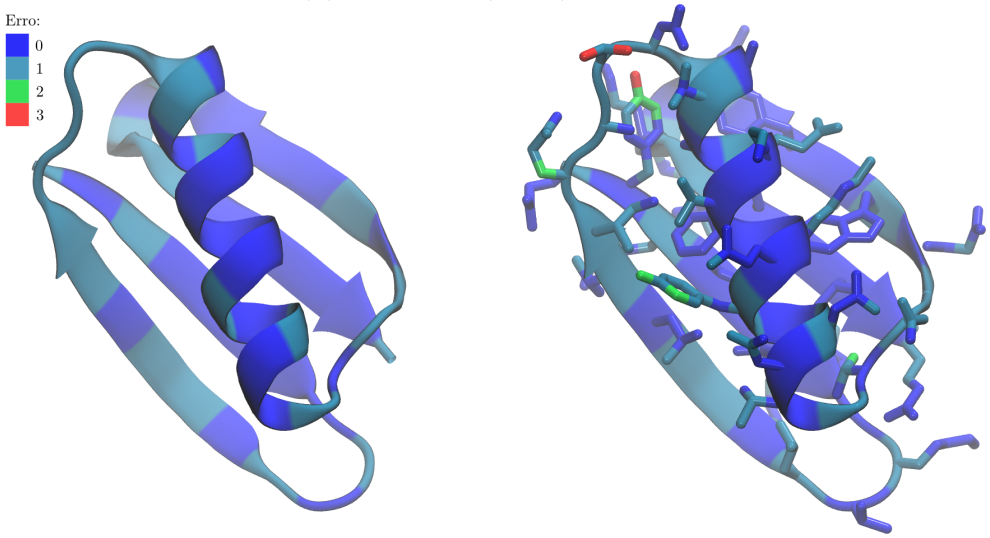
dos átomos preditos erroneamente para estas estruturas são atribuídos à camada vizinha à correta; uma quantidade substancialmente menor é predita para uma posição separada por duas camadas, e poucos átomos são atribuídos à camada completamente oposta à real (apenas 5 para CsBc, 3 para ProtGSsp, nenhum para RePc).

Conforme discutido anteriormente, essas predições foram obtidas utilizando conjuntos de treinamento dos quais se eliminou, antes de cada predição, todas as sequências com 25% ou mais de identidade com a proteína sendo predita, de acordo com o cálculo realizado pelo programa Clustal[53]. Em relação às 3 proteínas selecionadas, o banco de treinamento já não continha nenhuma outra sequência com mais de 20% de identidade, e a tabela 3.2 mostra que os resultados não seriam substancialmente alterados mesmo que o limite fosse reduzido até 10%. Apenas a eventual redução do limite de identidade ao patamar de 5% seria capaz de reduzir a acurácia de duas das três proteínas para valores ligeiramente menores que 50%. Ainda assim, como mostra a tabela 3.3, essa diminuição na qualidade ainda poderia ser atribuída simplesmente ao fato de que este valor reduziria o banco de treinamento

(a) RePc (65 aa), 56%



(b) ProtGSsp (56 aa), 55.71%



(c) CsBc (66 aa), 55.98%

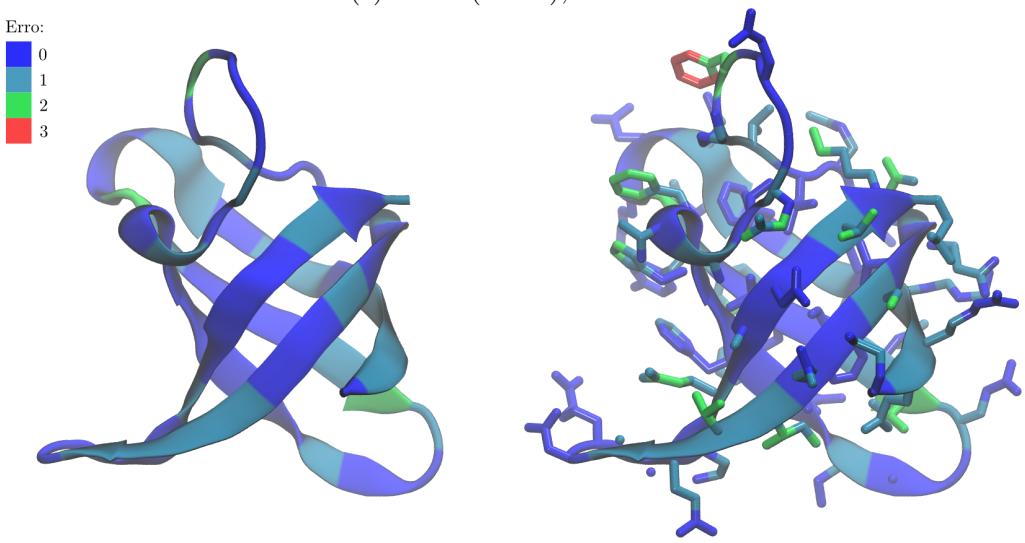


Figura 3.8: Estruturas das três proteínas selecionadas, coloridas de acordo com a diferença entre a camada de enterramento predita e a camada real para cada átomo. As cadeia principais estão representadas de acordo com os valores para C_α. Nas figuras da direita, também estão indicados os átomos das cadeias laterais. Acima de cada par de figuras, estão assinalados o número de aminoácidos e a porcentagem de átomos corretamente preditos para a proteína correspondente.

a uma pequena fração do seu tamanho original.

Proteína	25%	20%	15%	10%	5%
CsBc	55.98%	55.98%	56.37%	54.25%	49.81%
ProtGSsp	55.71%	55.71%	56.18%	57.11%	46.15%
RePc	56.00%	56.00%	56.38%	59.05%	59.81%

Tabela 3.2: Porcentagem de resíduos corretamente classificados por predições em que diferentes valores-limite de identidade de sequência são usados para eliminar proteínas do conjunto de treinamento.

Proteína	25%	20%	15%	10%	5%
CsBc	0	0	7	54	211
ProtGSsp	0	0	4	63	233
RePc	0	0	5	44	198

Tabela 3.3: Quantidade de sequências eliminada do banco de dados após a aplicação de cada valor-limite para a identidade de sequência máxima no conjunto de treinamento. O banco original contém 278 proteínas.

Essas análises demonstram, portanto, que as três proteínas escolhidas têm valores de acurácia que as situam no grupo das proteínas com melhores resultados de predição em nosso banco de dados e que esses resultados não são decorrentes de vieses na seleção de estruturas para treinamento do modelo ou de configurações particulares de estruturas secundárias. Com essas conclusões, a seleção das três proteínas encerra os objetivos da primeira parte desta tese.

A segunda parte do trabalho terá como tema principal o estudo de simulações do enovelamento que utilizarão as predições de enterramentos atômicos obtidas até aqui para tentar recuperar as estruturas nativas dessas três proteínas a partir de suas respectivas sequências de aminoácidos. Os resultados das simulações serão discutidos em detalhes no capítulo 5. Antes disso, porém, é necessário apresentar a metodologia que será utilizada nas simulações, e esse será o assunto do próximo capítulo.

Capítulo 4

Simulação de estruturas: Metodologia

Este capítulo introduz a segunda parte desta tese de doutorado, que tem como objetivo o desenvolvimento de uma metodologia computacional capaz de aplicar as predições de enterramentos apresentadas anteriormente para simular o enovelamento de proteínas selecionadas a partir de suas sequências de aminoácidos.

A metodologia de simulação aplicada neste trabalho é um algoritmo de dinâmica molecular em temperatura constante que atua sobre uma função de energia artificialmente construída, a qual inclui termos independentes de sequência já consagrados em algoritmos do gênero aliados a dois termos especialmente desenvolvidos para o problema tratado, um dos quais modela o enterramento atômico dependente de predições realizadas a partir da sequência de aminoácidos.

A implementação computacional que concretiza as ideias desenvolvidas a seguir é baseada em um programa escrito originalmente por Paul Whitford[58] em linguagem Fortran. O código foi originalmente adaptado com a adição de novos potenciais por Pereira de Araújo para o trabalho que primeiramente descreveu os enterramentos atômicos[22] e, desde 2010, tem sido mantido e aprimorado pelo autor desta tese.

4.1 Função de energia potencial

Um simulação de dinâmica molecular tem como objetivo reproduzir, ao longo de um intervalo de tempo, o comportamento de um sistema molecular modelado como um conjunto de massas pontuais no espaço cartesiano. As forças que atuam sobre cada átomo em cada instante de tempo são obtidas através da derivação de uma função de energia potencial. Ao longo da simulação, espera-se que o espaço conformacional seja explorado de modo a atingir uma conformação final de baixa energia.

No caso do algoritmo apresentado neste trabalho, denominado MDBury, a função de energia potencial implementada não tem o objetivo de ser fisicamente realista. Ao invés disso, ela é construída

de modo a ter o formato mais simples possível que incorpore as predições de enterramentos atômicos de modo a situar o mínimo energético suficientemente próximo à estrutura nativa.

Essa simplificação tem o benefício duplo de acelerar os cálculos computacionais e de permitir uma separação clara, no potencial, entre a porção dependente de sequências (derivada das predições de enterramentos atômicos) e a porção independente da sequência. Esta última modela apenas as forças mínimas necessárias para que uma topologia proteica genérica possa ser mantida pelo sistema e não é, por si só, capaz de distinguir entre estruturas nativas e não-nativas. Esse aspecto do algoritmo é importante para permitir a validação da hipótese central deste trabalho, que é o papel preponderante dos enterramentos atômicos no código do enovelamento.

O objetivo das simulações que empregarão o potencial descrito a seguir é, partindo de estruturas completamente desenoveladas, chegar a conformações que tenham o enovelamento correto e a mesma topologia geral da estrutura nativa. O fato de a função não modelar explicitamente todas as possíveis forças que podem contribuir para a estabilização da estrutura terciária impede que ela possa ser usada para promover ajustes finos em estruturas além de um certo grau de semelhança com a conformação nativa. Uma etapa posterior de refinamento poderá implementada futuramente, como ocorre em outros algoritmos de predição de estruturas, com o uso de um segundo potencial dedicado especificamente a este fim.

A função de energia potencial utilizada pelo MDBury tem o seguinte formato:

$$\begin{aligned} V &= V_{\text{ligações}} + V_{\text{ângulos}} + V_{\text{diedrais}} \\ &\quad V_{\text{repulsão}} + \underbrace{V_{\text{enterramentos}} + V_{\text{LH}}}_{\bigstar} \end{aligned} \tag{4.1}$$

Isto é, a energia potencial total do sistema é calculada como a soma de vários termos independentes. Os dois últimos termos, destacados acima, foram desenvolvidos especificamente para o nosso algoritmo. Os demais têm formato similar a outros comumente empregados em aplicações de mecânica molecular.

A definição do termo $V_{\text{enterramentos}}$ depende de uma classificação de tipos atômicos que pode ser derivada a partir de predições de enterramento que partem da sequência de aminoácidos, de acordo com a metodologia descrita nos capítulos anteriores. Todos os demais termos são independentes de sequência e têm o mesmo formato geral para qualquer estrutura proteica.

4.1.1 Termos ligantes

Os três primeiros termos do potencial têm a função de assegurar a manutenção de uma topologia realista para as ligações covalentes da proteína. Esses termos são projetados como potenciais de Hooke,

de maneira análoga a massas pontuais unidas por molas, que é um formato comumente encontrado em termos dessa natureza em potenciais de mecânica molecular[59]. Os mínimos de energia nesses potenciais se localizam nas conformações em que as distâncias e ângulos formados entre os átomos ligados covalentemente se encontram em seus valores ótimos, e as constantes de mola são ajustadas de modo a não permitir que esses valores desviem além do realisticamente esperado (figura 4.1).

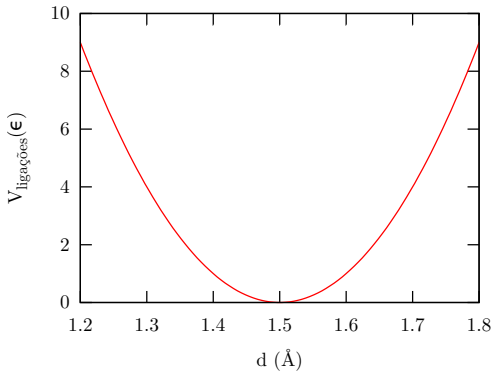


Figura 4.1: Potencial $V_{\text{ligações}}$, que mantém a distância da ligação covalente. O mínimo de energia está localizado na distância ótima, que, neste caso, é de 1,5Å. Os outros dois termos ligantes têm um formato análogo.

Os três termos ligantes são

$$V_{\text{ligações}} = \sum_{\text{ligações}} k_d(d - d_0)^2 \tag{4.2}$$

$$V_{\text{ângulos}} = \sum_{\text{ângulos}} k_\theta(\theta - \theta_0)^2 \tag{4.3}$$

$$V_{\text{diedrais}} = \sum_{\text{diedrais}} k_\chi(\chi - \chi_0)^2 \tag{4.4}$$

onde d é a distância entre um par de átomos, θ é o ângulo formato entre três átomos e χ é o ângulo diedral formado entre quatro átomos (figura 4.2). Os parâmetros referentes aos valores ótimos correspondentes, d_0, θ_0 e χ_0, são obtidos pelo programa a partir do observado em estruturas estendidas padronizadas. As constantes de mola têm os valores $k_d = 100\,\epsilon\text{Å}^{-2}$ $k_\theta = 20\,\epsilon\,\text{rad}^{-2}$ e $k_\chi = 10\,\epsilon\,\text{rad}^{-2}$, onde ϵ é a nossa unidade de energia, definidos de modo a tornar consistentemente alto o custo energético do desvio de qualquer distância ou ângulo em relação aos valores ótimos, em relação aos outros termos do potencial.

Os termos ligantes são aplicados apenas em conjuntos de átomos cuja configuração relativa se deseja manter aproximadamente constante ao longo da simulação. Desse modo, a equação 4.2 é aplicada a todos os pares de átomos unidos covalentemente, e a equação 4.3, a todos os vértices formados por três átomos unidos covalentemente. A equação 4.4, entretanto, é aplicada em apenas nos ângulos diedrais que devem permanecer fixos, como a ligação peptídica e os grupos planares das cadeias laterais dos

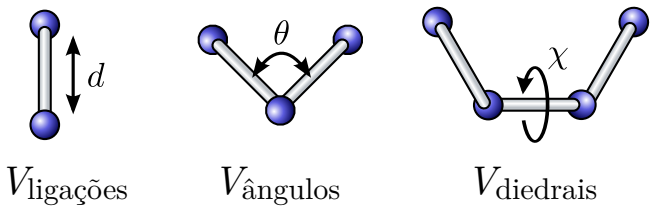


Figura 4.2: Termos ligantes do campo de força.

aminoácidos aromáticos, assim como nos diedrais formados pelos átomos N, C_α, C e C_β, de modo a manter a quiralidade correta da cadeia lateral.

4.1.2 Repulsão atômica

A repulsão eletrônica entre os orbitais de átomos próximos no espaço (repulsão de Pauli) é modelada em nosso potencial através do seguinte termo (figura 4.3):

$$V_{\text{repulsão}} = \sum_{\text{pares}} \epsilon_{\text{rep}} \left(\frac{\sigma_{\text{rep}}}{d} \right)^{12} \tag{4.5}$$

onde d é a distância entre dois átomos e $\epsilon_{\text{rep}} = 1.0\epsilon$.

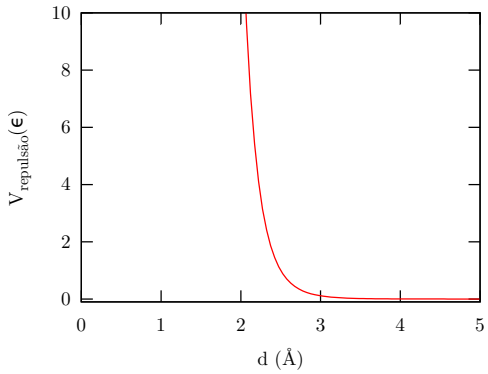


Figura 4.3: O termo da repulsão atômica, que modela a repulsão de Pauli entre átomos próximos no espaço.

Este termo corresponde à parte repulsiva do potencial conhecido como Lennard-Jones, que é tipicamente utilizado em simulações moleculares para modelar tanto a repulsão de Pauli como a atração de van der Waals entre pares de átomos não-ligados. Nossa implementação busca manter a simplicidade da porção independente de sequência, não incluindo a parte atrativa.

O termo de repulsão atômica é aplicado a todos os pares de átomos que estão separados por mais de duas ligações covalentes, desde que não pertençam ao mesmo diedral plano. A constante σ_{rep} corresponde ao raio atômico e tem o valor $\sigma_{\text{rep}} = 2.5$Å para todos os átomos, exceto no caso da repulsão entre carbonos C_β e o oxigênio da carbonila da cadeia principal, quando assume o valor $\sigma_{\text{rep}} = 3$Å. Esse último ajuste foi necessário para impedir o surgimento de algumas estruturas secundárias não

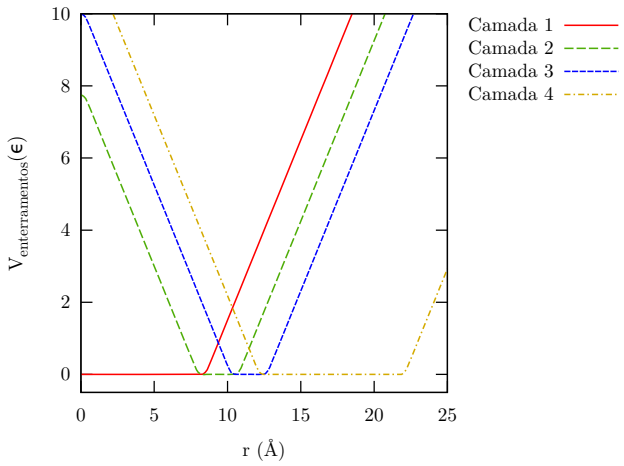


Figura 4.4: Exemplo de uma possível configuração para o potencial dos enterramentos em uma proteína dividida em quatro camadas. Cada linha representa a variação da energia em função do enterramento para átomos classificados em uma camada diferente.

realistas, formadas em simulações nas quais todos os átomos tinham o mesmo raio atômico.

4.1.3 Enterramentos

Os dados de entrada do MDBury podem incluir, para cada átomo i, um valor esperado de enterramento r_i^* e um intervalo de tolerância δ_i, medidos em Å. O termo de enterramentos é constituído por um somatório aplicado a todos os átomos para os quais esses dois parâmetros são definidos:

$$V_{\text{enterramentos}} = \sum_{\text{átomos}} B(r_i) \tag{4.6}$$

onde r_i é o enterramento do átomo i (isto é, sua distância até o centro geométrico da proteína).

A função $B(r_i)$ é construída de modo a ter o valor igual a zero quando r_i se encontrar entre o intervalo $(r_i^* - \delta_i, r_i^* + \delta_i)$ e a crescer linearmente à medida que r_i se afasta desta janela (figura 4.4). Como consequência, átomos que estão dentro do intervalo especificado não sofrem qualquer força oriunda deste termo, e átomos que estão fora sofrem uma força constante de $\pm 1\epsilon\text{Å}^{-1}$ em sua direção. Adicionalmente, uma pequena região de quadrática de tamanho $\delta_q = 0.5$Å, necessária para que a função seja diferenciável em todos os pontos, é aplicada em torno do intervalo definido.

A forma completa da função $B(r_i)$ é a seguinte:

$$B(r_i) = \begin{cases} -a_1 r^2 + b_1 & \text{para} \quad r \leq \delta_q \\ -a_2 r + b_2 & \text{para} \quad \delta_q < r \leq r_i^* - \delta_i - \delta_q \\ a_3 (r - r_3)^2 & \text{para} \quad r_i^* - \delta_i - \delta_q \leq r < r_i^* - \delta_i \\ 0 & \text{para} \quad r_i^* - \delta_i \leq r < r_i^* + \delta_i \\ a_4 (r - r_4)^2 & \text{para} \quad r_i^* + \delta_i < r \leq r_i^* + \delta_i + \delta_q \\ a_5 r - b_5 & \text{para} \quad r > \delta_q \end{cases} \tag{4.7}$$

onde $a_1 \ldots a_5$ e $b_1 \ldots b_5$ são definidos em termos de r_i^*, δ_i e δ_q.

Nas simulações que serão relatadas no capítulo seguinte, todos os átomos são classificados em um de quatro tipos atômicos, correspondentes a quatro camadas de enterramento, de acordo com predições realizadas a partir da sequência utilizando o programa HmmPred. Em termos do raio de giro esperado para a proteína, R_g, esses quatros tipos atômicos correspondem, respectivamente, aos valores $(r^*/R_g, \delta/R_g) = (0.378, 0.378)$, $(0.859, 0.103)$, $(1.051, 0.089)$, e $(1.57, 0.43)$. O raio de giro é estimado a partir do número de resíduos N da proteína, de acordo com $R_g \approx 2.7\sqrt[3]{N}$Å.

4.1.4 Ligações de hidrogênio

Na estrutura da cadeia principal de uma proteína, átomos de oxigênio pertencentes à carboxila de um resíduo são capazes de formar ligações de hidrogênio com os átomos de hidrogênio ligados ao grupamento amina de outro resíduo, um tipo de ligação que permite que a proteína se enovele nos padrões locais regulares conhecidos como estruturas secundárias. A formação de ligações de hidrogênio no MDBury é modelada por um novo potencial, especialmente desenvolvido para essas simulações, cujo formato se baseia em vetores formados pelas posições dos átomos envolvidos.

Como a topologia usada pelo programa não inclui átomos de hidrogênio, a ligação de hidrogênio é modelada como uma força atrativa que atua diretamente entre átomos de oxigênio e nitrogênio na cadeia principal. As ligações de hidrogênio que podem ser formadas entre átomos pertencentes a cadeias laterais, por enquanto, ainda não são consideradas pelo programa.

Em uma proteína globular, os resíduos com cadeia lateral hidrofóbica tendem a se encontrar mais enterrados na estrutura e a formar ligações de hidrogênio entre si; resíduos polares, por outro lado, tendem a ficar expostos ao solvente e formar ligações de hidrogênio com este. Como nosso modelo não inclui átomos de água explícitos, esse fenômeno é modelado como um potencial que aplica uma penalidade de ϵ_{hb} (medido em ϵ, nossa unidade de energia) para pares de átomos potencialmente doadores (O) e aceptores (N) da cadeia principal que se encontram enterrados na estrutura sem

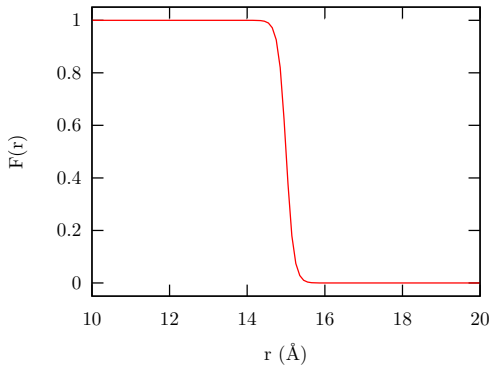


Figura 4.5: A função de $F(r)$, para $\mu_r = 15$Å e $\beta_r = 10$Å^{-1}.

formar ligações de hidrogênio. O limite de enterramento para esta penalidade, em Å, é definido por um parâmetro μ_r. Em outras palavras, há duas maneiras de um destes pares de átomos alcançar o estado de menor energia: deslocar-se até uma distância do centro maior que μ_r, ou se manterem enterrados e alinhados em uma configuração ótima, compatível com uma ligação de hidrogênio.

Se o termo dos enterramentos não for utilizado, a consequência deste potencial é simplesmente o deslocamento de todos os resíduos para uma posição de enterramento superior a μ_r, formando uma envoltório que circunda o raio atômico determinado por esse valor. Quando o termo dos enterramentos é aplicado, entretanto, a força que conduz os átomos aos enterramentos preditos passa a competir com essa tendência de exposição indiscriminada dos resíduos, e os átomos enterrados passam a satisfazer a restrição do termo das ligações de hidrogênio alinhando-se na configuração apropriada.

O termo das ligações de hidrogênio depende, portanto, tanto da posição relativa dos átomos envolvidos entre si, como da posição de cada um deles em relação ao centro da proteína (enterramento). Essas dependências são modeladas através de várias aplicações de funções contínuas projetadas de modo a retornar valores muito próximos a 1 ou 0 para a maior parte dos possíveis valores de entrada.

4.1.4.1 Funções $F(\alpha)$

A função $F(\alpha)$, tal como aplicada neste trabalho, tem o seguinte formato:

$$F(\alpha) = \frac{1}{1 + \exp(\beta_\alpha(\alpha - \mu_\alpha))} \tag{4.8}$$

onde β_α é um valor em torno do qual a saída da função transmuta de 1 para 0, de forma abrupta porém contínua (figura 4.5). Em outras palavras, para $\alpha < \beta_\alpha$, $F(\alpha) \approx 1$; para $\alpha > \beta_\alpha$, $F(\alpha) \approx 0$. A declividade da transição contínua entre 1 e 0 ao redor do valor de β_α é modelada por μ_α . O formato dessa função é baseado na função de Fermi, usada no estudo da estrutura eletrônica no estado sólido.

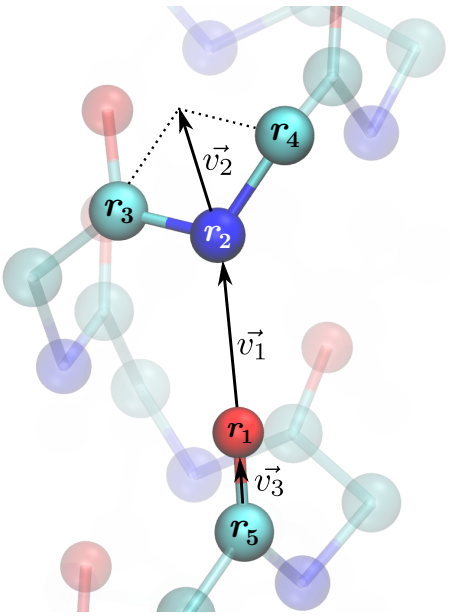


Figura 4.6: As cinco coordenadas atômicas e os três vetores que participam da definição da ligação de hidrogênio.

4.1.4.2 Definição de V_{LH}

A conformação ótima para a caracterização da ligação de hidrogênio, de acordo com o modelo adotado pelo potencial, depende das coordenadas de cinco átomos: o oxigênio aceptor da carbonila (1), o nitrogênio doador (2), os dois átomos adjacentes a este nitrogênio (3, 4) e o carbono adjacente ao oxigênio aceptor (5). Estas coordenadas definem três vetores: $\vec{v_1} = \vec{r_2} - \vec{r_1}$, $\vec{v_2} = \vec{r_3} + \vec{r_4} - 2\vec{r_2}$ e $\vec{v_3} = \vec{r_1} - \vec{r_5}$. Em termos desses vetores, $h = |\vec{v_1}|$ é o módulo de $\vec{v_1}$, η é o ângulo entre $\vec{v_1}$ e $\vec{v_2}$, e θ é o ângulo entre $\vec{v_1}$ e $\vec{v_3}$ (figura 4.6).

Para cada par de átomos doador e aceptor, i e j, a ligação de hidrogênio é modelada como o produto de três funções $F(\alpha)$ aplicadas a h, η e θ:

$$\lambda_{ij}(h,\eta,\theta) = F(h)F(\eta)F(\theta) \tag{4.9}$$

sendo os parâmetros das respectivas funções $F(\alpha)$ definidos da seguinte forma: $\mu_h = 3\text{Å}$, $\beta_h = 10\text{Å}^{-1}$, $\mu_\eta = 0.5\,\text{rad}\ (29°)$, $\beta_\eta = 100\,\text{rad}^{-1}$, $\mu_\theta = 0.7\,\text{rad}\ (40°)$, $\beta_\theta = 100\,\text{rad}^{-1}$. O valor de $\lambda_{ij}(h,\eta,\theta)$, portanto, é aproximadamente igual a 1 quando são cumpridas as três condições para o par i,j: $h < \mu_h$, $\eta < \mu_\eta$ e $\theta < \mu_\theta$; caso alguma delas não seja verdadeira, $\lambda_{ij}(h,\eta,\theta) \approx 0$. De acordo com a definição anteriormente fornecida para funções de $F(\alpha)$, valores significativamente intermediários entre 0 e 1 são encontrados apenas em um pequeno intervalo de transição.

O número total de ligações de hidrogênio formado por cada átomo doador ou aceptor i é dado pela

soma de todas as possíveis pontes de hidrogênio em que ele está envolvido:

$$\Lambda_i = \sum_j \lambda_{ij} \tag{4.10}$$

A contribuição energética total para cada átomo i, então, é dada por

$$E_i(\Lambda_i,r_i) = \frac{1}{2}\epsilon_{hb} f(r_i,\Lambda), \tag{4.11}$$

onde $\mu_r = 15\text{Å}$ e $\beta_r = 10\text{Å}^{-1}$, $\epsilon_{hb} = 5\epsilon$ e

$$f(r,\Lambda) = \begin{cases} F(r)(1-\Lambda) & \text{para } \Lambda \leq 0.95 \\ 0 & \text{para } \Lambda > 1.05 \end{cases} \tag{4.12}$$

com uma região quadrática intermediária para $0.95 < \Lambda < 1.05$, de modo a manter a diferenciabilidade em $\Lambda = 1$. Quando o procedimento de *annealing* é aplicado, o valor de ϵ_{hb} varia linearmente entre 0ϵ e 5ϵ ao longo da trajetória. A inserção gradual das ligações de hidrogênio modelada com o *annealing* tem como objetivo evitar a formação prematura de estruturas secundárias que prendam a estrutura em um mínimo local, antes que a simulação tenha a oportunidade de explorar uma região maior do espaço conformacional.

O valor constante de $f(r,\Lambda) = 0$ para $\Lambda > 1.05$ representa uma inovação em relação à primeira versão deste potencial, descrita em Pereira de Araújo (2009)[22] e evita contribuições energéticas favoráveis que poderiam surgir em conformações em que o mesmo átomo participa de mais de uma ligação de hidrogênio – uma situação fisicamente impossível em cadeias principais de proteínas.

O termo das ligações de hidrogênio, então, é dado por:

$$V_{\text{LH}} = \sum E_i(\Lambda_i,r_i) \tag{4.13}$$

Ou seja, a energia total das ligações de hidrogênio é dada pela soma das contribuições energéticas de cada um dos possíveis átomos doadores ou aceptores, de acordo com suas orientações relativas no espaço.

Conforme será discutido no capítulo seguinte, a satisfação ou não das restrições estabelecidas por esse termo tem um papel fundamental na análise da qualidade das estruturas produzidas pelo MDBury.

4.2 Algoritmo de dinâmica molecular

O sistema modelado pelo MDBury nas simulações descritas neste trabalho consiste em uma cadeia polipeptídica composta apenas por átomos pesados (sem hidrogênios), em um meio sem solvente explícito. O algoritmo executado consiste em uma simulação de dinâmica molecular aplicada ao potencial descrito nas seções anteriores.

Em métodos de dinâmica molecular, o comportamento do sistema é reproduzido ao longo do tempo de acordo com a aplicação das leis de Newton sobre cada um de seus átomos, representados como partículas pontuais. Uma vez que é impossível resolver analiticamente todo o sistema de equações associado ao cálculo da estrutura de uma macromolécula, esse tipo de simulação procede através de métodos de integração numérica, dividindo a trajetória em vários pequenos estágios, separados por um intervalo de tempo Δt. A força total atuante em cada partícula a cada instante t é calculada como a soma vetorial, nesse instante, de todas as forças que atuam sobre ela, de acordo com o potencial utilizado. Assumindo que essas forças permanecem constantes durante o intervalo Δt, é possível determinar suas respectivas acelerações e, com isso, suas posições e velocidades no instante $t + \Delta t$. O resultado da simulação é uma trajetória que especifica como as posições e velocidades de todas as partículas envolvidas variam ao longo do tempo.

Uma simulação de dinâmica molecular pode ser realizada de maneira mais simples no chamado *ensemble* microcanônico, em que a energia total do sistema é constante. Entretanto, para que os resultados da simulação sejam compatíveis com o comportamento de um sistema biológico, é frequentemente desejável que, ao invés disso, a temperatura do sistema seja mantida constante, no que é conhecido como o *ensemble* canônico. Esta propriedade pode ser obtida através do acoplamento do sistema a um banho térmico, que é um sistema externo com temperatura constante, capaz de regular a temperatura do sistema de interesse através da transferência de energia. Um algoritmo que gera um sistema com temperatura média constante em simulações de dinâmica molecular é conhecido como termostato[60]. Nosso algoritmo usa o método chamado termostato de Berendsen[61], que será descrito na seção 4.2.2.

A figura 4.7 esquematiza, em linhas gerais, a sequência de passos que caracteriza o algoritmo de dinâmica molecular implementado pelo programa MDBury.

4.2.1 Forças, velocidades e posições

O sistema modelado pelo MDBury consiste em um conjunto de N átomos, sendo cada átomo i definido por uma massa m_i, uma posição $\vec{s_i}$ no espaço tridimensional e uma velocidade $\vec{v_i}$. As velocidades $\vec{v_i}$

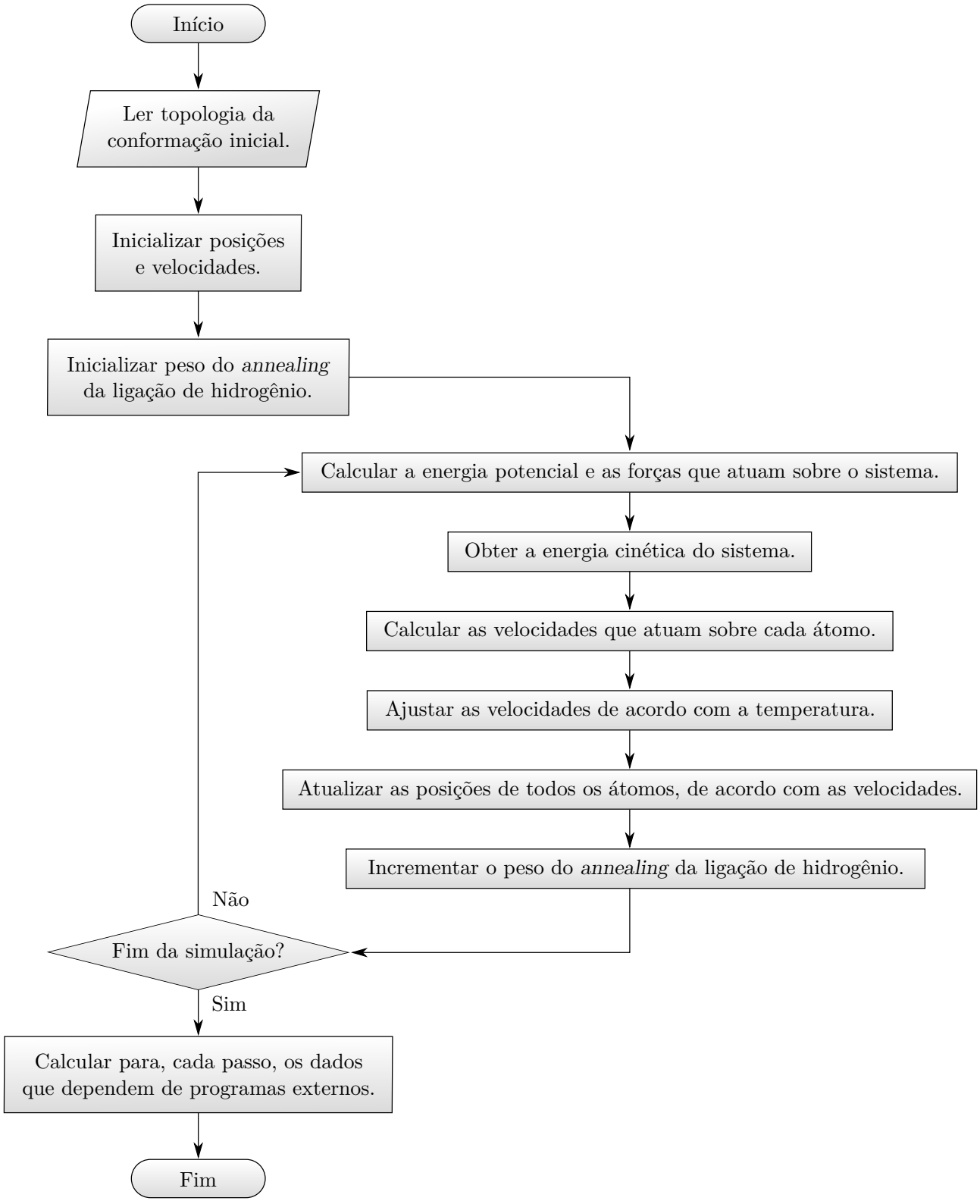


Figura 4.7: Algoritmo de simulação molecular adotado pelo MDBury.

são definidas aleatoriamente para a conformação inicial, e as posições iniciais $\vec{s_i}$ são lidas a partir de um arquivo em formato PDB que contém uma conformação completamente estendida da proteína modelada, gerada a partir de sua sequência de aminoácidos a partir de um script para o programa PyMOL[62].

Após todas as devidas inicializações, o programa entra em um *loop* principal que será repetido tantas vezes quantas forem o número de passos especificado nos dados de entrada. A principal computação realizada em cada passo é o cálculo da energia potencial V_i de cada átomo do sistema, assim como a força correspondente que atua sobre ele, $\vec{F_i}$, que é usada para atualizar a velocidade $\vec{v_i}$ de acordo com a segunda lei de Newton, multiplicada por um fator de escalonamento (equação 4.14). Em seguida, as posições de cada átomo são atualizadas de acordo com as velocidades calculadas (equação 4.15):

$$\vec{v_i}(t + \Delta t) = \lambda \left(\vec{v_i}(t) + \frac{\vec{F_i}}{m_i} \Delta t \right) \tag{4.14}$$

$$\vec{s_i}(t + \Delta t) = \vec{s_i}(t) + \vec{v_i}(t) \Delta t \tag{4.15}$$

onde λ é o fator de escalonamento dado pelo termostato de Berendsen, que será discutido a seguir, e $\Delta t = 0.005\tau$ é o intervalo de integração utilizado pelo algoritmo, sendo τ a unidade de tempo determinada pelas nossas unidades de distância, Å, massa, m, e energia, ϵ, isto é, $1\tau = 1\text{Å}\sqrt{m/\epsilon}$.

4.2.2 Termostato de Berendsen

A temperatura instantânea T de um sistema molecular generalizado pode ser definida em qualquer ponto no tempo como

$$T = \frac{2}{k_B N_{df}} K \tag{4.16}$$

onde K é a energia cinética instantânea, e N_{df} é o número de graus de liberdade internos do sistema. Para um sistema modelado no vácuo, como é o caso das simulações descritas neste trabalho, $N_{df} = N - 6$, sendo N o número total de átomos[60].

A energia cinética instantânea do sistema é dada por

$$K = \frac{1}{2} \sum_{i=1}^{N} m_i \left| \vec{v_i} \right|^2 \tag{4.17}$$

onde as velocidades internas $\vec{v_i}$ são calculadas a partir das velocidades $\vec{v_i}$ através da exclusão de qualquer componente ao longo dos graus de liberdade externos, neste caso, o vetor coordenada do centro de massa do sistema, o momento angular sobre o centro de massa e o tensor de inércia do sistema relativo ao centro de massa[60]. No nosso sistema, o valor de m é unitário para todos os

átomos.

No MDBury, o sistema simulado está acoplado a um banho térmico que o mantém a uma temperatura média constante, definida como $T_0 = 1\epsilon$. Este acoplamento é executado por um algoritmo conhecido como termostato de Berendsen[61;63]. No termostato de Berendsen, as velocidades do sistema são reescalonadas a cada intervalo de integração, de modo que a mudança de temperatura é proporcional à diferença entre a temperatura atual T e a temperatura desejada T_0:

$$\frac{dT}{dt} = \frac{1}{\tau}(T_0 - T) \tag{4.18}$$

onde τ é o parâmetro que determina o grau de acoplamento entre o banho e o sistema. Nas simulações descritas neste trabalho, para manter a simplicidade dos cálculos, $\tau = 1$. Este método tem o efeito de criar um decaimento exponencial do sistema em direção à temperatura desejada:

$$T = T_0 - Ce^{-t/\tau} \tag{4.19}$$

A equação 4.18 também implica em

$$\Delta T = \frac{\Delta t}{\tau}(T_0 - T) \tag{4.20}$$

Que leva à definição do fator de escalonamento λ, usando na equação 4.14:

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_0}{T} - 1\right)} \tag{4.21}$$

4.2.3 Compilação de resultados

Além das informações já registradas pelo programa ao longo da execução, dois outros dados que serão importantes na análise dos resultados precisam ser calculados separadamente para cada conformação registrada, utilizando programas externos. O desvio médio quadrático (*root mean square deviation*, RMSD) de cada conformação em relação à estrutura nativa é calculado de acordo com o algoritmo de McLachlan[64], conforme implementado pelo programa ProFit[65]. A razão de resíduos que adotam estruturas secundárias regulares (α-hélices ou folhas β) em cada conformação é calculada de acordo com as definições empregadas pelo programa DSSP[1].

O próximo capítulo descreve as trajetórias e as conformações finais obtidas a partir de simulações executadas com o MDBury para as três proteínas de diferentes classes estruturais selecionadas anteriormente em nosso banco de dados de predições. Os resultados obtidos serão discutidos, em

grande parte, em termos da comparação entre os dois parâmetros descritos acima e a satisfação de diferentes termos da função de energia potencial, de modo a tentar compreender o comportamento das simulações e as razões pelas quais diferentes trajetórias são ou não capazes de atingir conformações próximas à nativa utilizando o método apresentado.

Capítulo 5

Simulação de estruturas: Resultados e Discussão

O algoritmo de dinâmica molecular apresentado no capítulo anterior foi utilizado para realizar simulações do enovelamento proteico das três estruturas selecionadas anteriormente no banco de dados, utilizando as predições descritas no final do capítulo 3 como parâmetros do termo dos enterramentos atômicos. Este capítulo descreve essas simulações e analisa os seus resultados.

Todas as trajetórias de simulação descritas a seguir partiram de conformações estendidas das proteínas e empregaram o mecanismo de *annealing* do termo das ligações de hidrogênio, que tem como efeito aumentar linearmente o peso deste termo ao longo da trajetória. Os resultados descritos configuram predições computacionais de estruturas proteicas obtidas de maneira *ab initio*, ou seja, em última análise, apenas a partir das respectivas sequências de aminoácidos.

A discussão que segue tem dois objetivos principais. É necessário, primeiramente, avaliar se os enterramentos atômicos preditos são capazes de conduzir a simulação do enovelamento a conformações próximas à nativa para estas proteínas, que é o objetivo principal deste trabalho. Adicionalmente, a natureza das trajetórias observadas na simulação será analisada em função das características do algoritmo implementado, com o propósito de compreender os motivos e as circunstâncias nas quais o método apresentado pode ser capaz de atingir os seus objetivos.

Os resultados apresentados neste capítulo também são discutidos em um artigo, cujo texto também integra esta tese, que foi aceito em outubro de 2013 para publicação na revista *Proteins: Structure, Function, and Bioinformatics*.

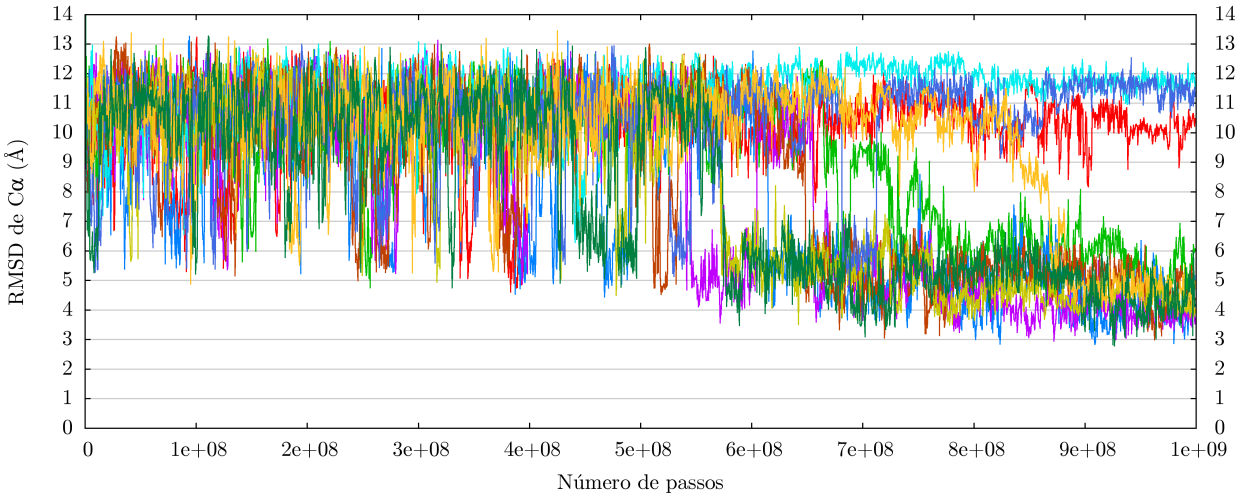


Figura 5.1: Evolução do RMSD de C_α em relação à estrutura nativa ao longo de dez trajetórias independentes da simulação do enovelamento da proteína efetora RxLR (RePc), partindo de conformações estendidas. Cada linha representa uma trajetória.

5.1 Proteína efetora RxLR (RePc)

A primeira estrutura escolhida em nosso conjunto de testes para simulação é a proteína efetora RXLR de *Phytophthora* (RePc), que tem uma topologia formada por quatro α-hélices (código PDB: 3zr8)[55]. *Phytophthora* é um importante gênero de micro-organismos eucarióticos causador de infecções em plantas. Proteínas conhecidas como efetoras são secretadas por esses patógenos para acelerar o processo de infecção das células hospedeiras[66]. Uma vez que os hospedeiros potenciais de *Phytophthora* incluem várias espécies importantes para a agricultura, seus efeitos são frequentemente capazes de causar danos econômicos substanciais[67].

Em nossas simulações, dez trajetórias independentes de simulação foram executadas para a proteína RePc, todas partindo de conformações completamente estendidas. A figura 5.1 ilustra a variação ao longo do tempo do desvio quadrático médio (*Root Mean Square Deviation* - RMSD) da posição dos átomos C_α de cada uma dessas trajetórias em relação à estrutura nativa. Sete das dez trajetórias executadas resultaram em um enovelamento correto, com a formação apropriada das hélices que compõem a topologia nativa e valores de RMSD de C_α entre 3Å e 5Å.

Para possibilitar uma análise mais detalhada do comportamento da simulação, a figura 5.2 destaca uma das trajetórias que alcançou uma conformação similar à nativa, comparando a variação ao longo do tempo de dois termos característicos de nosso potencial: o termo dos enterramentos e o termo das ligações de hidrogênio. É claramente perceptível que esses dois termos têm comportamentos bastante distintos: enquanto o primeiro se mantém constante em torno de 80ϵ ao longo de toda a simulação, o segundo, a princípio, cresce linearmente, como consequência do procedimento de *annealing* que lhe

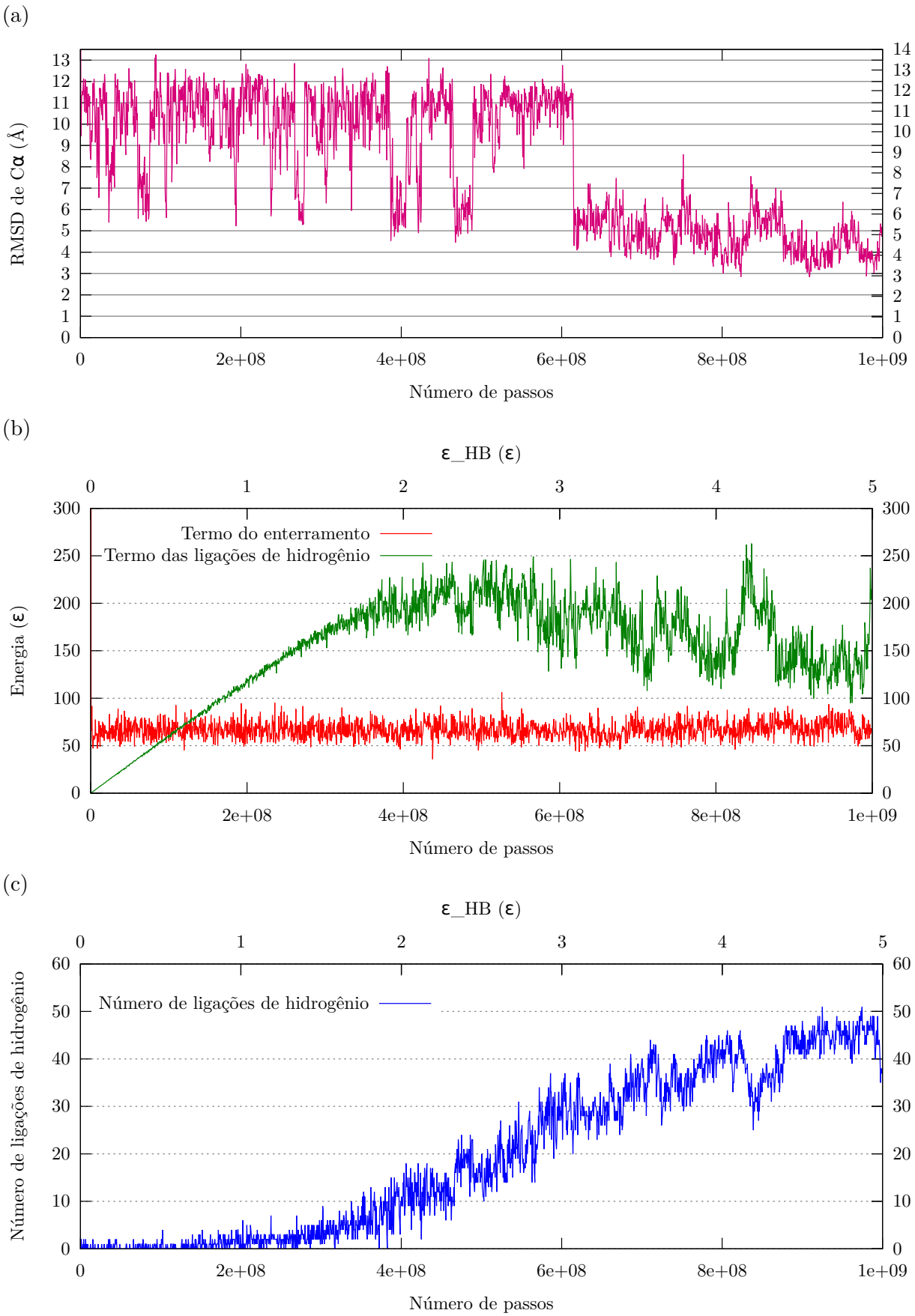


Figura 5.2: Detalhamento da variação de diferentes medidas ao longo de uma única trajetória da simulação da proteína RePc. O eixo horizontal superior indica o peso do termo das ligações de hidrogênio, ϵ_{hb}. O eixo vertical indica: (a) o RMSD de C_α em relação à estrutura nativa; (b) os termos de energia do enterramento e das ligações de hidrogênio; (c) o número de ligações de hidrogênio formadas pela estrutura.

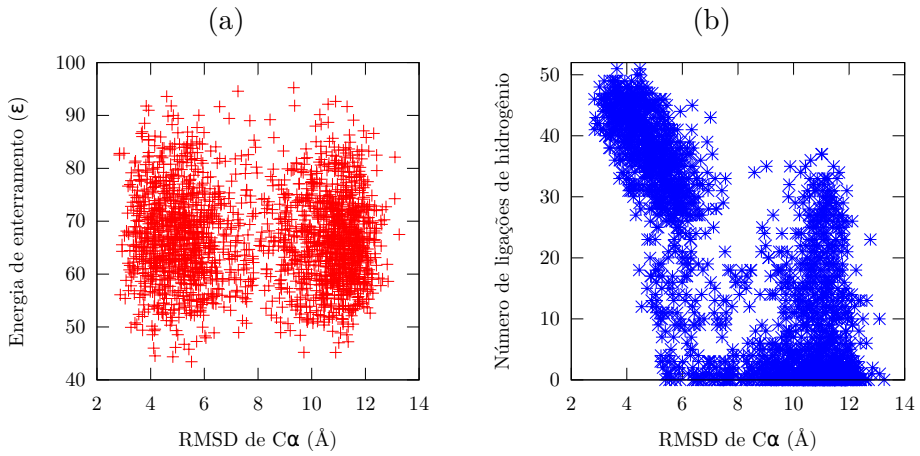


Figura 5.3: Para a mesma trajetória apresentada na figura 5.2, cada ponto representa uma conformação. O eixo horizontal indica o RMSD de C_α em relação à estrutura nativa, e o eixo vertical indica: (a) a energia do termo do enterramento; (b) o número de ligações de hidrogênio formadas pela estrutura.

é aplicado, e depois tende a se estabilizar em torno de 150ϵ (figura 5.2-b). O número de ligações de hidrogênio formadas pela estrutura a cada passo, de acordo com a definição estipulada pelo respectivo potencial, também começa crescendo linearmente e depois acompanha o inverso das variações do termo energético (figura 5.2-c).

Essas observações fornecem um indicativo importante sobre o papel dos diferentes aspectos do potencial na obtenção dos resultados da simulação. O fato de que o termo de enterramento se mantém praticamente inalterado ao longo de toda a trajetória indica que, por si só, ele não deve ser capaz de distinguir a conformação nativa dos enovelamentos incorretos alcançados ao longo da simulação.

Como confirma a figura 5.3-a, não há qualquer correlação, nessa trajetória, entre os valores alcançados pelo potencial de enterramentos e os respectivos RMSDs de C_α obtidos em relação à estrutura nativa. Por outro lado, a figura 5.3-b mostra claramente que as estruturas que são capazes de formar um alto número de ligações de hidrogênio ao longo da trajetória atingem conformações com RMSD consistentemente mais baixo que as demais.

A figura 5.4 estende essa análise para todas as dez trajetórias, comparando para os últimos 10% passos de cada uma delas os valores médios do termo das ligações de hidrogênio V_{PH} e do RMSD de C_α em relação à estrutura nativa. Como pode ser visto na figura 5.4-a, o termo V_{PH} fornece um excelente indicador de qualidade da predição: com exceção de apenas um caso, as trajetórias que obtiveram valores energéticos mais baixos que as demais (entre 140ϵ e 150ϵ) para este termo foram as mesmas que alcançaram os menores valores de RMSD, por volta de 4Å (o ponto com RMSD $\approx$ 2Å corresponde à estrutura nativa relaxada, descrita adiante). A única trajetória que terminou com um valor médio para V_{PH} menor que 160ϵ e alto RMSD corresponde a uma estrutura final que tem as quatro hélices corretamente formadas, mas dispostas em um arranjo mútuo invertido, formando

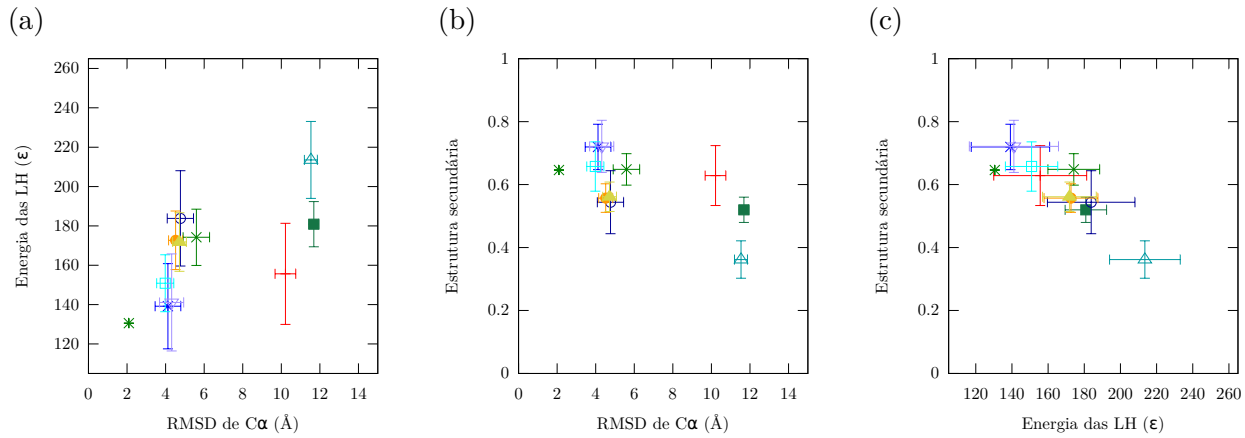


Figura 5.4: Parâmetros calculados para as trajetórias da simulação da proteína RePc. Cada ponto corresponde à média dos 10% últimos passos de uma trajetória independente, com exceção do ponto mais à esquerda nos dois primeiros gráficos, que corresponde aos valores da estrutura nativa relaxada. Barras de erro indicam os desvios-padrão correspondentes. Conforme indicado nos rótulos, os eixos horizontal e vertical de cada gráfico podem corresponder ao RMSD de C_α em relação à estrutura nativa (a,b), ao valor da energia potencial das ligações de hidrogênio (a,c) ou à fração de resíduos que adotam estrutura secundária regular (b,c), conforme calculado pelo DSSP [1].

uma topologia simétrica à da estrutura nativa (figura 5.5-d). Não houve nenhum caso em que uma trajetória foi capaz de simultaneamente satisfazer o termo das ligações de hidrogênio e chegar a uma topologia distinta tanto da nativa quanto de sua imagem espetacular.

Esses resultados sugerem uma metodologia que permite selecionar a melhor estrutura predita entre todas as trajetórias, de maneira a prescindir de comparações com a conformação nativa para a seleção da estrutura final. Primeiro, selecionamos a trajetória que alcançou o menor valor médio para V_{PH} em seus passos finais. Dentro dessa trajetória, não é conveniente utilizar o mesmo critério para selecionar a melhor conformação, porque a aplicação do *annealing* tende a obscurecer a relação entre valor energético e a efetiva satisfação das restrições correspondentes. Em vez disso, selecionamos a conformação que formou o maior número de ligações de hidrogênio, utilizando a energia apenas como eventual critério de desempate. A figura 5.5-c mostra a estrutura selecionada de acordo com essa metodologia, em comparação com a estrutura de menor RMSD (figura 5.5-b), a estrutura nativa (5.5-a) e com uma estrutura selecionada a partir da trajetória que resultou em uma imagem especular (5.5-d).

Além das estruturas resultantes de simulações *ab initio*, foi produzida também, para fins de comparação, uma conformação da RePc que denominamos "estrutura nativa relaxada". Essa estrutura é o resultado de uma simulação muito curta (5000 passos) realizada a partir da estrutura nativa obtida experimentalmente, utilizando valores nativos em quatro camadas para o termo dos enterramentos, sem *annealing* das ligações de hidrogênio. A estrutura assim obtida tem um RMSD de C_α aproximadamente igual a 2Å em relação à conformação nativa original e não apresenta qualquer alteração

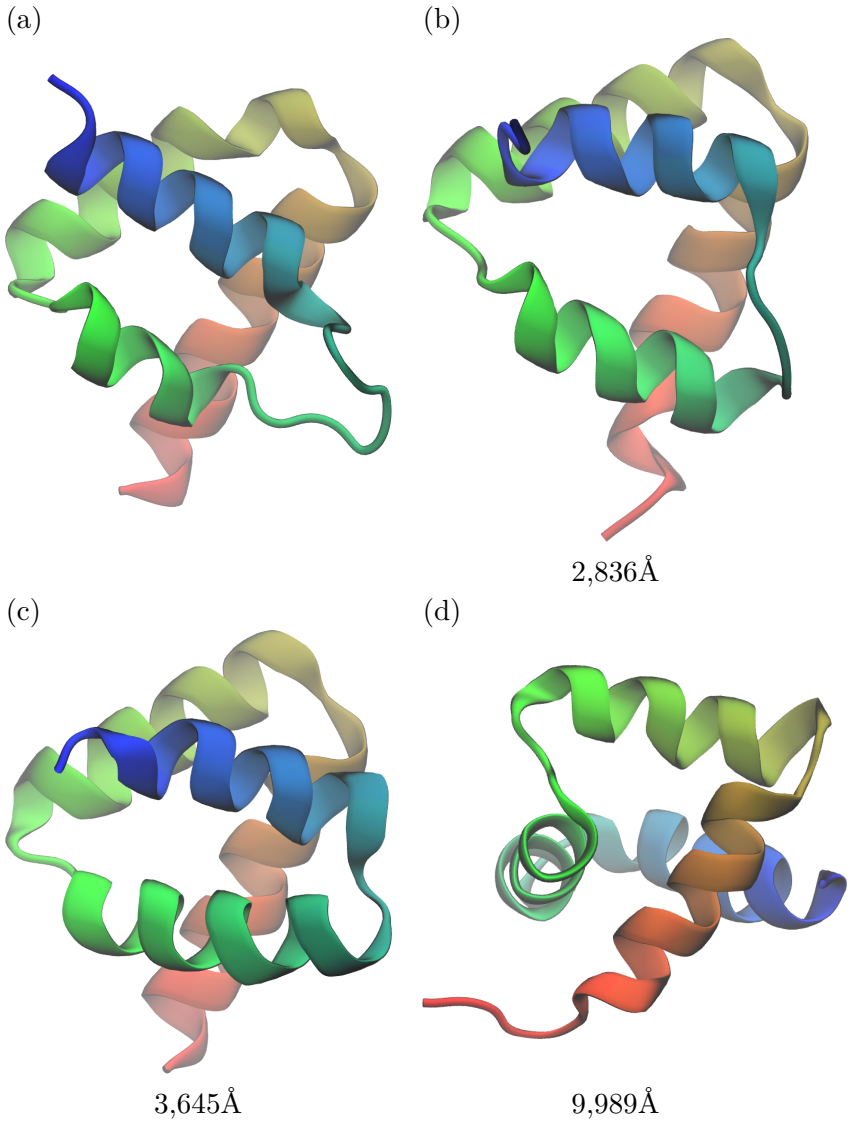


Figura 5.5: (a) Estrutura nativa da proteína efetora RxLR – RePc; (b) Estrutura predita com menor RMSD de C_α em relação à nativa; (c) Estrutura predita selecionada de acordo com critérios de qualidade independentes da conformação nativa; (d) Estrutura "espelho" com alto grau de formação de ligações de hidrogênio e alto RMSD. Os RMSDs de C_α estão indicados abaixo de cada estrutura.

significativa em sua topologia, sendo caracterizada principalmente por ajustes finos na conformação das ligações covalentes e pequenas variações nas orientações relativas dos átomos que participam de ligações de hidrogênio.

Esses ajustes são importantes para a discussão dos resultados porque, por exemplo, os ângulos e distâncias formados entre átomos doadores e aceptores de ligações de hidrogênio na estrutura nativa determinada experimentalmente podem não corresponder exatamente aos valores mais estritos definidos pelo nosso potencial. Dessa forma, calcular o valor desse termo diretamente na estrutura nativa original resultaria em valores artificialmente altos, que não são comparáveis com aqueles que podem ser alcançados pelo algoritmo em simulações *ab initio*. Adicionalmente, devido às aproximações adotadas pelo potencial, o relaxamento da estrutura nativa tem o efeito de sinalizar o RMSD mínimo teórico que poderia a princípio ser alcançado pelo nosso algoritmo, fornecendo também um parâmetro de comparação para os valores de RMSD efetivamente atingidos pelas simulações.

Conforme indicado na figura 5.4-a, o valor do termo das ligações de hidrogênio V_{PH} na estrutura nativa relaxada é igual a 130,55ϵ, que está dentro da faixa de valores atingida pelas conformações finais da trajetória de menor energia. Esse valor é praticamente idêntico ao atingido pela estrutura de menor RMSD nessa trajetória (figura 5.5-b), que tem $V_{\mathrm{PH}} = 131{,}31\epsilon$. Para a conformação selecionada de acordo com os critérios descritos anteriormente (figura 5.5-c) $V_{\mathrm{PH}} = 94{,}79\epsilon$.

5.2 Proteína G de estreptococos (ProtGSsp)

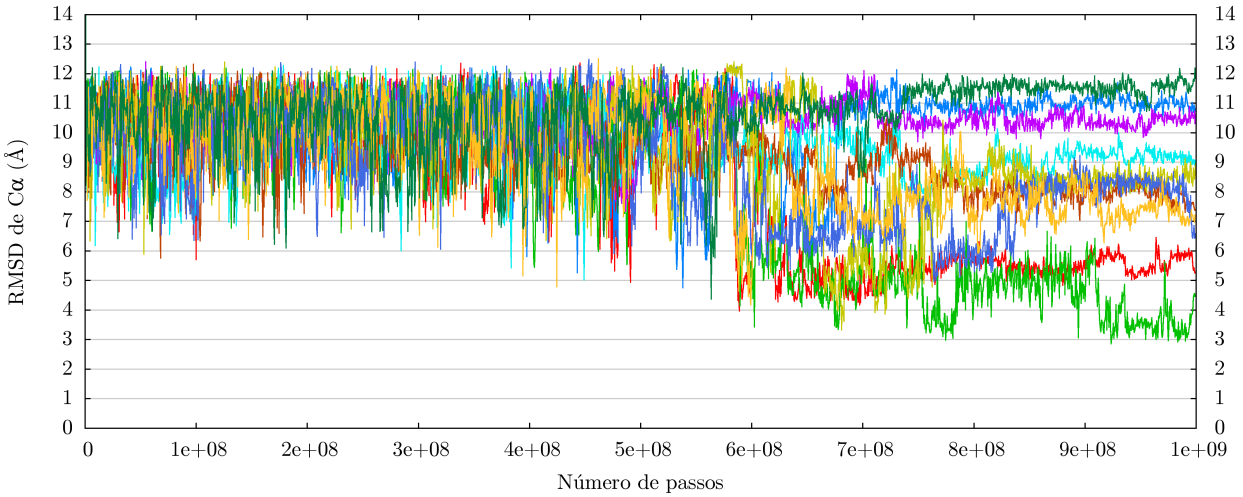


Figura 5.6: Evolução do RMSD de C_α em relação à estrutura nativa ao longo de dez trajetórias independentes da simulação do enovelamento da proteína G de estreptococos (ProtGSsp), partindo de conformações estendidas. Cada linha representa uma trajetória.

A segunda estrutura selecionada para simulação neste trabalho é o domínio β1 da proteína G (código PDB: 3fil)[56], que tem uma topologia formada por uma α-hélice e quatro folhas β. A proteína

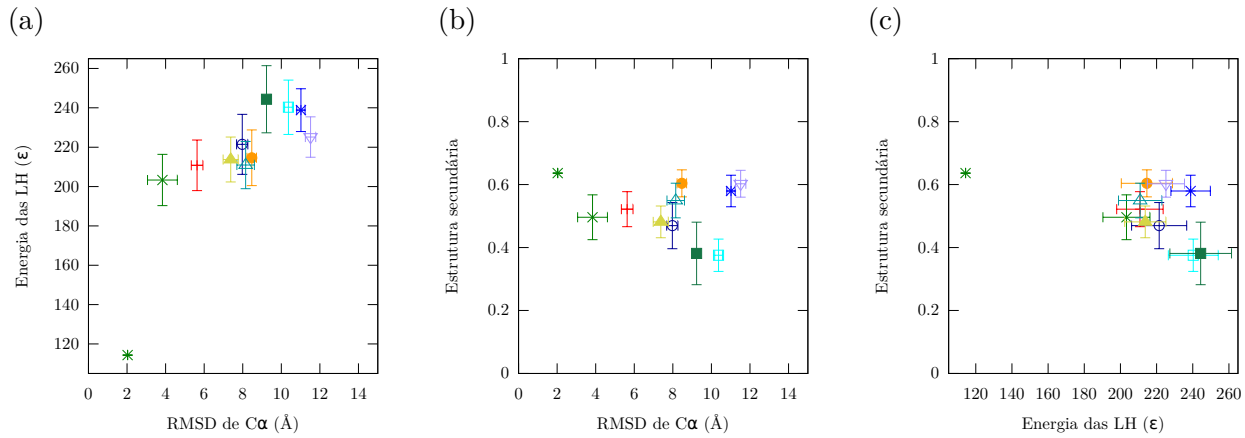


Figura 5.7: Parâmetros calculados para as trajetórias da simulação da ProtGSsp. São seguidas as mesmas convenções da figura 5.4.

G é uma molécula expressa em certos grupos de bactérias estreptococos com a função de se ligar a imunoglobulinas, e é extensivamente utilizada na indústria biotecnológica para a purificação de anticorpos[68]. O domínio β1 da proteína G é comumente empregado como estrutura modelo em estudos do enovelamento proteico e de predição de estruturas de proteínas[41;69].

Para essa estrutura, a distribuição dos valores de RMSD de C_α nos resultados de simulação foi mais heterogênea que o observado para a RePc, com duas das dez trajetórias alcançando valores finais entre 3Å e 6Å (figura 5.6). Assim como no caso anterior, os valores médios finais de V_{PH} de cada trajetória são capazes de distinguir entre diferentes valores de RMSD (figura 5.7), sendo que os dois menores valores de energia correspondem de maneira apropriada aos melhores valores de similaridade. Ao contrário do que aconteceu com a RePC, entretanto, os valores atingidos para o termo das ligações de hidrogênio V_{PH}, mesmo na trajetória de menor energia (aproximadamente entre 190ϵ e 220ϵ), são significativamente mais altos que os calculados para a estrutura nativa relaxada (114.35ϵ). Esse resultado indica que conformações mais próximas à nativa poderiam, a princípio, ser atingidas por simulações que utilizam o potencial adotado, possivelmente com melhores predições de enterramento, mas sem a necessidade de modificações no termo das ligações de hidrogênio.

A figura 5.8-c mostra a estrutura selecionada de acordo com os mesmos critérios usados para a RePc, em comparação com a estrutura de menor energia (figura 5.8-b) e com a estrutura nativa (figura 5.8-a).

5.3 Proteína de choque frio Bc-Csp de *Bacillus caldolyticus* (CsBc)

A terceira estrutura selecionada para simulação neste trabalho, com o código PDB 1c9o[57], é a proteína de choque frio Bc-Csp da bactéria *Bacillus caldolyticus* (CsBc). Proteínas de choque frio são uma família de proteínas de estrutura altamente conservada, encontrada em organismos tão diversos como

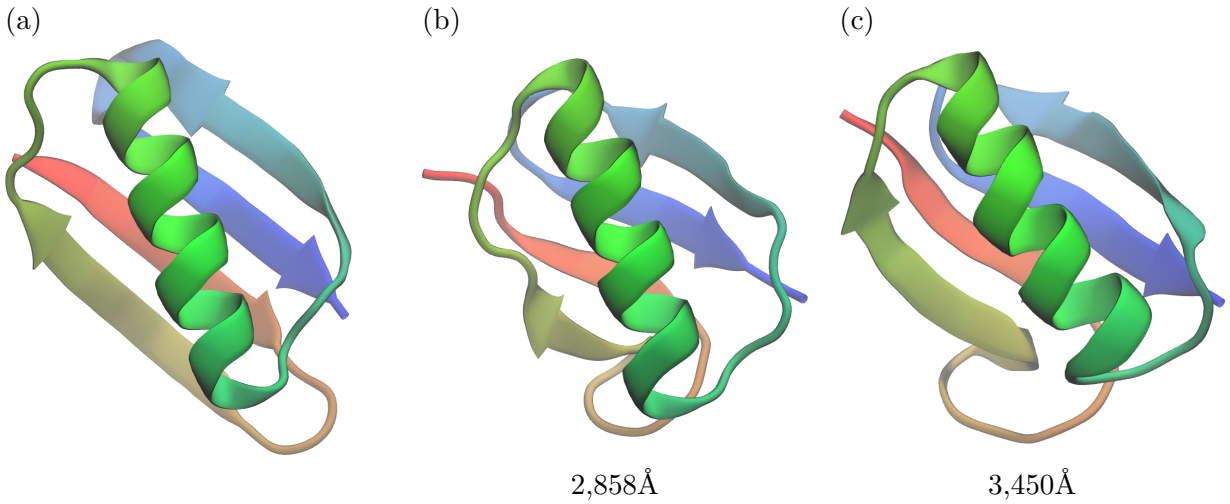


Figura 5.8: (a) Estrutura nativa da proteína G de estreptococos – ProtGSsp; (b) Estrutura predita com menor RMSD de C_α em relação à nativa; (c) Estrutura predita selecionada de acordo com critérios de qualidade independentes da conformação nativa. Os RMSDs de C_α estão indicados abaixo de cada estrutura.

bactérias e vertebrados. Essas proteínas, que têm uma topologia em forma de barril β, têm a ação induzida principalmente como resposta a quedas bruscas de temperatura no ambiente celular, quando se ligam a cadeias simples de DNA ou RNA para regular os níveis de transcrição e tradução genética em um estado de estresse térmico[70].

O fato de folhas β serem estabilizadas por interações não-locais dificulta a formação de estruturas parciais estáveis ao longo da simulação e faz com que conformações errôneas fiquem mais facilmente presas em mínimos locais. Provavelmente por esses motivos, esta proteína foi a mais desafiadora de nosso conjunto-teste. Das dezoito trajetórias executadas, três resultaram em valores de RMSD de C_α entre 5Å e 6Å (figura 5.9).

Ao contrário do que ocorreu com as outras duas proteínas, entretanto, o valor médio final da energia das ligações de hidrogênio nesse caso não é capaz de distinguir entre trajetórias que resultaram ou não em conformações próximas à nativa (figura 5.10-a).

Uma análise visual das trajetórias que terminaram em conformações com alto RMSD, entretanto, revela que essas estruturas formam uma quantidade muito pequena de estruturas secundárias regulares, e nenhuma delas poderia ser realisticamente confundida com uma proteína real (figura 5.11). O cálculo da razão de resíduos que adotam α-hélices ou folhas β, realizado de maneira independente de nosso potencial com o auxílio do programa DSSP[1], confirma essa observação. De fato, como pode ser observado na figura figura 5.10-b, apenas duas das trajetórias apresentam em suas porções finais uma porcentagem de resíduos maior que 50% participando de estruturas secundárias regulares, e estas são exatamente as duas trajetórias de menor RMSD. Nas simulações anteriores, podemos notar que uma das trajetórias de RePc e duas das trajetórias de ProtGSsp (figuras 5.4-b e 5.7-b) também haviam

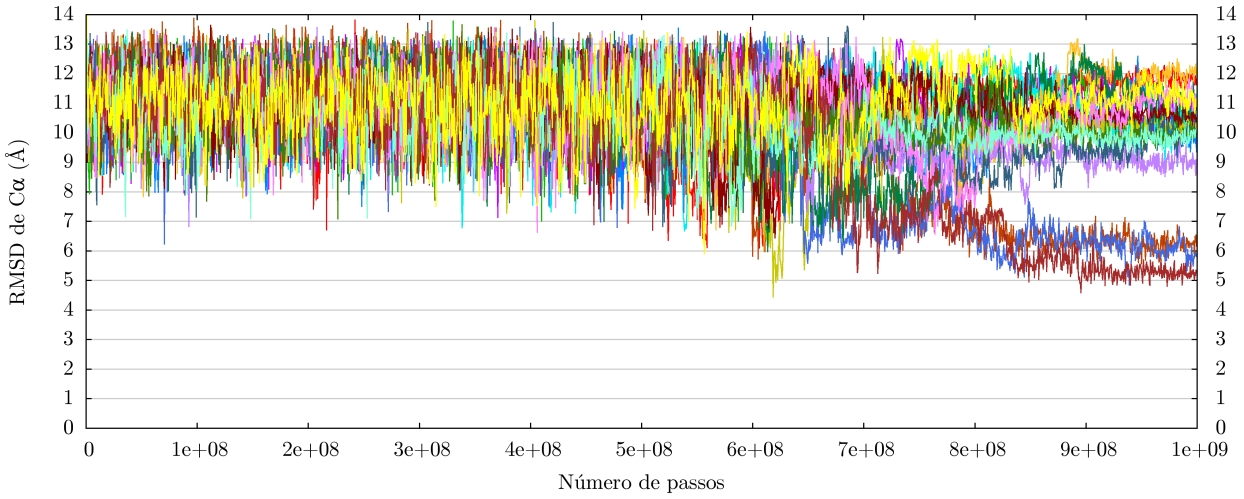


Figura 5.9: Evolução do RMSD de C_α em relação à estrutura nativa ao longo de dezoito trajetórias independentes da simulação do enovelamento da proteína de choque frio Bc-Csp de *Bacillus caldolyticus* (CsBc), partindo de conformações estendidas. Cada linha representa uma trajetória.

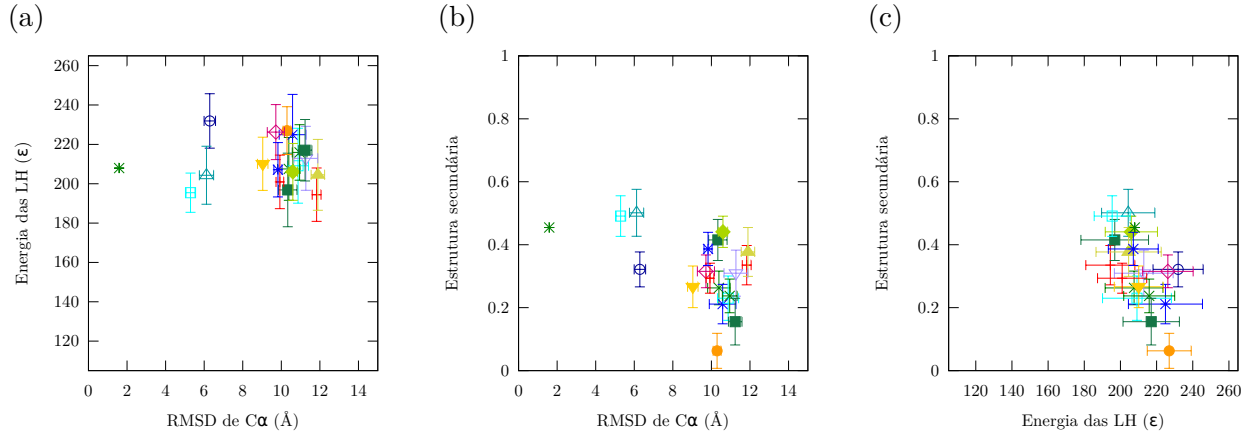


Figura 5.10: Parâmetros calculados para as trajetórias da simulação da proteína Bc-Csp. São seguidas as mesmas convenções da figura 5.4.

apresentado uma baixa fração de estruturas secundárias, mas ao contrário do que ocorreu com CsBc, essas mesmas trajetórias já haviam sido corretamente distinguidas das demais pelo valor da energia das ligações de hidrogênio.

Se eliminarmos, então, as trajetórias de CsBc que não foram capazes de formar estruturas secundárias regulares e selecionarmos entre as restantes a melhor conformação de acordo com os mesmos critérios anteriormente usados, chegamos à conformação disposta na figura 5.12-c.

5.4 Discussão

Os resultados apresentados nas seções anteriores descrevem a obtenção de conformações semelhantes à nativa em simulações *ab initio* do enovelamento proteico que combinam predições de enterramento realizadas a partir da sequência com restrições estruturais independentes da sequência, as quais estão

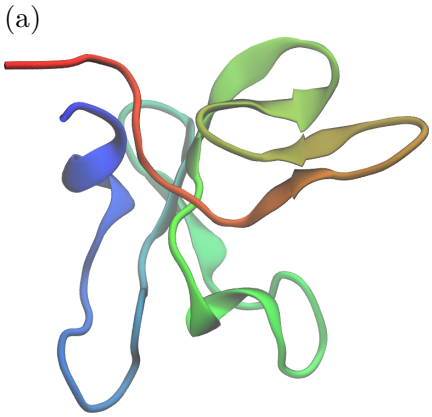


Figura 5.11: Um exemplo típico de estrutura não-realista, com alto RMSD e baixa energia de ligações de hidrogênio, formada para a proteína Bc-Csp em algumas trajetórias das simulações do enovelamento.

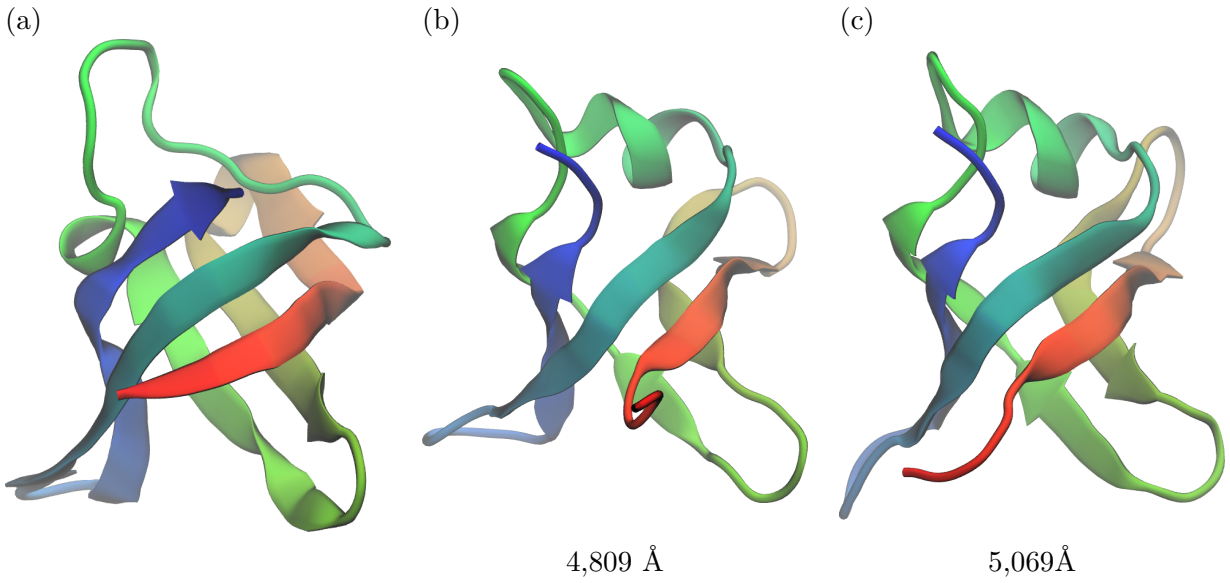


Figura 5.12: (a) Estrutura nativa da proteína de choque frio Bc-Csp de *Bacillus caldolyticus* – CsBc; (b) Estrutura predita com menor RMSD de C_α em relação à nativa; (c) Estrutura predita selecionada de acordo com critérios de qualidade independentes da conformação nativa. Os RMSDs de C_α estão indicados abaixo de cada estrutura.

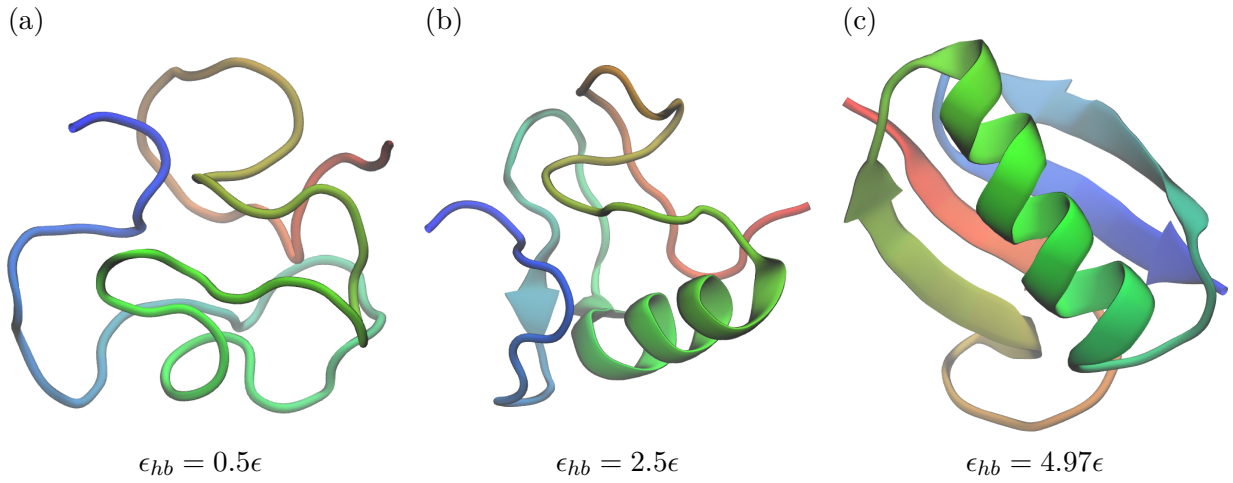


Figura 5.13: Três conformações diferentes assumidas pela proteína ProtGSssp ao longo de uma trajetória, à medida que cresce o peso do termo das ligações de hidrogênio ϵ_{hb}.

relacionadas principalmente à topologia das ligações covalentes à formação de ligações de hidrogênio. Esses resultados validam a nossa hipótese inicial de que informações de enterramento codificadas na sequência podem se constituir, ao menos para algumas classes de estruturas, no único elemento dependente de sequência necessário para determinar a conformação nativa de um proteína. Estudos que englobem uma gama maior de tipos de sequências e de estruturas proteicas serão necessários para uma avaliação mais completa do alcance total do método proposto.

Para que uma metodologia de simulação molecular possa ser aplicada com sucesso à predição estrutural de novas proteínas, ela precisa ser capaz de não apenas chegar a bons resultados em uma fração significativa das trajetórias executadas, mas também de distinguir as predições corretamente alcançadas sem ter a estrutura nativa disponível para comparação. Um padrão interessante que emergiu em nossos resultados é que a formação das ligações de hidrogênio, e não a satisfação do termo de enterramentos, revelou-se como o principal indicador de qualidade das predições.

As figuras 5.1, 5.6 e 5.9 mostram a evolução do RMSD de C_α em relação às estruturas nativas de cada conformação ao longo das diferentes trajetórias. Um padrão claramente perceptível nos três gráficos é que o RMSD oscila sempre de maneira substancial ao longo da primeira metade da simulação, e depois disso tende a se estabilizar para cada trajetória em torno de um intervalo curto de valores.

Esse comportamento decorre do fato de que, no início da simulação, embora o termo dos enterramentos esteja atuando normalmente, o termo das ligações de hidrogênio está desligado devido ao *annealing*. Como consequência, a proteína se compacta rapidamente, e a maior parte dos átomos é conduzida com sucesso às camadas preditas, sem formar estruturas secundárias regulares. Tais condições geram uma amostragem do espaço conformacional que, embora restrita, ainda é insuficiente para permitir a clara distinção da estrutura nativa. Em seguida, à medida que o termo das ligações de hidrogênio se torna mais importante na simulação, o espaço conformacional disponível à trajetória se

torna cada vez mais restrito, e a variação entre as diferentes conformações adotadas é reduzida (figura 5.13). Em alguns casos, conforme visto, a conformação final adotada pela trajetória tem topologia idêntica à nativa; em outros, é formada uma estrutura errônea. Os últimos exemplos costumam ser facilmente distinguidos dos primeiros pela pouca formação de estruturas secundárias.

Esse cenário sugere uma explicação para os papeis de ambos os termos na geração da estrutura final: as restrições impostas pelos enterramentos têm o efeito prático de limitar o espaço de busca disponível para o algoritmo às conformações em que cada átomo está localizado em sua camada predita ou próximo a ela. Essas restrições efetivamente bloqueiam todas as outras topologias diferentes da nativa que seriam, a princípio, fisicamente plausíveis para a sequência dada. Ainda assim, o potencial resultante continua compatível com um número proibitivamente alto de padrões de enovelamento incorretos. A seleção do padrão nativo dentro desta amostra do espaço conformacional, então, é realizada por meio da aplicação do potencial das ligações de hidrogênio, o qual remove da busca todas as conformações que não formam estruturas secundárias regulares.

Essa perspectiva também explica por que a energia dos enterramentos comumente se mantém constante ao longo de toda a simulação (figura 5.2-b) e por que, com exceção de uma topologia simétrica à estrutura nativa, não foi observada em nenhuma trajetória o surgimento de conformações finais errôneas mas ricas em estruturas secundárias.

Apesar desses sucessos, também é importante destacar o fato de que, na proteína com topologia de barril β CsBc, algumas trajetórias resultaram em conformações que conseguem satisfazer as restrições impostas pelo termo das ligações de hidrogênio sem, no entanto, formar topologias dominadas por estruturas secundárias regulares. O surgimento dessas conformações implica que a definição de ligações de hidrogênio implementada pelo nosso potencial poderia ser ainda mais estrita, sugerindo uma direção para possíveis melhoramentos em nosso campo de força.

Capítulo 6

Conclusões e Perspectivas

Este trabalho apresentou a primeira versão do que pretende vir a se tornar futuramente uma metodologia completa para predição *ab initio* de estruturas de proteínas a partir de sequências de aminoácidos, desenvolvida a partir da hipótese de que o código do enovelamento está escrito na linguagem dos enterramentos atômicos, os quais constituem a única informação dependente de sequência necessária para determinar a estrutura nativa de uma proteína.

Os primeiros estudos que apontaram para a validade dessa hipótese[21;22] foram publicados poucos anos antes do início deste projeto e serviram como motivação para a sua proposta. Esses trabalhos já haviam demonstrado que enterramentos atômicos obtidos a partir da estrutura nativa são suficientemente informativos para guiar uma simulação do enovelamento de uma proteína até a sua conformação tridimensional correta e levantaram a proposta de que resultados similares poderiam ser obtidos utilizando predições de enterramento obtidas unicamente a partir da sequência.

Nossa metodologia foi dividida em duas etapas independentes. Primeiramente, foi desenvolvido um algoritmo de aprendizado supervisionado para predição de enterramentos atômicos a partir de sequências de aminoácidos, baseado em um método originalmente concebido para predição de estruturas secundárias. Os resultados obtidos com esse algoritmo foram avaliados por meio de comparações com estudos informacionais sobre a relação entre sequências e enterramentos, realizados concomitantemente a este trabalho como parte do projeto de mestrado desenvolvido em nosso laboratório por Juliana Ribeiro Rocha[49]. Esses estudos permitiram que se chegasse à conclusão de que o algoritmo de predição desenvolvido tem uma performance próxima ao ótimo, no sentido de que ele é capaz de extrair praticamente toda a informação sobre enterramentos que pode ser concebivelmente codificada na sequência, de acordo com o modelos de dado adotado.

A próxima etapa, então, foi a construção de um método que permitisse que essas predições pudessem ser usadas como potenciais dependentes de sequência em uma simulação do enovelamento proteico.

O algoritmo de dinâmica molecular usado para essas simulações se baseou na mesma metodologia empregada nos estudos originais que envolveram enterramentos nativos, com uma implementação que foi continuamente refinada e aprimorada ao longo deste trabalho. O procedimento resultante foi capaz de obter e distinguir com sucesso estruturas próximas à nativa para três proteínas globulares pertencentes a diferentes classes estruturais, a partir de trajetórias que iniciaram de conformações estendidas.

Esses resultados representam, em última análise, a validação da ideia inicial que motivou todo o trabalho, demonstrando que, ao menos para um conjunto restrito de proteínas, enterramentos atômicos podem atuar como os únicos intermediários informacionais entre sequência e estrutura. Esperamos que a metodologia descrita possa ser vista como o marco inicial de uma nova abordagem em métodos de predição de estruturas de proteínas, assim como uma importante contribuição à nossa compreensão dos mecanismos que regem o enovelamento proteico.

6.1 Perspectivas

Durante a trajetória percorrida até a obtenção dos resultados descritos neste trabalho, diversas oportunidades para refinamentos do método foram mapeadas, e vários novos caminhos emergiram como potenciais vias de investigação.

Algumas das linhas de pesquisa relacionadas e este projeto que podem ser exploradas em trabalhos futuros são:

(1) O desenvolvimento de um processo que permita correlacionar a qualidade das predições de enterramento às distribuições de probabilidades emitidas pelo algoritmo de predição, assim como a identificação dos mínimos requisitos de qualidade de predição necessários para possibilitar a determinação da estrutura nativa de uma proteína.

(2) Uma investigação das diversas possibilidades de partição do volume da proteína em camadas, possivelmente com variações para cada átomo, dependendo de parâmetros locais, de modo a permitir o melhor aproveitamento da informação contida em predições com diferentes graus de qualidade.

(3) O estudo dos diversos fatores que influenciam a formação de estruturas secundárias regulares nos resultados de simulação. Os padrões obtidos em nossos resultados indicam que estes poderiam ser melhorados se ângulos diedrais associados a regiões inacessíveis do gráfico de Ramachandran fossem melhor penalizados pelo potencial. Um dos caminhos para esse objetivo pode advir de ajustes no efeito do tamanho dos átomos no potencial de repulsão, em particular no impedimento à estabilização de algumas conformações locais.

(4) A simulação de proteínas com cadeiais maiores e o estudo de possíveis modificações necessárias à metodologia para a predição de enterramentos em estruturas compostas por múltiplos domínios.

Apêndice A

Entropia e transinformação

Uma avaliação rigorosa da proposta de que enterramentos possam constituir o intermediário informacional entre sequências e estruturas de proteínas requer que sejamos capazes de quantificar o que entendemos como a informação contida na sequência, a informação contida em enterramentos e o relacionamento entre ambas. O arcabouço que permite que tais análises sejam feitas pode ser fornecido pela Teoria da Informação[71], cujos princípios foram formulados pela primeira vez 1948 por Claude E. Shannon[51] como uma maneira de estudar os limites práticos de operações de transmissão e armazenamento de dados, e hoje encontram aplicações em áreas tão diversas como criptografia, linguística, estatística e neurociência.

De acordo com os preceitos da Teoria da Informação, a quantidade de informação emitida por uma fonte pode ser medida de acordo com a imprevisibilidade dos valores que sua saída pode assumir. Considere, por exemplo, uma mensagem composta por uma sequência de símbolos discretos X, que são lidos sequencialmente por um receptor, sendo que X é capaz de assumir qualquer valor x pertencente a um alfabeto $\mathcal{X}$. Quanto mais uniforme for a distribuição de probabilidades $p(x) = Pr\{X = x\}$, mais imprevisível será o próximo símbolo lido a cada momento, e podemos intuitivamente considerar que uma quantidade maior de informação estará sendo transmitida a cada leitura. Essa intuição é capturada pela definição da **entropia** $H(X)$:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{A.1}$$

Na convenção comumente aplicada em textos de Teoria da Informação, todos os logaritmos são aplicados em base 2, e a entropia $H(X)$ calculada dessa maneira é medida em bits.

A correlação entre os valores assumidos por duas variáveis X e Y pode ser medida pela **entropia**

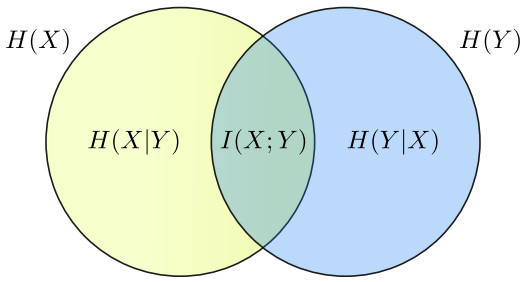


Figura A.1: Relação entre entropia, entropia condicional e transinformação.

condicional $H(Y|X)$:

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \tag{A.2}$$

Por fim, a **transinformação** $I(X;Y)$ mede a quantidade de informação que a variável X contém a respeito da variável Y, e é dada por:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{A.3}$$

Pode ser demonstrado que essas três grandezas estão relacionadas entre si (figura A.1):

$$I(X;Y) = H(X) - H(X|Y) \tag{A.4}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{A.5}$$

$$I(X;Y) = I(Y,X) \tag{A.6}$$

Medidas de transinformação entre sequências de aminoácidos e enterramentos atômicos estimadas de acordo com metodologia desenvolvida por Juliana Rocha[49] são usadas no capítulo 2 deste trabalho como parâmetros de avaliação de qualidade de um método proposto para predição de enterramentos a partir da sequência.

Apêndice B

Proteínas usadas nos bancos de treinamento

As tabelas seguintes listam as proteínas que integram os bancos de treinamento derivados das duas versões do PDBSelect usadas neste trabalho, após as filtragens descritas no início do capítulo 2, na página 15. Cada tabela está organizada em ordem crescente de tamanho da proteína, e as colunas indicam, respectivamente, o código PDB, a cadeia, o número de aminoácidos e a fração de resíduos que adotam conformações em α-hélices e folhas β, de acordo com as definições adotadas pelo programa DSSP[1].

C.1 Proteínas usadas apenas na primeira versão (Novembro de 2009)

PDB	C	Tam	α	β
1bon	A	20	0.2000	0.0000
1l2y	A	20	0.3500	0.0000
2jof	A	20	0.4000	0.0000
2osl	P	20	0.3500	0.0000
1kcn	A	21	0.1905	0.0000
1lv4	A	21	0.0000	0.0000
1pp5	A	21	0.0000	0.1905
1q2j	A	21	0.0000	0.0000
1rpb	A	21	0.0000	0.0000
1ter	A	21	0.3810	0.0000
1zpx	A	21	0.1905	0.0000
1hp9	A	22	0.5000	0.0000
1sut	A	22	0.0000	0.0000
2ajw	A	22	0.2727	0.0000
2bbp	A	22	0.0000	0.0000
1meq	A	23	0.4783	0.0000
1rij	A	23	0.4348	0.0000
1wqe	A	23	0.7391	0.0000
3c0t	B	23	0.6087	0.0000
1oig	A	24	0.0000	0.1667
1ozb	I	24	0.0000	0.0000
2fhw	A	24	0.6667	0.0000
2k98	A	24	0.7500	0.0000
1hvw	A	25	0.0000	0.1600
1wo7	A	25	0.1600	0.0000
1yym	S	25	0.3600	0.2400
1znf	A	25	0.2800	0.0000
2fqc	A	25	0.3600	0.0000
2vqe	U	25	0.2500	0.0000
1cnn	A	26	0.0000	0.0000
1f3k	A	26	0.0000	0.0000
2hgo	A	26	0.0000	0.3462
1ans	A	27	0.0000	0.0000
1fu3	A	27	0.0000	0.1481
1ixt	A	27	0.0000	0.2963
1kcp	A	27	0.0000	0.0000
2nx6	A	27	0.1481	0.0000
1j5m	A	28	0.0000	0.0000
1mea	A	28	0.0000	0.0000
1v5a	A	28	0.0000	0.1429
1wm8	A	28	0.0000	0.0000
2juy	A	28	0.0000	0.1429
2nx7	A	28	0.2857	0.0000
1fry	A	29	0.0000	0.0000
1kal	A	29	0.0000	0.1379
1s8k	A	29	0.2414	0.3103
1zmi	B	29	0.0000	0.4138
2gx1	A	29	0.0000	0.0000
1df6	A	30	0.0000	0.2000
1emx	A	30	0.0000	0.0000
1zfo	A	30	0.0000	0.0000
1ztn	A	30	0.0000	0.0000
1dsv	A	31	0.0000	0.1290
1nj3	A	31	0.0000	0.1290
1pvz	A	31	0.2258	0.1935
1rmk	A	31	0.0000	0.1290
1wt8	A	31	0.0000	0.1290
1g1z	A	32	0.0000	0.1250
1hp3	A	32	0.0000	0.1875
1zmq	A	32	0.0000	0.5938
3i5w	B	32	0.0000	0.5938
1ie6	A	33	0.0000	0.0000
1nix	A	33	0.0000	0.1818
1x5v	A	33	0.0000	0.3333
2knp	A	33	0.0000	0.3333
1c2u	A	34	0.0000	0.0000
1c4e	A	34	0.0000	0.1765
1i26	A	34	0.0000	0.1176
1lu8	A	34	0.0000	0.2353
1s6x	A	34	0.0000	0.2353
2a2v	A	34	0.0000	0.0000
1c49	A	35	0.2000	0.2286
1lmr	A	35	0.0000	0.2286
1sis	A	35	0.2286	0.2857
1tsk	A	35	0.2000	0.2286
1z96	B	35	0.7143	0.0000
2e2s	A	35	0.0000	0.2571
1eit	A	36	0.0000	0.3611
1iyc	A	36	0.0000	0.1667
1q3j	A	36	0.0000	0.2222
2cbh	A	36	0.0000	0.3333
1bgk	A	37	0.4865	0.0000
1hd6	A	37	0.6216	0.0000
1m2s	A	37	0.2973	0.2162
1p8b	A	37	0.0000	0.2973
1bnb	A	38	0.0000	0.3947
1dkc	A	38	0.0000	0.3514
1erp	A	38	0.5789	0.0000
1n5g	A	38	0.4737	0.0000
1xsw	A	38	0.0000	0.0000
3i1p	4	38	0.0000	0.1053
1bbg	A	40	0.2000	0.2000
1erd	A	40	0.4500	0.0000
1mpz	A	41	0.0000	0.0000
1wmt	A	41	0.2500	0.1500
2e2f	A	41	0.2683	0.2195
1wxn	A	42	0.0000	0.5000
2vut	L	42	0.2381	0.0952
1ozz	A	44	0.1818	0.3409
3ddt	C	44	0.1136	0.2045
1ed7	A	45	0.0000	0.2727
1atx	A	46	0.0000	0.3478
1jkz	A	46	0.2391	0.3478
1ti5	A	46	0.2391	0.3261
2cwb	A	46	0.6304	0.0000
1gps	A	47	0.2128	0.3617
1n4n	A	47	0.2128	0.2766
1pxe	A	48	0.0000	0.0000
1ssl	A	48	0.1042	0.1250
3bsu	B	48	0.3333	0.2917
1bk8	A	50	0.1800	0.3600
1ixs	A	50	0.7000	0.0000
1qld	A	50	0.0800	0.3200
2kji	A	50	0.4800	0.0000
3c6w	A	51	0.0784	0.1961
1faq	A	52	0.0000	0.1923
1tot	A	52	0.1346	0.3077
2jrj	A	52	0.1923	0.1923
2nsv	A	52	0.4038	0.0000
1g1x	H	53	0.3800	0.0000
2d49	A	53	0.0000	0.2453
2kak	A	53	0.0755	0.0000
1x3p	A	54	0.0741	0.0000
2odx	A	54	0.0000	0.2778
2vxd	A	54	0.7222	0.0000
1iw4	A	55	0.1273	0.1636
1w0u	A	55	0.6182	0.0000
2cs7	A	55	0.2182	0.2545
2k57	A	55	0.0000	0.4909
2zjr	W	55	0.3273	0.1091
1y62	A	56	0.1429	0.2500
2hdd	B	56	0.6964	0.0000
1dxs	A	57	0.5789	0.0000
2k9h	A	57	0.2807	0.0000
1t50	A	58	0.3448	0.0000
1wj0	A	58	0.0690	0.1724
1y2y	A	58	0.0000	0.0000
2vn5	D	58	0.4107	0.0000
1dwl	A	59	0.0000	0.2373
1fex	A	59	0.5254	0.0000
1k61	B	59	0.6610	0.0000
1y6u	A	59	0.3729	0.2034
1z60	A	59	0.1525	0.1695
2b7e	A	59	0.6780	0.0000
2glo	A	59	0.6102	0.0000
1cxw	A	60	0.0667	0.2333
1hx2	A	60	0.0000	0.2000
1ow5	A	60	0.7333	0.0000
1w70	A	60	0.0000	0.4333
1x41	A	60	0.5333	0.0000
2fpe	H	60	0.0000	0.3833
2j05	B	60	0.0000	0.2373
2k35	A	60	0.2167	0.1667
2vje	C	60	0.1500	0.2000
1y0m	A	61	0.0000	0.4098
2fry	A	61	0.0000	0.2623
2kcr	A	61	0.1803	0.1639
1isu	B	62	0.0968	0.0645
1oyi	A	62	0.5161	0.1129
1trl	A	62	0.6613	0.0000
1sq8	A	64	0.6406	0.0000
2dt6	A	64	0.5938	0.0000
2iyb	G	64	0.0781	0.2188
2juf	A	64	0.1094	0.3438
2k2p	A	64	0.3906	0.2344
3hvz	F	64	0.2031	0.2969
1g25	A	65	0.2000	0.1231
2asc	A	65	0.1538	0.4154
2k85	A	65	0.7231	0.0000
2kfd	A	65	0.5846	0.0000
2sn3	A	65	0.1231	0.1846
3bd1	A	65	0.5846	0.0000
1c9o	B	66	0.0000	0.5455
1u1s	A	66	0.1515	0.4394
1zlj	H	66	0.6364	0.0000
2bzt	A	66	0.5455	0.0000
1u3o	A	67	0.0000	0.2985
2fk4	A	67	0.3433	0.1493
1x2p	A	68	0.0000	0.3088
2k6a	A	68	0.0000	0.3824
1c4q	E	69	0.1739	0.4783
1dpu	A	69	0.4638	0.1304
1ryk	A	69	0.6522	0.0000
1y7y	A	69	0.6377	0.0000
2qif	A	69	0.3188	0.3333
1b22	A	70	0.4429	0.0000
1mms	B	70	0.4714	0.0857
2b97	B	70	0.1429	0.4000
1d3b	C	71	0.0986	0.5352
1h64	1	71	0.0845	0.6056
1hr0	W	71	0.0704	0.1972
2aj0	A	71	0.2958	0.3239
2jve	A	71	0.1690	0.3662
1lea	A	72	0.5417	0.0833
1t0z	A	72	0.2778	0.2222
1yj1	B	72	0.1528	0.2917
1ztg	A	72	0.4583	0.2778
2k42	A	72	0.4861	0.0556
1dtj	D	73	0.4462	0.3077
1vjq	A	73	0.2055	0.2329
2dvw	B	73	0.5890	0.0000
3h2v	H	73	0.2466	0.3562
1hst	A	74	0.4730	0.1351
1kjs	A	74	0.6081	0.0000
1wfw	A	74	0.0000	0.2973
2faz	B	74	0.2162	0.3108
2gle	A	74	0.5676	0.0000
2k52	A	74	0.0000	0.5135
2osq	A	74	0.2432	0.3108
1apj	A	74	0.1486	0.2973
1wve	C	75	0.4533	0.0000
2mst	A	75	0.2400	0.2000
1bg8	A	76	0.5526	0.0000
1c01	A	76	0.0000	0.3684
1h75	A	76	0.3289	0.2237
1pk3	A	76	0.5526	0.0000
3e19	A	76	0.1733	0.4133
3i31	A	76	0.1316	0.3026
1eo0	A	77	0.6883	0.0000
1j7r	A	77	0.5065	0.0000
1pqs	A	77	0.2857	0.2338
1ri9	A	77	0.1169	0.2208
1y3j	A	77	0.2597	0.2468
2b8i	A	77	0.7662	0.0000
2bwf	B	77	0.1558	0.3247
2g9o	A	77	0.2597	0.3117
1uv7	B	78	0.3378	0.3378
2ker	A	78	0.0000	0.3590
2kfk	A	78	0.2308	0.3077
3hgl	A	78	0.5641	0.0000
1krn	A	79	0.0000	0.1266
1x3u	A	79	0.6582	0.0000
2dvh	A	79	0.5570	0.0000
2p5m	B	79	0.4177	0.2785
3bs9	A	79	0.2405	0.4051
3cu4	A	79	0.4430	0.0000
1imt	A	80	0.0000	0.2875
1ttn	A	80	0.1875	0.3125
1wwh	D	80	0.2500	0.3375
2a7y	A	80	0.1875	0.3250
2k8s	B	80	0.2500	0.3625
2rcf	E	80	0.0625	0.3875
2zqe	A	80	0.3625	0.2500
3bqp	A	80	0.6375	0.0000
3dnj	A	80	0.5375	0.2250
1dcj	A	81	0.2840	0.2469
2fa8	C	81	0.3457	0.2963
2hnu	A	81	0.0000	0.3704
1ezg	A	82	0.0000	0.2195
1kx2	A	82	0.3580	0.0000
1puz	A	82	0.6951	0.0000
1wmh	B	82	0.2195	0.3659
1x3x	B	82	0.2927	0.2439
2dob	A	82	0.5976	0.0000
3e8v	A	82	0.0000	0.6341

1pht	A	83	0.1084	0.3855	1cqy	A	99	0.0000	0.5152	1u5f	A	111	0.1176	0.3529	2bkm	A	128	0.6328	0.0000
1wxc	B	83	0.1644	0.3562	1d1n	A	99	0.0000	0.3673	3d1b	A	111	0.7207	0.0000	2k5t	A	128	0.2578	0.3281
2amw	A	83	0.4458	0.0000	1opc	A	99	0.3232	0.2424	1rdo	2	112	0.1786	0.3036	2ohw	B	128	0.3672	0.1797
2ef8	B	83	0.6265	0.0000	1qb5	H	99	0.2525	0.3737	2pi0	B	112	0.2946	0.1786	2rcd	A	128	0.2812	0.2656
2hl7	A	83	0.5000	0.0488	2j9w	B	99	0.7273	0.0000	2z6f	A	112	0.0357	0.5893	3fzw	B	128	0.2812	0.3984
3a39	A	83	0.1205	0.1566	2v3m	A	99	0.0842	0.3474	2zop	F	112	0.7768	0.0000	2d9z	A	129	0.1318	0.2636
1f53	A	84	0.0000	0.3214	2v76	D	99	0.1818	0.4545	3hh1	C	112	0.3462	0.2115	2vq8	A	129	0.1641	0.3281
3kat	A	84	0.5119	0.0000	3cpq	A	99	0.4747	0.1717	1rlg	A	113	0.4602	0.1681	3a6s	A	129	0.1667	0.4524
1grx	A	85	0.4588	0.2118	1hdf	B	100	0.0000	0.3500	1v9y	A	113	0.2115	0.3462	3c9g	B	129	0.2946	0.2403
1hxv	A	85	0.0706	0.3529	1j6q	A	100	0.0000	0.3800	2b7l	D	113	0.3009	0.2301	2ds0	B	130	0.0000	0.4077
1yo3	A	85	0.3882	0.2471	1jbi	A	100	0.0800	0.3000	2atp	D	115	0.0000	0.5398	2hy5	A	130	0.3769	0.1923
2o30	A	85	0.0000	0.5663	2of5	L	100	0.6200	0.0000	2qm2	B	115	0.4174	0.1304	3ct6	A	130	0.4000	0.1692
2ytc	A	85	0.2118	0.2824	2zxj	B	100	0.3300	0.2400	3ic5	A	115	0.3739	0.2087	1lmi	A	131	0.0000	0.4809
3eus	A	85	0.6000	0.0000	1m2d	B	101	0.1980	0.2178	1fo0	B	116	0.0000	0.5625	1uap	A	131	0.1450	0.3282
1jfw	A	86	0.0000	0.0465	1nqj	A	101	0.0000	0.6598	1h8u	B	116	0.2069	0.3276	1ypq	A	131	0.1603	0.3130
1k5o	A	86	0.5059	0.0000	1ztx	E	101	0.0000	0.5347	1jx7	F	116	0.3190	0.1897	2q22	A	131	0.2748	0.1985
1na3	B	86	0.8256	0.0000	2qzi	A	101	0.2376	0.3663	1rlk	A	116	0.4569	0.2500	3fgy	B	131	0.2977	0.4046
1zkh	A	86	0.1744	0.3140	3gxw	A	101	0.2366	0.3011	2bgp	A	116	0.0000	0.5172	3i3f	A	131	0.2519	0.2748
2pls	L	86	0.0930	0.2558	3hpw	B	101	0.1300	0.3800	2k8e	A	116	0.2328	0.3103	2oa2	A	132	0.0530	0.3106
1p4w	A	87	0.6437	0.0000	3iwh	A	101	0.1980	0.1782	3hty	A	116	0.0000	0.5745	2pi2	D	132	0.2097	0.4677
1ci4	B	88	0.6364	0.0000	1bhu	A	102	0.0392	0.1765	1pm4	A	117	0.0000	0.4957	2zp2	B	132	0.1301	0.2764
1sjq	A	88	0.2500	0.2386	1my7	B	102	0.0000	0.4608	1vyn	A	117	0.2222	0.3333	3grd	A	132	0.2273	0.3712
1wmi	C	88	0.2841	0.3409	2gmq	A	102	0.0800	0.3600	2h8c	D	117	0.3009	0.2743	3kht	A	132	0.3333	0.1894
2afd	A	88	0.4432	0.0000	2gol	A	102	0.5980	0.0000	2p8i	D	117	0.3077	0.2051	1efn	D	133	0.3238	0.0952
1dgn	A	89	0.6629	0.0000	2hgl	A	102	0.2255	0.2647	1qfo	A	118	0.0000	0.4103	1o8v	A	133	0.1278	0.5564
2cx6	A	89	0.4157	0.1798	2ida	A	102	0.1275	0.1471	2f9h	B	118	0.1017	0.3390	3dlo	A	133	0.4286	0.2331
2eaq	A	89	0.2022	0.4270	2jua	A	102	0.7647	0.0000	2ijq	B	118	0.6356	0.0339	3i7u	A	133	0.2782	0.4812
2hm2	Q	89	0.6067	0.0000	1pul	A	103	0.6699	0.0000	2re2	A	118	0.1780	0.2542	3ilh	A	133	0.4211	0.2030
3jtn	A	89	0.4157	0.2472	1yv4	A	103	0.1650	0.4078	1m8n	B	119	0.0000	0.6555	1is3	A	134	0.0000	0.6119
2bjd	A	90	0.2444	0.4333	2h3l	A	103	0.0971	0.3689	1t1j	A	119	0.5294	0.1429	1j3w	A	134	0.4478	0.2239
2he4	A	90	0.1444	0.3444	2pv2	A	103	0.3398	0.3204	2alj	A	119	0.1849	0.2437	1w6n	A	134	0.0000	0.6045
3d9t	A	90	0.3778	0.0889	3chb	G	103	0.2136	0.3592	2j48	A	119	0.4034	0.1765	2bnl	B	134	0.7442	0.0000
3fxt	A	90	0.3000	0.2889	3k2y	A	103	0.1262	0.4078	2qjz	B	119	0.4538	0.0000	2ejx	A	134	0.3657	0.4328
1m5z	A	91	0.1538	0.3736	1d4t	A	104	0.1827	0.2596	2rhk	B	119	0.2353	0.4118	3b9c	A	134	0.0000	0.5821
2ife	A	91	0.3077	0.2527	2q30	F	104	0.0000	0.4231	2z3q	A	119	0.5652	0.0522	3dmc	A	134	0.2707	0.4060
1rl1	A	92	0.0000	0.4130	1jhb	A	105	0.4381	0.1714	1gy7	D	120	0.2250	0.4917	1cfe	A	135	0.3037	0.1630
3bn4	F	92	0.3478	0.3587	1jiw	I	105	0.0381	0.5143	2wat	B	120	0.3583	0.2750	1snc	A	135	0.2444	0.2963
1cy5	A	93	0.7391	0.0000	1nrv	A	105	0.1782	0.2574	1oko	D	121	0.0000	0.4463	2h2r	A	135	0.1333	0.3852
1l5p	B	93	0.1290	0.3011	2hsh	A	105	0.4095	0.2762	2jba	B	121	0.3802	0.2479	1thn	C	136	0.3235	0.2353
1pn5	A	93	0.5699	0.0000	2htf	A	105	0.3524	0.1429	2qfd	J	121	0.0000	0.4050	1wyp	A	136	0.4926	0.0000
1t1v	B	93	0.4194	0.1935	2vkc	A	105	0.4571	0.1619	2qgv	J	121	0.2810	0.2066	1z1s	A	136	0.2426	0.3897
1t23	A	93	0.0968	0.3978	1mfw	A	106	0.1132	0.1981	3ff4	A	121	0.2645	0.1983	2f0x	A	136	0.1471	0.3162
1wiu	A	93	0.0000	0.5161	2frg	P	106	0.0000	0.6038	1ijy	B	122	0.3279	0.0328	2gwf	A	136	0.3456	0.1250
3czc	A	93	0.2796	0.2366	3fg7	B	106	0.1981	0.2358	2cua	A	122	0.0000	0.4590	1hqz	9	137	0.3139	0.2336
1fsh	A	94	0.3617	0.1064	3fza	A	106	0.5094	0.1604	2i2q	A	122	0.2705	0.3197	1ktg	A	137	0.2701	0.4380
2ywq	D	94	0.3371	0.3258	1yfn	B	107	0.1215	0.3832	2qv0	B	122	0.4262	0.2049	1nmn	B	137	0.3305	0.2966
3iez	A	94	0.2660	0.4362	2ppp	A	107	0.1121	0.3832	3glv	B	122	0.3525	0.1967	2cdo	B	137	0.0000	0.5109
1im3	P	95	0.0000	0.6316	1hkf	A	108	0.0000	0.5463	1tuj	A	123	0.5854	0.0325	2duw	A	137	0.3139	0.1533
1mn8	A	95	0.5158	0.0000	2nzh	B	108	0.1296	0.3704	2dqa	A	123	0.4715	0.0813	2g7b	A	137	0.1314	0.5620
1qa9	B	95	0.0000	0.4421	2pd2	B	108	0.3611	0.2222	2qgo	A	123	0.2087	0.3826	2j1v	B	137	0.0000	0.4338
1vbv	A	95	0.1282	0.3718	2qzt	A	108	0.3796	0.2870	2quo	A	123	0.0488	0.4797	3f08	A	137	0.2482	0.3796
1w4m	A	95	0.4000	0.1368	3e0z	D	108	0.5278	0.0000	2r4i	A	123	0.2683	0.5203	3grn	B	137	0.2409	0.3431
2ofq	A	95	0.0000	0.4105	3fov	A	108	0.2718	0.4175	2zay	B	123	0.4228	0.1951	1q1u	A	138	0.0000	0.4420
2uzg	A	95	0.2105	0.1684	1t0g	A	109	0.2294	0.1835	3blz	K	123	0.2358	0.4878	1v06	A	138	0.1087	0.3043
2ozi	A	96	0.0000	0.4792	2q3w	A	109	0.0000	0.3761	3cnb	C	123	0.4472	0.2033	2nn8	A	138	0.0000	0.6087
3dl3	A	96	0.0417	0.4167	2r2y	A	109	0.1193	0.4771	3cxg	A	123	0.2683	0.3496	2qs7	A	138	0.2847	0.1898
1ovy	A	97	0.2990	0.1753	1bcp	K	110	0.1091	0.3818	1dy5	A	124	0.1840	0.3280	3d9n	A	138	0.7174	0.0000
2hgf	A	97	0.0928	0.2887	1ejf	A	110	0.0000	0.5091	2p57	A	124	0.5081	0.0726	3esm	A	138	0.0000	0.4745
2w50	B	97	0.6082	0.0000	1exh	A	110	0.0000	0.5455	3eqz	A	124	0.4032	0.2258	2ch4	W	139	0.1007	0.4245
3jst	A	97	0.4227	0.1959	1krs	A	110	0.1727	0.3000	1dbw	B	125	0.4000	0.1920	2z7e	A	139	0.4101	0.1799
1rzx	A	98	0.1531	0.4184	1vc1	A	110	0.3636	0.2727	3crn	B	125	0.5040	0.1040	3e7n	A	140	0.3357	0.2071
1x45	A	98	0.1429	0.2857	2jp0	A	110	0.0545	0.2727	2q3l	B	126	0.3984	0.3333	1gux	B	141	0.5887	0.0284
2aib	B	98	0.5408	0.0612	2k24	A	110	0.3636	0.3091	2vzp	A	126	0.0000	0.5397	1oz9	A	141	0.5106	0.2199
2grg	A	98	0.1837	0.3469	2oiz	D	110	0.0000	0.2636	3hrl	A	126	0.4316	0.1263	2ofc	B	141	0.1064	0.5319
2i45	A	98	0.0722	0.4227	3b5t	E	110	0.0000	0.3636	1lfo	A	127	0.1181	0.5512	2yu6	A	141	0.1986	0.2908
2pmu	A	98	0.3367	0.2245	3gnj	D	110	0.3545	0.1364	1qmp	D	127	0.4127	0.2063	2z5v	A	141	0.2837	0.1277
3fc7	B	98	0.2041	0.3673	1h4x	A	111	0.4144	0.3243	1qu9	A	127	0.2520	0.2205	2jda	B	142	0.0000	0.5493
3fm8	A	98	0.0000	0.4286	1kpf	A	111	0.2432	0.2252	2cjt	D	127	0.0000	0.6378	2zs0	B	142	0.6761	0.0000

3heb	B	142	0.3944	0.1127	1uuy	A	161	0.4348	0.2112	1ydm	C	183	0.2678	0.1311	2gef	B	215	0.1618	0.3039
1fsl	B	143	0.6993	0.0000	1xoy	A	161	0.0497	0.3478	2ejb	A	183	0.4181	0.1977	2q0s	G	215	0.3674	0.1302
1mhq	A	143	0.6853	0.0000	2ift	B	161	0.2733	0.2981	2np5	D	183	0.6855	0.0000	2jig	B	216	0.1269	0.3046
1wka	A	143	0.1748	0.2587	2ph0	A	161	0.2547	0.3106	2v73	A	183	0.0437	0.4481	2o8l	A	216	0.0694	0.4120
1z6b	B	143	0.1791	0.3955	3hwj	A	161	0.0000	0.4800	1uww	A	184	0.0000	0.3443	3cp7	B	216	0.0972	0.3287
2e4m	C	143	0.0000	0.4825	1tp9	A	162	0.2593	0.2963	1zr3	D	184	0.4022	0.2120	2g6y	D	217	0.0372	0.5535
2fy6	A	143	0.2657	0.1958	3bjn	A	162	0.3889	0.1296	2fhp	B	184	0.2626	0.2626	3flp	A	217	0.0369	0.4470
2j1a	A	143	0.0000	0.5070	3k9w	A	163	0.4233	0.1779	3cpm	A	184	0.2717	0.2283	2d3y	A	219	0.3653	0.1233
1kng	A	144	0.2986	0.2292	1b5l	A	164	0.7386	0.0000	1oej	A	187	0.2620	0.2888	2i3d	B	219	0.2557	0.2009
2gl9	D	144	0.2431	0.3264	3e8m	D	164	0.3354	0.1341	1rhy	B	187	0.3536	0.3039	2ij9	B	219	0.4378	0.1475
2z5c	A	144	0.2891	0.3359	1via	B	165	0.5032	0.1419	2o08	B	187	0.5989	0.0000	2oqz	A	220	0.1481	0.2130
1f08	B	145	0.3586	0.2828	3fet	A	165	0.2121	0.3091	1jhs	A	188	0.2021	0.3883	2qmo	A	220	0.3545	0.1136
1jke	A	145	0.2069	0.3862	1oxk	A	166	0.6329	0.0000	1ukf	A	188	0.4149	0.2660	3bo6	A	220	0.3000	0.1636
1o1x	A	145	0.3586	0.1310	1v9t	B	166	0.1205	0.3554	2h29	B	188	0.3936	0.1596	3guy	A	220	0.4244	0.1756
1v7m	V	145	0.4966	0.0552	2hzq	A	166	0.0783	0.4096	2pnm	A	188	0.1538	0.2692	1tqj	F	221	0.3134	0.1982
2z8a	A	145	0.7034	0.0000	2p12	A	166	0.1627	0.3735	3kdh	C	188	0.2553	0.3351	1fdp	C	222	0.0754	0.3668
3ijm	B	145	0.2828	0.2276	1g12	A	167	0.5030	0.0659	1i8a	A	189	0.0000	0.4656	1oq1	A	223	0.0359	0.3677
1id0	A	146	0.2945	0.2671	1vjf	A	167	0.2364	0.2242	2q4m	A	189	0.0898	0.3054	2ps1	B	224	0.3929	0.2545
2prx	A	146	0.1681	0.2941	1yn9	C	167	0.4000	0.1187	2ccm	A	190	0.5947	0.0211	3h7j	B	224	0.0357	0.5670
1aoh	B	147	0.0272	0.5578	2fkb	C	167	0.1677	0.3473	2veb	A	190	0.6000	0.0000	2p7i	A	225	0.3200	0.2311
1jl7	A	147	0.7075	0.0000	3ci6	B	167	0.3373	0.1867	2r2a	B	191	0.0526	0.1947	3c3y	B	225	0.4178	0.1956
2hvw	A	147	0.3265	0.2721	1ohu	B	168	0.5241	0.0000	3bdv	A	191	0.3508	0.0890	1o1z	A	226	0.3451	0.2257
2p5d	A	147	0.1781	0.3151	3eti	F	168	0.3571	0.2262	1fvg	A	192	0.2552	0.1302	1a7t	B	227	0.2731	0.3216
2pdr	D	147	0.2245	0.2585	1x1r	A	169	0.3491	0.2485	2cc0	A	192	0.3958	0.1771	1g8a	A	227	0.2291	0.3128
1b93	A	148	0.3986	0.2230	1lg7	A	170	0.1636	0.3576	3cr3	B	192	0.5104	0.0000	1w2y	B	229	0.6368	0.0189
1dzk	A	148	0.0878	0.5000	1zde	A	170	0.0807	0.4845	3er6	A	192	0.3750	0.2031	2hqz	A	229	0.0176	0.5022
1oj6	C	148	0.6622	0.0000	2bem	C	170	0.0412	0.3765	1gqp	B	194	0.1105	0.4105	1zos	A	230	0.3304	0.3130
2ibg	F	148	0.2015	0.2015	3dmn	A	170	0.2454	0.1963	1nxm	B	194	0.0412	0.3608	1eix	D	231	0.4242	0.1602
1aqz	B	149	0.0972	0.3264	1knq	A	171	0.4737	0.1579	1yll	A	195	0.0000	0.5812	1sgj	C	231	0.4286	0.1688
1gtz	E	149	0.3758	0.1611	1mka	A	171	0.1520	0.3509	2o2n	A	195	0.6483	0.0000	2ac7	A	231	0.3484	0.2896
2ehg	A	149	0.3758	0.2483	1oqv	B	171	0.3450	0.2164	2w1j	A	195	0.1897	0.3026	2f6u	A	231	0.3680	0.1775
2hf6	A	149	0.3960	0.2349	1wba	A	171	0.0000	0.4269	3c7m	B	195	0.4923	0.1128	3e5t	A	231	0.0000	0.5502
2nxp	A	149	0.4899	0.0671	2fco	A	171	0.3419	0.2645	1ioo	B	196	0.3265	0.1888	1o6e	A	232	0.3142	0.2301
3a0y	B	150	0.3733	0.3200	2zmf	B	172	0.3430	0.2267	2fbt	A	196	0.3929	0.1786	2h5y	C	232	0.3707	0.2414
3clj	A	150	0.6871	0.0000	3bx6	A	172	0.1744	0.4186	1cex	A	197	0.3452	0.1472	1zi8	A	233	0.3348	0.2017
1a6m	A	151	0.7417	0.0000	3dsz	A	172	0.0988	0.4244	2gz4	D	197	0.5381	0.0305	2w6r	A	233	0.2667	0.3244
1ouw	B	151	0.0000	0.6333	1sxr	A	173	0.2601	0.1792	3imm	C	197	0.0203	0.3096	3cu2	A	234	0.3205	0.2137
1vkb	A	151	0.0946	0.3446	1v0a	A	173	0.0000	0.3801	1jss	B	199	0.2312	0.4070	1sfl	B	236	0.3632	0.1923
2olp	B	151	0.7152	0.0000	3bo7	C	173	0.1176	0.2647	1t9f	A	199	0.0000	0.4101	2ywv	A	237	0.2996	0.3376
2otm	C	151	0.3046	0.1523	1g5q	L	174	0.3103	0.2299	1yad	D	200	0.4241	0.1885	2zzj	A	238	0.0420	0.5420
3bbb	B	151	0.4040	0.1589	2oh1	B	174	0.2644	0.2586	1ydg	A	201	0.4328	0.1791	3khi	A	238	0.4578	0.0578
3hyq	A	151	0.2053	0.2914	2q4i	A	174	0.0230	0.4655	1z8h	A	202	0.4703	0.1386	2bd0	A	240	0.4167	0.1708
2flh	C	152	0.2368	0.4474	2yzk	C	174	0.3614	0.2892	1ln1	A	203	0.2414	0.4039	1h7e	B	241	0.3154	0.2324
2vof	C	152	0.6291	0.0000	3ck2	A	174	0.1149	0.3161	3i3q	A	203	0.1330	0.2956	1uay	A	241	0.4232	0.1618
3h2d	A	152	0.2829	0.2368	1iaz	B	175	0.1086	0.4686	3e7d	B	204	0.4314	0.0980	2apj	D	241	0.2971	0.1799
1elk	B	153	0.6536	0.0000	2au7	A	175	0.1714	0.3143	1q40	D	205	0.1742	0.2022	3d2l	C	241	0.2282	0.3278
2q4n	A	153	0.0000	0.5686	1m4i	B	176	0.2330	0.3636	2vtw	B	205	0.0195	0.4390	1agj	A	242	0.1364	0.3347
1am7	A	154	0.3791	0.0458	1wlt	B	176	0.0568	0.4545	1b09	C	206	0.0437	0.4078	2bkx	A	242	0.3058	0.1694
2ob0	A	154	0.2403	0.3312	2acf	C	176	0.3523	0.1818	2q52	B	206	0.6505	0.0000	3hft	A	242	0.2810	0.1116
3eyt	B	154	0.2143	0.2208	2gui	A	176	0.3807	0.1875	1h2e	A	207	0.3527	0.1739	1jzt	A	243	0.2922	0.1481
1nbc	B	155	0.0000	0.5419	2iwr	A	176	0.3409	0.2443	1jy5	B	207	0.2500	0.2353	3enz	F	243	0.3128	0.2922
1dyo	B	156	0.0000	0.4167	3e8l	C	176	0.0000	0.3807	3fhg	A	207	0.6812	0.0000	3gne	A	243	0.0000	0.4527
1mxi	A	156	0.2885	0.1603	3ey5	A	176	0.3239	0.1875	1gtv	B	208	0.5048	0.1202	1yb1	B	244	0.4979	0.1564
1p6o	A	156	0.4551	0.1859	1a3c	A	178	0.3155	0.3393	3eja	B	208	0.0192	0.3125	2wje	A	244	0.4180	0.1352
1wwz	A	157	0.3631	0.3503	2dre	A	178	0.0000	0.3064	3f0p	A	208	0.3846	0.2115	1oe4	A	245	0.4531	0.1143
1xww	A	157	0.4013	0.1401	2q3f	B	178	0.3146	0.1910	1q0p	A	209	0.2953	0.1969	1qq7	A	245	0.4653	0.1061
1zng	A	157	0.0828	0.4713	3gxb	A	178	0.3164	0.2203	1qre	A	210	0.1048	0.3190	2dr3	C	245	0.3106	0.2936
2f4w	A	157	0.3784	0.1824	1isp	A	179	0.3073	0.1788	1eyq	A	212	0.3868	0.3066	3f1l	A	245	0.4122	0.1347
3fiq	B	157	0.1083	0.4522	1jwq	A	179	0.4022	0.2179	2cz2	A	212	0.4360	0.0758	1xa1	A	246	0.3347	0.2408
3ily	B	157	0.4296	0.1197	1mjn	A	179	0.3184	0.2235	3fzx	A	212	0.0190	0.5095	1ybf	A	246	0.2822	0.1743
1xju	B	158	0.5256	0.0769	1nsl	D	180	0.2611	0.2944	1q6o	A	213	0.4460	0.2019	2fim	B	246	0.0897	0.3462
1ytl	D	158	0.4051	0.1646	2zu1	A	180	0.0722	0.5500	3bl5	F	213	0.4350	0.1200	3hoj	A	247	0.3806	0.1498
2fia	B	159	0.3459	0.2642	3dou	A	181	0.3239	0.2386	3giu	B	213	0.2582	0.2019	1sg4	A	249	0.4900	0.1606
1svp	A	160	0.0000	0.4375	2gmw	A	182	0.3407	0.1538	1em2	A	214	0.2336	0.3505	2ako	A	250	0.3942	0.1535
1ww7	C	160	0.0250	0.5437	2j2j	A	182	0.0000	0.4066	1h0h	B	214	0.2196	0.2150	2gtr	C	250	0.4920	0.1480
2qgx	D	160	0.3688	0.1688	2rba	B	182	0.3516	0.1429	2d39	A	214	0.1190	0.2810	3ddo	A	250	0.2645	0.2810
3c8o	A	160	0.1688	0.3125	1cr5	C	183	0.1591	0.4148	1x1z	A	215	0.4744	0.1814	1p1x	B	251	0.4741	0.1434
1paq	A	161	0.6522	0.0000	1p3y	1	183	0.3953	0.1628	2dho	A	215	0.2465	0.2977	2dur	A	251	0.0199	0.4104

3ds8	A	253	0.2055	0.1660	2fy7	A	277	0.2342	0.2045	2vpn	B	310	0.4839	0.1935	3fd5	A	369	0.3216	0.2485
1sby	A	254	0.4173	0.1890	3fcx	B	279	0.3406	0.2138	3b50	A	310	0.4452	0.1935	2jhf	B	374	0.2353	0.2299
3bf7	B	254	0.3819	0.1417	1xru	B	281	0.0786	0.2929	2o7r	A	311	0.3377	0.2143	3g62	A	375	0.2533	0.1280
2iip	C	256	0.3651	0.2579	2hte	A	281	0.2747	0.3077	2e6f	B	312	0.3526	0.1987	2qgy	A	376	0.3936	0.1702
3ho6	B	256	0.2480	0.2683	2c1w	C	282	0.3382	0.1985	2pt0	B	313	0.3291	0.1534	1gw1	A	379	0.3138	0.1250
1ro7	A	257	0.2879	0.1323	1wb4	B	283	0.2898	0.1873	2aeb	A	314	0.3567	0.1529	3euo	B	379	0.3641	0.2190
1xg5	D	257	0.4514	0.1440	2r11	A	283	0.3121	0.1489	3k1u	A	314	0.0000	0.3726	2hih	B	386	0.3446	0.1321
2pif	B	257	0.2047	0.1181	1f0n	A	284	0.3768	0.1901	3em5	A	315	0.3238	0.1587	2ib8	A	391	0.3171	0.2199
1oaa	A	259	0.4479	0.1583	1uoc	B	286	0.3984	0.1328	3h63	C	315	0.3238	0.1905	3h12	B	392	0.3852	0.1658
1qgi	A	259	0.5212	0.0541	2afy	X	286	0.1273	0.1745	1ijq	B	316	0.0000	0.4756	2zu9	B	393	0.3573	0.0859
3cpg	A	259	0.3661	0.1299	2gfq	A	288	0.2535	0.2361	2ixo	B	316	0.5304	0.0128	2aq5	A	395	0.0127	0.3899
1j6o	A	260	0.3923	0.1538	3b4u	B	288	0.4549	0.1285	2v84	A	319	0.3386	0.1724	3h2g	A	396	0.3737	0.1907
2a40	E	260	0.2500	0.2577	3isr	D	289	0.1522	0.3183	1ys1	X	320	0.3625	0.1812	2i49	A	398	0.3769	0.1457
1uwc	B	261	0.2874	0.2414	1zby	A	291	0.4605	0.0687	3hbc	A	320	0.1608	0.2154	1ru4	A	400	0.0175	0.2675
2h0q	A	261	0.3065	0.1073	3cij	A	291	0.3265	0.2268	2dvt	D	324	0.4224	0.1211	2bwr	B	401	0.0000	0.3367
1m3u	A	262	0.4160	0.1374	3daq	C	292	0.4555	0.1404	1y2k	B	326	0.6378	0.0000	2vba	D	404	0.3515	0.1931
1xu9	D	262	0.4942	0.1467	3fsu	E	292	0.4893	0.1429	3epw	B	326	0.4202	0.1595	3ivy	A	408	0.4489	0.1047
2cki	B	262	0.1374	0.1450	1fxo	G	293	0.3311	0.2253	3dss	B	327	0.4434	0.0000	1lnl	A	409	0.2787	0.1051
2cmg	B	262	0.2634	0.2977	1w3i	A	293	0.4778	0.1229	2q4h	B	330	0.2500	0.2078	1jnd	A	419	0.2693	0.1845
3k31	A	262	0.4122	0.1450	3bcz	D	293	0.3276	0.1775	3dxt	A	331	0.2659	0.2568	3b9t	A	419	0.0859	0.2888
2g8o	B	263	0.2852	0.2510	2fu9	B	294	0.2368	0.2180	2afb	B	332	0.3242	0.1896	2qfr	A	424	0.1321	0.2759
2qjv	A	263	0.0152	0.3498	1hl2	A	295	0.5017	0.1254	1ri6	A	333	0.0000	0.5495	2nvp	A	430	0.3907	0.0884
1ak0	A	264	0.5758	0.0303	3d7r	A	296	0.3186	0.0814	1y7t	A	333	0.4128	0.1835	1ia6	A	431	0.4753	0.0235
1mhm	A	264	0.2165	0.3661	3g9t	A	296	0.3898	0.1593	2w18	A	333	0.0000	0.5852	2vdu	B	435	0.0505	0.4734
3en0	B	265	0.2491	0.2830	2qpq	B	297	0.3767	0.1918	2hzl	A	334	0.4731	0.1437	1qwo	A	436	0.3862	0.1103
2cb4	C	266	0.1004	0.2124	2ciw	A	298	0.3523	0.0336	2e3b	A	336	0.3601	0.0595	1oyg	A	440	0.0886	0.3886
2dwu	B	266	0.4662	0.1805	2qc5	A	298	0.0000	0.4933	1pby	B	337	0.0000	0.5015	3kez	A	447	0.3647	0.0403
1x8h	A	268	0.2863	0.2467	2o4j	A	299	0.6266	0.0290	2pqm	A	338	0.3550	0.1568	2osx	A	449	0.2784	0.1849
3bmo	B	268	0.4382	0.1474	3gn6	D	300	0.2862	0.1886	1ceo	A	340	0.3934	0.1471	2ij2	B	457	0.4878	0.1109
2zvr	B	269	0.4601	0.1445	1vju	A	302	0.3973	0.2568	1s4n	B	340	0.3452	0.1250	1gkp	D	458	0.3013	0.2096
2q02	C	270	0.3778	0.1444	3cl6	A	302	0.3742	0.1192	2bfd	B	341	0.3483	0.1952	3gue	B	475	0.2435	0.2543
2oo3	A	271	0.3731	0.1940	1uf5	A	303	0.2277	0.2706	1ek6	B	345	0.3971	0.1826	2ipi	D	492	0.2500	0.2154
1jtd	B	273	0.0000	0.4286	1zl0	B	303	0.2937	0.2211	2isn	A	345	0.2186	0.2814	1wdp	A	493	0.3387	0.1014
1wxc	A	273	0.3333	0.0147	3dz1	A	303	0.4053	0.0532	2q49	B	345	0.2725	0.1942	3cgh	A	507	0.3964	0.0197
2ggs	A	273	0.3700	0.1575	3ijd	A	303	0.4757	0.1458	3ib0	A	345	0.2749	0.1842	2hj0	B	511	0.2133	0.1957
2gqp	A	273	0.3223	0.2464	1h1n	B	304	0.3388	0.1776	2a7n	A	353	0.3966	0.1275	1hdh	A	525	0.2838	0.1448
2hvm	A	273	0.2930	0.1905	3hr1	A	305	0.6557	0.0000	1wvg	A	356	0.4561	0.1331	2olr	A	535	0.2542	0.2355
2w1v	A	274	0.2555	0.2774	1rq2	B	306	0.3864	0.2407	2hds	A	358	0.3240	0.2235	1uwk	A	554	0.4170	0.1318
3bxp	A	274	0.3062	0.1512	3hjv	A	306	0.4412	0.1111	1jdw	A	360	0.2333	0.1972	2q9o	B	559	0.0787	0.3506
1w2f	A	275	0.3164	0.2073	3c4u	A	307	0.5256	0.1433	2oui	A	360	0.2556	0.2278	4ubp	C	570	0.2544	0.1754
2bji	A	275	0.3650	0.2409	1uxj	C	308	0.4352	0.2126	2jep	B	362	0.3343	0.1298	2ad6	A	571	0.0403	0.3538
3f1s	B	275	0.0992	0.3015	1u60	A	309	0.4757	0.1197	2d8l	A	363	0.5152	0.0441					

C.2 Proteínas usadas em ambas as versões (Novembro de 2009 / Março de 2012)

PDB	**C**	**Tam**	**α**	**β**															
1t2y	A	25	0.0000	0.0000	3e21	A	40	0.6750	0.0000	1zuu	A	56	0.0000	0.3750	2p5k	A	63	0.5556	0.1587
3gj4	D	25	0.0000	0.1600	3e7r	L	40	0.2500	0.2000	1ft8	E	57	0.4130	0.1304	3eg3	A	63	0.0000	0.4127
3gj3	B	27	0.0000	0.1481	1fd3	A	41	0.1463	0.2683	2vpb	A	57	0.2982	0.1053	1to2	I	64	0.1746	0.2222
1lu0	B	29	0.0000	0.0000	3d36	C	42	0.6190	0.0000	1g6x	A	58	0.1379	0.2414	1ucs	A	64	0.0625	0.1250
3e4h	A	29	0.0000	0.2069	1mkc	A	43	0.0000	0.0930	1k78	I	58	0.5517	0.0000	3fju	B	65	0.1692	0.1846
3g9y	A	29	0.0000	0.1379	1q9b	A	43	0.0930	0.0930	1oot	A	58	0.0000	0.4310	1no1	C	66	0.5152	0.0000
1j5l	A	30	0.1333	0.0000	1bhp	A	45	0.3778	0.1333	1zuy	A	58	0.0000	0.4138	1sj1	B	66	0.2727	0.2424
3e8y	X	30	0.3333	0.3333	1ejg	A	46	0.4130	0.0870	2rh2	A	58	0.0000	0.4561	1tg0	A	66	0.0000	0.4394
1mzw	B	31	0.5806	0.0000	2f3c	I	46	0.2391	0.1957	1oai	A	59	0.6271	0.0000	1ryq	A	67	0.0952	0.3333
1clv	I	32	0.1250	0.1250	1ptq	A	50	0.0800	0.2000	1d0d	A	60	0.1333	0.2000	1tuk	A	67	0.4925	0.0000
2ho2	A	33	0.0000	0.3939	1tc3	C	51	0.5686	0.0000	2v1q	A	60	0.0000	0.4000	1z3e	B	67	0.5373	0.0000
1wy3	A	35	0.6000	0.0000	2gkt	I	51	0.1961	0.1765	1dtd	B	61	0.0984	0.4590	2o9s	A	67	0.0000	0.4030
1rju	V	36	0.0000	0.0000	4sgb	I	51	0.0000	0.1961	1i2t	A	61	0.8197	0.0000	1sf0	A	68	0.0882	0.1471
2jyp	A	36	0.0000	0.0000	1gvd	A	52	0.5577	0.0000	1kq1	W	61	0.1803	0.5410	1vbw	A	68	0.1618	0.2059
2nls	A	36	0.1944	0.3333	1jjd	A	52	0.0962	0.0769	1npi	A	61	0.1311	0.1967	1di2	B	69	0.4754	0.2295
1oc0	B	37	0.1081	0.0000	1yk4	A	52	0.0000	0.2115	2f4m	B	61	0.6557	0.0000	1gcq	C	69	0.0000	0.3623
2eq7	C	37	0.3514	0.0000	2ra2	F	53	0.0000	0.1887	1syx	F	62	0.1452	0.2581	1whz	A	69	0.3768	0.1739
1fre	A	39	0.0000	0.0000	2fdn	A	55	0.0727	0.1818	2iim	A	62	0.0000	0.3871	1kw4	A	70	0.5714	0.0000
1p9g	A	40	0.1000	0.2000	2gnc	B	55	0.0000	0.4727	2nn4	A	62	0.6935	0.0000	2vqc	A	70	0.4429	0.1714
2erl	A	40	0.6750	0.0000	3c4s	B	55	0.0000	0.5818	1cse	I	63	0.1746	0.2698	1c75	A	71	0.4789	0.0000
					3fil	A	55	0.2545	0.4000	1f94	A	63	0.0000	0.5397	2ewt	A	71	0.6197	0.0000

2hip	B	71	0.0000	0.1408	2qg1	A	89	0.1685	0.3371	2w2x	D	107	0.1869	0.4299	3e9v	A	120	0.4667	0.1667
1cc8	A	72	0.2917	0.3611	3fdr	A	89	0.1461	0.3034	1a2p	C	108	0.2037	0.2315	3egn	A	120	0.3167	0.2417
1wm3	A	72	0.1528	0.3333	1yd3	A	90	0.4333	0.2333	1bkr	A	108	0.5370	0.0000	3jql	A	120	0.4202	0.0672
2vc8	A	72	0.0000	0.5139	2e3h	A	90	0.0000	0.4024	1gmx	A	108	0.4167	0.1852	1ufy	A	121	0.2149	0.2810
3d2w	A	72	0.2083	0.3750	3g27	A	90	0.3049	0.0976	1kaf	A	108	0.3333	0.3796	2bh4	X	122	0.3636	0.0413
1hlq	C	74	0.0811	0.0541	1nu4	A	91	0.3516	0.2527	2ptt	B	108	0.0000	0.4444	2gkg	A	122	0.4344	0.1967
1ok0	A	74	0.0000	0.4459	2gpi	A	91	0.3407	0.2527	2pvb	A	108	0.5047	0.0000	1c44	A	123	0.3659	0.2358
2zc2	B	75	0.6267	0.0000	2o71	A	91	0.6484	0.0000	1ycc	A	109	0.3704	0.0000	2o7a	A	123	0.6897	0.0000
3g9o	B	75	0.2933	0.0800	1qys	A	92	0.3261	0.4022	2j6b	A	109	0.2569	0.2569	1gu2	A	124	0.4919	0.0645
2rhf	A	76	0.6842	0.0000	1y8x	B	92	0.2391	0.2500	3b33	A	109	0.2752	0.2294	1ob9	A	124	0.3065	0.3548
3b7h	A	76	0.6053	0.0000	2yxm	A	92	0.0000	0.5326	3h9w	A	109	0.2569	0.3119	3ef4	C	124	0.1210	0.3306
1t8k	A	77	0.5455	0.0000	1bc8	C	93	0.3656	0.1290	1m1f	B	110	0.1792	0.2925	1nwz	A	125	0.2320	0.2800
1vcc	A	77	0.1429	0.3247	1n0q	B	93	0.5269	0.0000	1smo	B	110	0.0364	0.5091	2ehp	B	125	0.3200	0.2080
2i5u	A	77	0.6494	0.0000	1wri	A	93	0.1398	0.3333	2hdv	A	110	0.1635	0.3365	3ehg	A	125	0.3680	0.2640
2k51	A	77	0.0000	0.1818	2fhz	B	93	0.1613	0.3333	2pu9	A	110	0.5273	0.0000	3fsa	A	125	0.1463	0.3171
2q5w	D	77	0.2338	0.2857	2gzv	A	93	0.1630	0.3370	1nyc	A	111	0.0721	0.5405	1fgy	A	126	0.1111	0.3730
3f52	A	77	0.7013	0.0000	3dr0	C	93	0.5269	0.0000	1xg8	A	111	0.3670	0.2110	1ooh	B	126	0.5794	0.0000
2fu2	A	78	0.7949	0.0000	1qzm	A	94	0.5532	0.0638	1qc6	A	112	0.1204	0.4722	2fuf	A	126	0.1984	0.3095
1kp6	A	79	0.2785	0.2532	2fe5	A	94	0.1489	0.3404	2euc	A	112	0.5648	0.0000	2hsb	A	126	0.6984	0.0000
1xmk	A	79	0.5190	0.1266	1c5e	B	95	0.0526	0.3158	2mcm	A	112	0.0000	0.5804	3gwr	A	126	0.1840	0.3440
2fi0	A	79	0.4684	0.0759	1xmt	A	95	0.2526	0.2842	2vh3	A	112	0.5446	0.0000	1g6g	A	127	0.0945	0.3780
3ibw	A	79	0.1899	0.3544	1jo0	B	96	0.4167	0.2083	3d6i	B	112	0.2963	0.2685	2gxf	D	127	0.2203	0.3390
1bdo	A	80	0.0000	0.4500	1lni	A	96	0.1146	0.2396	3f14	A	112	0.1786	0.4911	2jbx	B	127	0.6429	0.0000
1mj4	A	80	0.2125	0.2000	1mwp	A	96	0.1146	0.3438	3ffy	A	112	0.2946	0.1964	2py2	A	127	0.1339	0.3307
1usm	A	80	0.3846	0.2436	2cbp	A	96	0.0625	0.3333	1ifr	A	113	0.0000	0.4690	2wcw	B	127	0.2764	0.3008
1zzk	A	80	0.3625	0.2500	3dml	A	96	0.2632	0.0632	1pz4	A	113	0.2301	0.2832	3g8k	A	127	0.1575	0.3071
2qcp	X	80	0.0000	0.5375	1r1q	A	97	0.1649	0.2062	1r26	A	113	0.3805	0.2478	3i2v	A	127	0.4320	0.1600
1cxy	A	81	0.2346	0.1975	1urr	A	97	0.2577	0.3814	2chh	A	113	0.0000	0.7080	1doi	A	128	0.2031	0.2500
1iqz	A	81	0.1852	0.1728	1w2l	A	97	0.3299	0.0619	2waq	G	113	0.0000	0.4867	1gpq	B	128	0.2656	0.3047
1rwj	A	81	0.1605	0.0988	2cwr	A	97	0.0000	0.6392	3eul	D	113	0.3067	0.2000	1jbe	A	128	0.4252	0.1732
1vqo	S	81	0.3210	0.3086	3cnr	A	97	0.2283	0.3587	3fxh	A	113	0.6018	0.0000	2hew	F	128	0.0469	0.5078
2bps	B	81	0.1111	0.2963	1kwi	A	98	0.1954	0.4023	4fiv	A	113	0.0531	0.4425	3cip	G	128	0.2031	0.3359
2fu4	A	81	0.5062	0.1235	2hlr	A	98	0.0000	0.2857	1o7i	B	114	0.0000	0.4912	3g8z	A	128	0.2283	0.4173
2bl8	B	82	0.6914	0.0000	2p9x	A	98	0.7755	0.0000	1pp7	U	114	0.4474	0.1930	1knm	A	129	0.0000	0.3023
2ffm	A	82	0.2073	0.2683	1bm8	A	99	0.3737	0.2727	1s3s	H	114	0.2315	0.1944	2d48	A	129	0.5814	0.0465
451c	A	82	0.4634	0.0000	1plc	A	99	0.0404	0.3535	1sau	A	114	0.4474	0.0702	2nsz	A	129	0.7442	0.0000
1bb9	A	83	0.1084	0.2892	1r77	A	99	0.0000	0.5051	2cxy	A	114	0.5439	0.0351	2vb1	A	129	0.3101	0.0620
1hq1	A	83	0.7013	0.0000	2ce0	A	99	0.4444	0.0000	3enu	A	114	0.0877	0.3421	2vuv	A	129	0.1783	0.3411
1ugi	B	83	0.2169	0.3735	2hpj	A	99	0.5455	0.0808	3f1p	A	114	0.2105	0.3596	1jf8	A	130	0.4000	0.1692
2a6s	D	83	0.3253	0.2771	2qjl	A	99	0.1515	0.3131	1vqo	O	115	0.4348	0.2348	1tx9	B	130	0.5308	0.0000
2bv2	A	83	0.0000	0.4217	2pd1	D	100	0.3200	0.3200	3gtz	C	115	0.2522	0.3043	1uuz	A	130	0.2326	0.2636
2ckx	A	83	0.6506	0.0000	2vlg	A	100	0.1915	0.3404	3h79	A	115	0.3478	0.2000	3nul	A	130	0.2769	0.2923
2d3d	A	83	0.5783	0.0000	1y0h	A	101	0.3168	0.2673	1fc3	C	116	0.6724	0.0000	1ifc	A	131	0.1145	0.5878
2gzg	A	83	0.5422	0.0000	2hj3	A	101	0.6931	0.0000	2fb6	A	116	0.4224	0.2069	1p0z	J	131	0.3206	0.2824
2zrr	A	83	0.7108	0.0000	3d33	A	101	0.0000	0.4947	2wb6	A	116	0.2807	0.3421	1t3y	A	131	0.3282	0.3206
3ge3	C	83	0.1687	0.3133	3gwl	B	102	0.7059	0.0000	3bj9	1	116	0.0000	0.5086	1tuh	A	131	0.3511	0.4275
2ywk	A	84	0.2143	0.3452	1xer	A	103	0.1359	0.2524	3cp5	A	116	0.5948	0.0000	2qsj	A	131	0.4344	0.1639
3ihs	B	84	0.2857	0.2143	2j73	A	103	0.0000	0.5534	3h9x	A	116	0.2500	0.2069	2v4x	A	131	0.5420	0.0458
1opd	A	85	0.3765	0.2706	1eaz	A	104	0.2039	0.3786	2cyj	A	117	0.2479	0.3333	3cxk	B	131	0.1221	0.2901
1wlz	A	85	0.6000	0.0000	1i0v	A	104	0.1635	0.2692	1f7l	A	118	0.3729	0.3305	3f8x	A	131	0.3231	0.3000
2p13	B	85	0.2024	0.1667	1jrm	A	104	0.2212	0.2115	1zma	A	118	0.4237	0.1949	3fk8	A	131	0.4046	0.1603
2qsb	A	85	0.6471	0.0000	2bt6	B	104	0.1731	0.2308	2b4a	A	118	0.3162	0.2222	1od3	A	132	0.0000	0.5496
2bkf	A	86	0.2907	0.3721	1lkk	A	105	0.1714	0.2571	2djf	A	118	0.0339	0.6186	1vqo	K	132	0.0758	0.3409
2zxy	A	86	0.4419	0.0000	1m9z	A	105	0.0000	0.4190	2mhr	A	118	0.6441	0.0000	2pie	A	132	0.0303	0.3485
1aba	A	87	0.3448	0.1839	2ov0	A	105	0.0000	0.4286	3a07	B	118	0.0000	0.3814	3ebt	A	132	0.2000	0.4154
1nh9	A	87	0.3457	0.3704	2p04	B	105	0.1619	0.3238	3fka	A	118	0.2288	0.3983	1gp0	A	133	0.0000	0.3759
1wmg	F	87	0.4828	0.0000	1a1x	A	106	0.0377	0.4811	1mai	A	119	0.2437	0.3109	2c60	A	133	0.2907	0.2674
1x6i	B	87	0.5862	0.0000	1btn	A	106	0.1792	0.3679	1srr	A	119	0.4622	0.2017	2hnf	A	133	0.0000	0.4504
2ip6	A	87	0.7011	0.0000	1ew4	A	106	0.3113	0.3113	1wou	A	119	0.3529	0.2269	3d5p	A	133	0.3233	0.2481
2nqw	A	87	0.2414	0.2299	1nnx	A	106	0.0426	0.5000	2g3r	A	119	0.0756	0.5210	3dmo	D	133	0.2857	0.2256
3g1j	B	87	0.0000	0.3793	1xlq	C	106	0.1226	0.1981	2p2e	A	119	0.0924	0.3109	3ga3	A	133	0.0526	0.3985
1gxu	A	88	0.2386	0.4091	2fhz	A	106	0.3491	0.1887	3fz4	A	119	0.3529	0.1429	3i7m	A	133	0.2632	0.1729
2bt9	C	88	0.0000	0.6818	2vq4	A	106	0.0000	0.5283	3h3h	A	119	0.3193	0.2521	1dqg	A	134	0.0000	0.3134
2cov	I	88	0.0000	0.6023	3dqy	A	106	0.0000	0.3491	1eaj	B	120	0.0000	0.4917	1lu4	A	134	0.3134	0.2313
1ay7	B	89	0.4719	0.1798	7fd1	A	106	0.1698	0.1132	1nez	H	120	0.0000	0.4917	1mc2	A	134	0.4016	0.0656
1h8e	H	89	0.0000	0.6180	1qad	A	107	0.1714	0.3048	1w0n	A	120	0.0000	0.6333	2vlq	B	134	0.3209	0.1045
1wv9	A	89	0.2614	0.1136	2i4a	A	107	0.3364	0.2804	3b7c	A	120	0.2250	0.2583	3f1t	C	134	0.1642	0.2910
2q3g	A	89	0.1461	0.4382	2i5f	A	107	0.1500	0.4300	3d1p	A	120	0.2833	0.1417	1e29	A	135	0.3630	0.0444

2fvv	A	135	0.2444	0.3852	2yv0	X	155	0.3548	0.2839	2dy0	B	181	0.3204	0.3425	3ein	A	207	0.5604	0.0773
3dm8	A	135	0.3259	0.4519	3a1b	A	155	0.2953	0.1275	2j8k	A	181	0.0497	0.0000	1tu7	A	208	0.5240	0.0913
3h16	D	135	0.3435	0.2061	1n1f	A	156	0.6364	0.0000	1mvl	A	182	0.4023	0.1667	1wzd	A	209	0.7273	0.0000
2okf	A	136	0.3769	0.2692	1nyk	A	156	0.0256	0.2756	2o9a	A	182	0.4011	0.2253	2ah5	A	209	0.3876	0.1627
1ccw	A	137	0.4161	0.2190	2ewr	A	156	0.3782	0.2692	3h8t	A	182	0.0220	0.5110	3boe	A	209	0.3158	0.1818
1tvg	A	137	0.0000	0.5294	2vxt	I	156	0.0385	0.5256	1pmh	X	183	0.0328	0.4863	1es9	A	212	0.4245	0.1132
1vkk	A	137	0.3212	0.2847	2fg1	A	157	0.2675	0.1465	3hp4	A	183	0.4590	0.1366	2bkr	A	212	0.4265	0.1659
1vh5	A	138	0.2029	0.3913	2nut	C	157	0.3285	0.2555	3igr	B	183	0.2404	0.3224	2hal	A	212	0.0519	0.4764
2r6q	A	138	0.0000	0.5797	2qim	A	157	0.2548	0.4522	1x8q	A	184	0.1087	0.4402	2vcl	A	212	0.1787	0.2705
2vac	A	139	0.3657	0.3284	2vy8	A	157	0.4052	0.1373	3a35	B	184	0.0489	0.4402	1lbu	A	213	0.3146	0.0798
2yvq	A	139	0.4148	0.1407	3gkm	A	157	0.2994	0.2675	1t0i	B	185	0.3427	0.1910	1sfs	A	213	0.3099	0.2066
3d3b	A	139	0.7122	0.0000	1y93	A	158	0.2278	0.1709	2vrn	A	185	0.3027	0.2324	2iuw	A	213	0.1748	0.3252
1j7d	A	140	0.2429	0.1643	2ywn	A	158	0.1656	0.2649	2zk9	X	185	0.2270	0.2757	3bhd	A	213	0.2796	0.3412
1jmv	A	140	0.4000	0.2500	3ip0	A	158	0.2848	0.2532	3gxr	A	185	0.5081	0.0324	3dtn	B	213	0.4178	0.2441
3eur	A	140	0.2643	0.2500	2hwy	B	159	0.5000	0.1121	1kmv	A	186	0.1838	0.3027	3eng	A	213	0.1127	0.2207
1b0b	A	141	0.7163	0.0000	2i6c	A	159	0.2848	0.2532	2p14	A	186	0.4462	0.1720	3gnz	P	213	0.1179	0.3302
1cg5	B	141	0.6950	0.0000	3ct5	A	159	0.5283	0.0000	3hyn	A	186	0.2097	0.1452	1u0a	D	214	0.0187	0.4907
1rg8	A	141	0.0000	0.3688	3f0d	D	159	0.2975	0.2975	1ysq	A	187	0.4286	0.2143	3bwy	A	214	0.4159	0.2290
2nwf	A	141	0.0567	0.3191	1r8s	A	160	0.3000	0.2188	2rk3	A	187	0.3763	0.2097	1jg1	A	215	0.2977	0.2512
2w72	A	141	0.6879	0.0000	2bzv	A	160	0.0000	0.3562	3eln	A	187	0.1337	0.3583	2uur	A	215	0.0616	0.4171
2wcj	A	141	0.5816	0.0000	2arc	A	161	0.2174	0.3416	1e5k	A	188	0.3032	0.2128	1k4i	A	216	0.3426	0.2083
1avg	I	142	0.0563	0.4085	2i6j	A	161	0.5404	0.1242	1wl8	A	188	0.2353	0.3048	2wag	A	217	0.2581	0.2074
2j3t	C	142	0.3262	0.1986	2ooi	A	161	0.1709	0.4367	2os0	A	188	0.2167	0.2500	1w66	A	218	0.2523	0.2798
2o1c	D	142	0.2113	0.3662	3d3s	D	161	0.2245	0.2517	2rci	A	188	0.2979	0.3617	3dlc	A	219	0.3744	0.2283
2zou	B	142	0.0000	0.3901	3fnc	B	161	0.3292	0.3354	3dcm	X	188	0.3723	0.1596	2pof	B	220	0.1689	0.1598
3bln	A	142	0.2324	0.3592	3gbw	A	161	0.0000	0.4688	1s99	A	189	0.2460	0.1818	1euv	A	221	0.3801	0.1719
3cg6	B	142	0.4923	0.1692	1lke	A	162	0.1266	0.4873	2dfb	A	189	0.0529	0.6138	2a6z	A	222	0.0180	0.4550
2hl0	A	143	0.2517	0.3776	2p8g	A	162	0.0926	0.3210	2nw0	A	189	0.2646	0.2857	2dgd	D	222	0.3874	0.1622
3f9s	A	143	0.2867	0.3077	1z4r	A	163	0.3067	0.2883	1m70	A	190	0.4474	0.0000	2nlr	A	222	0.0450	0.4910
1o8x	A	144	0.2797	0.2308	2gkp	A	163	0.3519	0.2840	1ucd	A	190	0.2842	0.2263	2oc5	A	222	0.7568	0.0000
2vt8	A	144	0.2778	0.2361	2ims	A	163	0.6135	0.0000	2iqy	A	190	0.0789	0.2526	1gxy	B	223	0.3274	0.1794
1ilr	1	145	0.0000	0.5103	2v1m	A	163	0.2761	0.1902	2jgb	A	190	0.2553	0.3032	1uai	A	223	0.0000	0.5202
1nww	A	145	0.2690	0.3655	1tt8	A	164	0.1280	0.3780	1qv1	A	191	0.5798	0.0213	3fci	A	223	0.3229	0.1076
1r9w	A	145	0.3741	0.3165	2aen	A	164	0.0610	0.5610	1vhu	A	191	0.3560	0.2094	1byi	A	224	0.3527	0.2009
2qeb	A	145	0.6345	0.0000	2q7b	A	164	0.2439	0.2805	2gzq	A	191	0.0995	0.2251	1pq5	A	224	0.0848	0.3304
1dm1	A	146	0.7397	0.0000	1wko	A	165	0.0970	0.3212	2wl1	A	191	0.0262	0.3665	1g61	B	225	0.2933	0.2444
1it2	B	146	0.6301	0.0000	1io0	A	166	0.4578	0.0904	3hmz	A	191	0.1466	0.3403	2ex4	B	226	0.3423	0.2703
1zce	A	146	0.1301	0.2945	2qf4	B	167	0.0000	0.4970	1ui0	A	192	0.4219	0.1094	2tps	A	226	0.4248	0.1549
1f4p	A	147	0.3061	0.2517	3bnw	B	167	0.1887	0.3774	3d06	A	192	0.0773	0.3591	2cws	A	227	0.0396	0.4802
1h97	A	147	0.6599	0.0000	3eyi	A	167	0.4531	0.1406	2qml	A	193	0.2021	0.1969	1nfp	A	228	0.4561	0.1886
2zex	A	147	0.0000	0.4626	3d3m	A	168	0.7284	0.0000	1i9s	A	194	0.3263	0.1000	1fj2	A	230	0.2926	0.1965
1k66	B	149	0.3691	0.1409	1bwn	B	169	0.1173	0.3333	1qhv	A	195	0.0000	0.3385	2gb4	B	230	0.2696	0.2696
1smb	A	149	0.4027	0.1477	1dvk	B	169	0.6358	0.0000	3c9q	A	195	0.2051	0.1846	3h3l	A	232	0.0345	0.2543
3edo	B	149	0.2685	0.2282	1rcf	A	169	0.2722	0.2189	2osa	A	196	0.6582	0.0000	3igs	B	232	0.3922	0.1078
3f2z	A	149	0.0000	0.3893	1ow1	A	170	0.3512	0.3333	1w0h	A	199	0.2714	0.1558	1k7c	A	233	0.3305	0.1416
2fom	B	150	0.0267	0.5000	1qft	B	170	0.1183	0.4142	1yd7	A	199	0.2765	0.2000	2f5t	X	233	0.2275	0.3991
2vvw	A	150	0.6800	0.0000	1r2q	A	170	0.3059	0.2824	2pvq	A	201	0.4826	0.0846	2jen	A	233	0.0429	0.5451
1f1e	A	151	0.5828	0.0265	3fj2	A	170	0.2000	0.3471	2vpt	A	201	0.4527	0.1393	2jfr	A	234	0.2991	0.2607
1v4p	A	151	0.2980	0.3179	1y43	B	171	0.0000	0.5556	2w15	A	201	0.3383	0.1642	2va1	F	234	0.3744	0.2026
2nl9	A	151	0.6643	0.0000	2nn5	A	171	0.3450	0.3567	1jm1	A	202	0.0792	0.2822	1tl2	A	235	0.0851	0.4085
2nrr	A	151	0.3662	0.2535	2wfi	A	172	0.1163	0.2907	1yuz	B	202	0.5446	0.0396	1qvz	A	236	0.3051	0.1059
2ob5	A	151	0.2384	0.1126	1cv8	A	173	0.2717	0.2890	2opi	A	202	0.3465	0.1634	1nls	A	237	0.0000	0.4599
1bvy	F	152	0.3421	0.1974	2pc1	A	173	0.2254	0.2486	2imf	A	203	0.5665	0.0985	2hyk	A	237	0.0000	0.4346
1byr	A	152	0.3750	0.2763	1oh4	A	174	0.0230	0.5920	1hdo	A	205	0.3024	0.2098	3cjw	A	237	0.6364	0.0000
1x3k	A	152	0.6974	0.0000	1rtt	A	174	0.3736	0.1782	1ix9	A	205	0.5220	0.1122	1lst	A	238	0.3403	0.1933
1gny	A	153	0.0392	0.5098	1y63	A	174	0.4852	0.1538	2c71	A	205	0.4098	0.1268	1p5z	B	241	0.5130	0.1130
1gwm	A	153	0.0000	0.4837	2scp	B	174	0.5690	0.0460	2czq	A	205	0.3268	0.1512	1ql0	B	241	0.1660	0.1950
1mfm	A	153	0.0261	0.3791	1kt7	A	175	0.0971	0.4571	2oxc	A	205	0.4098	0.1415	1tqh	A	242	0.4504	0.1405
2bk9	A	153	0.6732	0.0000	1z6m	A	175	0.3886	0.1029	2pu3	A	206	0.4638	0.1353	1z9t	A	242	0.1992	0.2282
2imj	D	153	0.2222	0.3203	2a2k	A	175	0.2907	0.1570	2vyo	A	206	0.3592	0.1602	1ymt	A	243	0.6949	0.0254
3eye	A	153	0.3464	0.2353	1d2s	A	176	0.0000	0.4971	3c6a	A	206	0.3367	0.1508	3cql	B	243	0.4033	0.0000
3gxh	B	153	0.3684	0.1250	1d4o	A	177	0.3277	0.1695	3fde	B	206	0.1408	0.2573	3i94	A	243	0.3539	0.2346
1j98	A	154	0.3247	0.2273	1eb6	A	177	0.5141	0.0791	1g66	A	207	0.3382	0.1449	1otk	B	244	0.6516	0.0000
2aqm	A	154	0.0000	0.4221	1kmq	A	177	0.3333	0.2542	1nf8	A	207	0.3575	0.1739	2h8g	B	244	0.3443	0.2541
2cak	A	154	0.0519	0.3831	2znr	A	178	0.2584	0.3483	1s9u	A	207	0.5495	0.0000	2w3z	A	244	0.3598	0.1506
3eqe	B	154	0.1299	0.2792	2bue	A	179	0.3017	0.2905	2vu6	A	207	0.1063	0.3671	3ega	A	244	0.0000	0.3063
2b18	A	155	0.3935	0.2194	1wc2	A	180	0.0833	0.2778	3bdi	A	207	0.2271	0.2029	3h11	A	244	0.2676	0.2535
2d37	A	155	0.1355	0.4774	1i4u	A	181	0.1602	0.4309	3dxy	A	207	0.3188	0.2029	1ah7	A	245	0.5796	0.0000

1k55	A	245	0.3143	0.2163	1ga8	A	281	0.2780	0.1516	1tca	A	317	0.3281	0.1230	1uas	A	362	0.2652	0.2348
3e0x	A	245	0.2776	0.1633	3irs	A	281	0.4128	0.1139	3ga7	A	318	0.2922	0.1396	1kwf	A	363	0.4573	0.0220
1o9g	A	249	0.3855	0.1406	1lyv	A	283	0.3286	0.1943	2zpu	A	319	0.4514	0.1505	1vyr	A	363	0.3030	0.1708
2vxn	A	249	0.3936	0.1566	1e9g	A	284	0.1444	0.2817	3bb7	A	321	0.2349	0.1841	1kq3	A	364	0.4863	0.1209
1zjz	A	251	0.4303	0.1474	1qtw	A	285	0.3860	0.1298	3ing	A	321	0.3281	0.1969	1oc7	A	364	0.3187	0.1071
2okg	B	251	0.3160	0.1720	1n62	F	286	0.3392	0.2587	1r6d	A	322	0.3882	0.1770	1r8g	A	368	0.3541	0.1076
1xqo	A	253	0.6403	0.0000	1oih	D	287	0.2105	0.2874	1bxo	A	323	0.1022	0.4489	1ga6	A	369	0.2873	0.2439
2bsy	A	253	0.0200	0.4760	1zgk	A	288	0.0000	0.4306	2ddx	A	324	0.3025	0.1975	3d59	A	372	0.3145	0.2016
1xm8	B	254	0.3465	0.2441	2cis	A	288	0.0451	0.4549	2z1e	A	324	0.2809	0.2776	1so7	A	374	0.0385	0.5110
3dha	A	254	0.1890	0.2874	2zpt	X	288	0.4236	0.1042	2wao	A	325	0.2369	0.3077	2qpx	A	376	0.4388	0.0878
1qgq	A	255	0.3167	0.2583	1rtq	A	291	0.3574	0.1924	1nuy	A	328	0.2844	0.2477	1qop	B	390	0.4282	0.1949
2yvt	A	256	0.2812	0.2656	1rhs	A	293	0.2935	0.1229	1odm	A	329	0.2705	0.1945	1kjq	B	391	0.3023	0.2894
3c70	A	256	0.3516	0.1797	1zwx	A	293	0.1972	0.2872	2ob3	B	330	0.4407	0.1064	1i24	A	393	0.3760	0.1790
1m2x	A	257	0.2511	0.3242	1h72	C	296	0.3615	0.2432	2qx3	B	330	0.0545	0.3576	1jfb	A	399	0.4687	0.0952
2qdx	A	257	0.2101	0.3113	1sdd	A	296	0.0332	0.2399	1gxm	B	332	0.3614	0.1054	1p1m	A	404	0.3193	0.2252
1lug	A	259	0.0659	0.2984	2whl	A	296	0.3492	0.1627	3imh	A	332	0.0331	0.4277	1xwt	A	404	0.4566	0.0918
2agk	A	260	0.3149	0.2468	1mj5	A	298	0.3266	0.1818	1k5c	A	333	0.0000	0.5315	2vfr	A	418	0.2081	0.2273
1ep3	B	261	0.1571	0.3525	1u4g	A	298	0.3691	0.1141	2p2s	A	334	0.2973	0.1892	1itx	A	419	0.2888	0.2053
2rb5	A	261	0.3180	0.1954	2a0m	A	298	0.3591	0.1678	2p9w	A	334	0.0329	0.4401	3h9m	A	425	0.2494	0.2843
2z4u	A	261	0.3716	0.2107	2w39	A	298	0.1544	0.3691	3das	A	334	0.0120	0.3413	1ug6	A	426	0.3920	0.1596
1arb	A	263	0.0570	0.3156	2ehz	A	299	0.1773	0.3278	3cbw	A	336	0.3006	0.1786	2v3i	A	433	0.1109	0.3418
2ggc	A	263	0.2433	0.2548	3cny	A	299	0.4080	0.1405	2qe8	B	337	0.0417	0.3571	2imz	B	438	0.0429	0.5714
1m40	A	265	0.3840	0.1711	8a3h	A	300	0.3567	0.1767	2z72	A	338	0.2633	0.2189	2w6p	B	445	0.3222	0.2655
1omz	A	265	0.2953	0.2362	1esc	A	302	0.3113	0.1225	1yfq	A	342	0.0234	0.4795	1z4v	A	448	0.0584	0.4449
1qop	A	267	0.4850	0.1391	1i1w	A	302	0.4106	0.1689	1qnr	A	344	0.3488	0.1657	2bf6	A	449	0.0179	0.4732
2p0l	A	269	0.2921	0.2172	3i10	A	302	0.3381	0.1295	1pwg	A	345	0.3304	0.1739	3eqa	A	458	0.4476	0.0590
1o4y	A	270	0.0630	0.4259	2cnq	A	304	0.3079	0.2318	1of8	B	347	0.3652	0.1623	1hx0	A	495	0.1818	0.1960
1qwy	A	270	0.0851	0.3362	1pa2	A	306	0.4314	0.0196	2abs	A	350	0.3519	0.2317	1n4w	A	499	0.2450	0.2008
3ch0	A	272	0.2509	0.1476	1m4l	A	307	0.3681	0.1629	3clm	A	352	0.5227	0.0966	1r7a	B	504	0.3274	0.1806
1gci	A	275	0.3048	0.1859	3fwk	A	308	0.3459	0.0788	1lc5	A	355	0.3577	0.1577	1wui	L	534	0.4064	0.1442
1wma	A	275	0.4691	0.1527	3cpx	A	309	0.2395	0.2168	1luc	A	355	0.4404	0.1651	2d1s	A	539	0.2857	0.2189
2ci1	A	275	0.2364	0.2473	1dcs	A	311	0.3121	0.2482	1m15	A	356	0.3764	0.1461	1gq1	B	559	0.1270	0.3435
1qqf	A	276	0.6123	0.0000	2b69	A	312	0.3055	0.1736	1qcx	A	359	0.0836	0.3565	3bb0	A	576	0.3906	0.0469
1ltz	A	277	0.4182	0.0982	1us0	A	314	0.3482	0.1182	2aml	B	360	0.4039	0.1421	2zux	B	582	0.0155	0.3505
2vha	B	277	0.3551	0.2319	3cu9	A	314	0.0127	0.4586	2oiz	A	360	0.0139	0.4722	1su7	A	633	0.5024	0.1074
1b8o	A	280	0.3036	0.2357	2w0i	A	315	0.2910	0.3134	3iar	A	360	0.4694	0.1222	1k3i	A	652	0.0000	0.3871
2imq	X	280	0.1357	0.3179	1hnj	A	317	0.3375	0.2429	1pe9	A	361	0.0803	0.3546					

C.3 Proteínas usadas apenas na segunda versão (Março de 2012)

PDB	C	Tam	α	β															
3uz6	X	25	0.2800	0.0000	3ddt	A	48	0.1250	0.1875	1k61	D	58	0.6552	0.0000	3bd1	C	64	0.6562	0.0000
2plx	B	26	0.6923	0.0000	3ngg	B	48	0.0000	0.1667	1kth	A	58	0.1379	0.2414	3hvz	B	64	0.2031	0.3125
1zmi	A	29	0.0000	0.4138	3qdr	B	49	0.0000	0.4286	2ovg	A	58	0.4483	0.2414	1icf	I	65	0.0923	0.1231
1zmm	A	31	0.0000	0.6129	3a02	A	51	0.7647	0.0000	3dfx	B	58	0.1724	0.0690	2yeo	A	65	0.1538	0.4154
2x9s	4	31	0.0000	0.3333	1rh6	B	52	0.3077	0.2885	2fma	A	59	0.2373	0.4915	2z30	B	65	0.0769	0.3692
3a9j	C	32	0.0000	0.1250	2b9s	B	52	0.4231	0.0000	2vn5	B	59	0.4386	0.0000	3ipf	B	65	0.0000	0.6923
3i5w	A	32	0.0000	0.5938	2uuy	B	52	0.0962	0.2692	3cjs	A	59	0.1017	0.3559	3oss	C	65	0.0000	0.5385
3qte	A	32	0.0000	0.5938	3c6w	C	52	0.0769	0.1154	3s6n	G	59	0.0000	0.5714	3zr8	X	65	0.6615	0.0000
3mjh	B	34	0.3235	0.1176	3o7a	A	52	0.0769	0.1154	1fas	A	61	0.0000	0.3934	1c9o	A	66	0.0000	0.6061
1omb	A	35	0.0000	0.1176	1g1x	C	53	0.4375	0.0000	1m8a	B	61	0.1639	0.2787	1y7y	B	66	0.6515	0.0000
1cbh	A	36	0.0000	0.3333	3ph0	C	53	0.7736	0.0000	1o7z	A	61	0.1500	0.2833	2hin	A	66	0.6364	0.0000
4cpa	I	37	0.0000	0.1622	3qmd	A	53	0.1698	0.0000	2gom	A	61	0.8033	0.0000	2xi8	A	66	0.5606	0.0909
1z96	A	38	0.7368	0.0000	3qwo	P	53	0.7736	0.0000	2yw8	A	61	0.0984	0.1639	2y3n	B	66	0.3115	0.0000
3odv	A	38	0.1579	0.3684	1igq	A	54	0.0000	0.5185	3kup	C	61	0.3000	0.3167	3dso	A	66	0.0000	0.4242
3r8s	4	38	0.0000	0.4211	2cs7	C	54	0.2222	0.2593	1isu	A	62	0.0968	0.0645	3fmy	A	66	0.6061	0.0000
3pis	A	39	0.1538	0.1026	3k2a	A	54	0.6852	0.0000	1v6p	A	62	0.0000	0.4355	3hfo	C	66	0.1667	0.4697
1fs1	A	41	0.4878	0.0000	1pfb	A	55	0.1273	0.4000	1vqo	G	62	0.8000	0.0000	3sb2	D	66	0.1538	0.4615
1px9	A	42	0.2381	0.2143	1w0u	B	55	0.6182	0.0000	1z7k	B	62	0.1774	0.2258	2y9u	A	67	0.5970	0.0000
3b0f	A	43	0.6279	0.0000	2hdd	A	55	0.7273	0.0000	2fpe	A	62	0.0000	0.3710	3i4o	B	67	0.0746	0.5522
2yvr	B	45	0.1111	0.2000	3alr	D	55	0.1818	0.0000	2oxl	A	62	0.7258	0.0000	3omt	B	67	0.5075	0.0000
1e8p	A	46	0.0000	0.3913	3pmt	A	55	0.0000	0.5455	2xmj	A	63	0.3492	0.2063	1x2i	A	68	0.6176	0.0000
2oo9	A	46	0.5435	0.0000	1y62	D	56	0.1429	0.2500	1uoy	A	64	0.0000	0.3594	2qif	B	68	0.3235	0.3382
2j9u	B	47	0.0851	0.1702	2cmp	A	56	0.7018	0.0000	2iyb	E	64	0.0781	0.2812	3cjs	C	68	0.2647	0.3529
3psm	B	47	0.2128	0.3191	3u23	A	56	0.0000	0.5000	2j05	A	64	0.0000	0.1875	1c4q	B	69	0.1739	0.4783
3bsu	A	48	0.3542	0.2917	2qtx	L	57	0.0000	0.5614	2xeu	A	64	0.2500	0.1562	2iy2	A	69	0.2174	0.3478
					1jo8	A	58	0.0000	0.3966	2xiw	A	64	0.1562	0.5469	3u5g	Z	69	0.3768	0.1014

1sfu	B	70	0.4714	0.1429	3i5r	A	81	0.1358	0.3951	1b34	B	93	0.1467	0.4933	1l8r	A	101	0.3960	0.1881
2b97	A	70	0.1429	0.4000	3pc7	A	81	0.2469	0.1235	1l5p	C	93	0.1290	0.3011	1m2d	A	101	0.2475	0.2178
2xmw	A	70	0.3382	0.3824	3tr3	B	81	0.3333	0.2593	1q8b	A	93	0.2473	0.2258	1nlq	D	101	0.0000	0.6000
3d2q	D	70	0.0571	0.0857	1ezg	B	82	0.0000	0.2195	1t1v	A	93	0.4194	0.1935	2hwv	A	101	0.3465	0.2178
3nar	A	70	0.7000	0.0000	3q6c	A	82	0.0882	0.4265	2rb8	A	93	0.0000	0.5161	2hy5	C	101	0.3069	0.2178
3ny3	A	70	0.0714	0.1429	1i71	A	83	0.0000	0.1205	2w0p	A	93	0.0000	0.4946	2j9w	A	101	0.7327	0.0000
3p1x	A	70	0.2857	0.1429	1luz	B	83	0.1084	0.4217	2xs2	A	93	0.1364	0.4318	2ptv	A	101	0.0000	0.4898
3t49	A	70	0.8429	0.0000	1q5y	D	83	0.2593	0.4815	3fgv	A	93	0.3441	0.2151	3czz	A	101	0.0000	0.5545
1d3b	K	71	0.0986	0.5352	2w7v	A	83	0.2048	0.3133	3hty	P	93	0.0000	0.5806	3hpw	A	101	0.1287	0.3762
1h64	Z	71	0.0986	0.5634	2y3y	A	83	0.2892	0.4699	3mop	K	93	0.4624	0.0000	1l3p	A	102	0.6961	0.0000
1spb	P	71	0.2676	0.3099	3a38	A	83	0.1205	0.1566	3nfk	B	93	0.1413	0.3478	1tul	A	102	0.0000	0.5098
1vih	A	71	0.3662	0.2394	3v3l	A	83	0.1446	0.3253	4a56	A	93	0.7634	0.0000	2oya	B	102	0.0980	0.2353
3bxu	A	71	0.3803	0.1268	2i0x	A	84	0.2500	0.2738	1g2r	A	94	0.3723	0.1702	2v76	B	102	0.2059	0.4412
3swn	B	71	0.0845	0.5493	2oy9	A	84	0.6429	0.0476	2e1f	A	94	0.6596	0.0000	3eoy	G	102	0.0392	0.5098
1t0z	B	72	0.2500	0.2083	2vqe	P	84	0.2048	0.3012	2fb0	A	94	0.3723	0.1809	3ge3	E	102	0.2549	0.3627
3rpf	C	72	0.1250	0.2361	2xcj	B	84	0.5000	0.0000	2fqp	D	94	0.0000	0.5484	3koj	B	102	0.0978	0.4457
1i27	A	73	0.4795	0.1370	3awu	B	84	0.1519	0.4051	2x1f	A	94	0.2660	0.2553	3kt9	A	102	0.0000	0.4510
1uj8	A	73	0.4658	0.0000	3o46	A	84	0.1786	0.3690	3fpn	B	94	0.1064	0.4255	3t5s	A	102	0.3469	0.2449
1vjq	B	73	0.1918	0.2192	3r27	A	84	0.2500	0.3333	3jst	B	94	0.4362	0.2021	1ag4	A	103	0.0388	0.3689
3aji	D	73	0.5890	0.0000	1ji7	C	85	0.5294	0.0000	3lo3	Q	94	0.3617	0.2872	3bs1	A	103	0.0680	0.4854
3nzl	A	73	0.5342	0.0000	1lpb	A	85	0.0706	0.2235	3pd7	B	94	0.2979	0.1383	3chb	H	103	0.2427	0.3398
3sd4	A	73	0.0000	0.3971	2xqq	D	85	0.3882	0.3176	3uze	D	94	0.0000	0.4359	3kg5	B	103	0.0000	0.4563
1ldd	A	74	0.5676	0.1081	3c0f	B	85	0.1765	0.2471	1bja	A	95	0.6105	0.0842	3nzn	B	103	0.3107	0.2136
3h31	A	74	0.1250	0.0556	3mgj	B	85	0.3176	0.3176	2d1p	F	95	0.2842	0.2000	3u01	A	103	0.1650	0.4078
3plu	B	74	0.1622	0.3649	1r8h	F	86	0.2500	0.2738	3d9t	B	95	0.3684	0.0842	1i1j	B	104	0.0000	0.3592
3t7l	A	74	0.1351	0.1892	2fa8	A	86	0.3256	0.2791	3h7h	B	95	0.2421	0.1789	1lkt	F	104	0.0000	0.3173
3zzp	A	74	0.4054	0.3108	3n5b	B	86	0.3721	0.3372	3iez	B	95	0.2632	0.4316	1nlx	F	104	0.8077	0.0000
1h3l	A	75	0.7867	0.0000	3o2e	A	86	0.3372	0.2674	3sqf	A	95	0.0421	0.3684	1x8d	D	104	0.3077	0.2692
2g0c	A	75	0.2286	0.2571	3so0	E	86	0.0698	0.4419	1bte	A	96	0.0435	0.4457	1xau	A	104	0.0000	0.5192
1bg8	B	76	0.5526	0.0000	3hzb	H	87	0.0690	0.4713	1lm8	C	96	0.4270	0.1461	2q30	B	104	0.0000	0.4231
1dtj	B	76	0.3281	0.3125	3ld7	A	87	0.0000	0.4023	1vgl	A	96	0.3333	0.2500	3bgu	B	104	0.3462	0.2788
1v5i	B	76	0.3026	0.3421	3lf5	B	87	0.2644	0.2529	2gff	A	96	0.3438	0.2917	3ktb	D	104	0.2308	0.1731
2ayd	A	76	0.0000	0.4342	3tbn	A	87	0.0000	0.1839	3a0s	A	96	0.2500	0.3958	3l5w	J	104	0.5281	0.0449
3kuc	B	76	0.1316	0.3684	4dm5	A	87	0.2989	0.3103	3mpc	A	96	0.0000	0.4479	3nbm	A	104	0.2981	0.2115
3lvk	B	76	0.2895	0.2895	1tig	A	88	0.3523	0.3523	3u7z	B	96	0.0729	0.4583	3o5n	G	104	0.1458	0.3750
3swn	A	76	0.0921	0.5395	1wmi	A	88	0.2841	0.3409	1v05	A	96	0.0000	0.4583	1if1	A	105	0.2952	0.0762
3tdu	D	76	0.5921	0.1842	2xet	B	88	0.0000	0.4545	1mn8	D	97	0.5052	0.0000	1jr8	B	105	0.6762	0.0000
3vke	A	76	0.4605	0.2895	3d8l	C	88	0.7159	0.0000	1vrv	A	97	0.4227	0.1856	1nrv	B	105	0.1782	0.2574
3a9f	B	77	0.5844	0.0000	3od8	H	88	0.2386	0.0455	3khf	B	97	0.1340	0.3814	1tqg	A	105	0.8381	0.0000
3a9j	B	77	0.1558	0.2727	4a8x	A	88	0.2273	0.3636	3us4	A	97	0.1237	0.2784	1z9m	B	105	0.0000	0.4762
3e19	C	77	0.1688	0.4026	1i8n	A	89	0.1124	0.3371	3v7q	B	97	0.5052	0.1649	2a7u	B	105	0.7048	0.0000
3p04	A	77	0.3377	0.1688	1oqj	B	89	0.2135	0.2472	3ygs	P	97	0.6598	0.0000	3hqx	A	105	0.0000	0.4476
3phx	B	77	0.1558	0.3247	2w4s	A	89	0.7241	0.0000	2aib	A	98	0.5612	0.0612	3kgk	A	105	0.3978	0.2043
3tdq	B	77	0.0519	0.4416	2zkz	D	89	0.5542	0.0482	2b0l	A	98	0.5306	0.1429	3lys	A	105	0.4952	0.0000
3u97	A	77	0.2078	0.2338	3ge2	A	89	0.0000	0.4831	2c3v	B	98	0.0000	0.4526	3m9j	A	105	0.4095	0.2571
1pk3	C	78	0.5256	0.0000	1f60	B	90	0.2000	0.4889	2h3l	B	98	0.1020	0.3878	1bwu	A	106	0.0000	0.4151
1y0n	A	78	0.3611	0.1667	2bjd	B	90	0.2667	0.4333	2qzi	D	98	0.1837	0.3469	2iwo	B	106	0.1700	0.3700
3swn	F	78	0.0000	0.5769	2dyj	B	90	0.4222	0.2222	3fm8	B	98	0.0000	0.4062	3eo6	A	106	0.0000	0.4245
3tjy	A	78	0.5513	0.0000	2qkl	B	90	0.6000	0.0000	3knb	A	98	0.0000	0.6122	3mjq	B	106	0.2075	0.3585
2ffg	B	79	0.2658	0.3165	2yy3	C	90	0.1778	0.3778	1qb5	D	99	0.2626	0.3636	3u12	B	106	0.1154	0.4519
2p5m	C	79	0.4304	0.3165	3bee	B	90	0.7111	0.0000	2i45	I	99	0.0612	0.4286	3zzy	A	106	0.1792	0.3585
2xnq	A	79	0.2051	0.3205	3g73	A	90	0.4000	0.1333	2pmu	F	99	0.3723	0.2447	1yfn	D	107	0.1308	0.3832
3iuw	A	79	0.1013	0.4430	3l46	A	90	0.2556	0.1556	3a2o	B	99	0.0404	0.4848	2bse	C	107	0.0000	0.4953
3md1	B	79	0.2405	0.3924	3mx7	A	90	0.0000	0.6556	3dlv	A	99	0.3011	0.2473	2hqs	E	107	0.3832	0.1682
1b34	A	80	0.1250	0.4625	3pe9	D	90	0.0000	0.5349	3eud	A	99	0.0404	0.3838	2pkd	F	107	0.0000	0.4860
2nzc	A	80	0.3125	0.3875	1mid	A	91	0.5604	0.0000	3gz7	A	99	0.2121	0.2828	3hze	A	107	0.3271	0.2056
2ofy	B	80	0.6625	0.0000	1v76	B	91	0.0440	0.3846	3luu	A	99	0.1444	0.2778	3q7r	B	107	0.3196	0.2577
3bqp	B	80	0.6375	0.0000	3hms	A	91	0.1099	0.4286	3ns6	A	99	0.3232	0.2424	3s8s	A	107	0.2804	0.2804
3fau	A	80	0.4583	0.2361	3ima	D	91	0.1977	0.5349	3oq2	A	99	0.2828	0.2727	1j0p	A	108	0.2130	0.0926
3k65	A	80	0.0875	0.0750	3ov8	A	91	0.6374	0.1538	1ejx	A	100	0.5000	0.1600	1q8d	A	108	0.5248	0.0000
3l9a	X	80	0.2625	0.3625	1s29	A	92	0.4239	0.0435	1mwq	A	100	0.3232	0.3131	2coq	A	108	0.0000	0.5556
3p5t	Q	80	0.2375	0.3125	1uad	D	92	0.0000	0.4674	2w7a	A	100	0.2700	0.2800	2d0o	D	108	0.3333	0.2500
3qq8	B	80	0.1250	0.3875	2axi	A	92	0.4239	0.2174	3gk5	A	100	0.3500	0.1600	2yad	A	108	0.1923	0.2692
1t4o	A	81	0.3704	0.2346	2c3h	H	92	0.0000	0.4348	3k0x	A	100	0.1313	0.4545	3bb6	A	108	0.0000	0.3519
1wwh	A	81	0.2593	0.3086	2hts	A	92	0.3596	0.1348	3nsw	G	100	0.0722	0.4124	3d1b	C	108	0.7407	0.0000
2hnu	E	81	0.0000	0.3704	2ozf	A	92	0.1630	0.3370	3pt3	A	100	0.4607	0.1910	3hnx	A	108	0.0000	0.5648
3dnj	B	81	0.5185	0.2469	2vpv	B	92	0.0000	0.5870	3rhb	A	100	0.5300	0.1700	3va4	B	108	0.0000	0.4167
3g5o	B	81	0.2716	0.3333	3pa6	C	92	0.3152	0.1739	3s8w	C	100	0.0000	0.3871	1nz0	A	109	0.4312	0.2018

1oj5	A	109	0.2642	0.3585	1tvd	A	116	0.0000	0.5089	1gy6	B	123	0.2439	0.4634	1zbf	A	132	0.3561	0.2273
2h28	A	109	0.0917	0.3211	2idl	B	116	0.3966	0.2672	1mb3	A	123	0.4492	0.1864	2f3l	A	132	0.0916	0.1221
2nqd	A	109	0.0000	0.4679	3eeh	A	116	0.2931	0.3276	2dqa	B	123	0.4715	0.0813	2hhg	A	132	0.3182	0.1439
2xfd	A	109	0.0000	0.5321	3es4	B	116	0.0000	0.4483	2h5n	D	123	0.6829	0.0000	3ak0	B	132	0.0000	0.6667
2y6x	A	109	0.7523	0.0000	3u4v	A	116	0.0862	0.3879	2zay	A	123	0.4472	0.2114	3d9r	A	132	0.2803	0.4470
2z0t	A	109	0.3303	0.2936	1b1u	A	117	0.3761	0.0342	3cxg	B	123	0.2683	0.3415	3grd	B	132	0.2197	0.3712
3l4h	A	109	0.3578	0.1193	1dqt	A	117	0.0000	0.5556	3do8	B	123	0.4065	0.1870	3obl	A	132	0.0000	0.5758
3lax	A	109	0.2617	0.2710	1elw	A	117	0.7949	0.0000	3f6g	A	123	0.2705	0.4016	3pmc	B	132	0.7727	0.0000
3lwc	A	109	0.0000	0.5096	1pm4	C	117	0.0000	0.5128	3kf8	D	123	0.1220	0.3496	3q63	A	132	0.2652	0.2879
1e50	Q	110	0.0000	0.4000	1rlg	B	117	0.4359	0.1624	3qzm	B	123	0.0000	0.5772	3ts9	A	132	0.7519	0.0000
1ejf	B	110	0.0000	0.5000	1zld	A	117	0.0388	0.5631	3zw5	B	123	0.1789	0.4065	1dyt	A	133	0.1805	0.3158
1jer	A	110	0.1182	0.2636	2ar5	A	117	0.3590	0.2051	1dy5	A	124	0.1855	0.3306	1ry9	D	133	0.3759	0.2932
1xed	C	110	0.0000	0.4623	2h8c	C	117	0.2857	0.2857	2htd	B	124	0.1855	0.4113	2gey	B	133	0.3008	0.4211
2qhl	A	110	0.0000	0.4727	2p8i	A	117	0.2991	0.2051	2hw4	A	124	0.2137	0.3675	2rgq	C	133	0.1955	0.3985
2qzt	B	110	0.3909	0.2545	3nsu	A	117	0.1091	0.3818	2rbg	A	124	0.3548	0.2500	1a78	B	134	0.0000	0.6343
3g48	A	110	0.0818	0.4636	1unq	A	118	0.2308	0.4188	3blz	A	124	0.2419	0.4758	1cnu	A	134	0.3233	0.3158
3ixs	G	110	0.2475	0.2970	2ope	C	118	0.4248	0.2035	3k1e	A	124	0.5645	0.0000	1p5u	C	134	0.0000	0.4046
3plw	A	110	0.4037	0.0917	2xod	A	118	0.3644	0.2203	3pkz	G	124	0.4715	0.1707	2bnl	F	134	0.7462	0.0000
1h4x	B	111	0.4414	0.2883	3bcw	A	118	0.0000	0.5763	1t1j	B	125	0.5040	0.1360	2fm8	A	134	0.3731	0.2090
1u5f	A	111	0.1261	0.3784	3evi	A	118	0.2458	0.2373	1vzi	A	125	0.0320	0.4400	2wnv	D	134	0.0000	0.5564
1xqa	B	111	0.2252	0.2252	3h7h	A	118	0.2712	0.1949	2ftb	A	125	0.1200	0.5600	3b9c	C	134	0.0000	0.5821
2dyn	A	111	0.1171	0.4324	3kdf	C	118	0.1111	0.3162	2jba	A	125	0.3680	0.2400	3d7a	B	134	0.3284	0.2910
3f1p	B	111	0.1441	0.3333	3mj0	A	118	0.1864	0.2797	3eqz	B	125	0.4000	0.2080	3l4a	A	134	0.5820	0.0328
3gnj	A	111	0.3604	0.1351	3p4h	A	118	0.0000	0.6496	3klr	A	125	0.1840	0.3520	3pg6	D	134	0.1866	0.3209
3gx8	A	111	0.4775	0.1622	3t6o	A	118	0.3390	0.2203	3pna	B	125	0.2960	0.3120	3q6a	H	134	0.2239	0.3284
3mez	A	111	0.0000	0.3514	1hy5	A	119	0.5126	0.0336	3ql9	A	125	0.2720	0.0960	1sen	A	135	0.3284	0.1269
1jli	A	112	0.5536	0.0000	1p28	A	119	0.6387	0.0000	3upv	A	125	0.8160	0.0000	1vl7	A	135	0.2593	0.3926
1qau	A	112	0.1339	0.4286	1pbj	A	119	0.2857	0.2605	3zxo	B	125	0.3200	0.2320	1xt5	A	135	0.0000	0.4741
2x4i	B	112	0.2692	0.2212	1qfo	C	119	0.0000	0.4138	1tp6	A	126	0.2778	0.4683	2y4z	A	135	0.6000	0.0444
3b64	A	112	0.3036	0.2411	1wk2	A	119	0.0714	0.4762	2gkm	B	126	0.6000	0.0000	3raz	A	135	0.2296	0.2074
3nuf	B	112	0.5625	0.0000	2gj3	B	119	0.2353	0.3277	2hq3	A	126	0.1746	0.2778	3rcc	B	135	0.0472	0.4016
3rnq	A	112	0.0000	0.5536	2hqt	H	119	0.5714	0.0000	2owa	B	126	0.4683	0.0714	3rmh	B	135	0.1716	0.3433
1hfo	F	113	0.2832	0.2566	2rhk	A	119	0.2353	0.4202	3ehc	C	126	0.2619	0.3492	4dfa	A	135	0.2636	0.2946
1v9y	B	113	0.2115	0.3462	2wno	A	119	0.0000	0.4454	3f7e	B	126	0.1746	0.3571	1eca	A	136	0.7132	0.0000
2b7l	C	113	0.3009	0.2301	3mfx	C	119	0.1770	0.1947	3l2a	A	126	0.4206	0.1349	2cdo	C	136	0.0000	0.5147
2h0e	B	113	0.0619	0.3097	3qor	B	119	0.0000	0.3277	3ljd	B	126	0.1746	0.3333	3bvp	B	136	0.4504	0.2061
2msb	B	113	0.1858	0.3186	3s0a	A	119	0.6387	0.0000	3m1x	A	126	0.2381	0.3095	3h87	A	136	0.5147	0.1176
3erj	B	113	0.4107	0.2321	3snk	A	119	0.3025	0.2353	3qxv	E	126	0.0000	0.4188	3p0t	B	136	0.3603	0.1838
3gwn	A	113	0.6283	0.0000	1ecs	B	120	0.1583	0.3917	1qmp	B	127	0.4228	0.2114	3ui4	A	136	0.2772	0.2772
3onh	A	113	0.1593	0.3363	2aj6	A	120	0.3667	0.2000	2vzp	B	127	0.0000	0.5669	1hqz	4	137	0.2993	0.2336
3s6e	A	113	0.2743	0.1593	2ckk	A	120	0.0000	0.4417	3fzw	A	127	0.2835	0.3937	2g1u	A	137	0.3630	0.2000
1rw1	A	114	0.4298	0.1404	2nvn	A	120	0.1833	0.3833	4a2v	A	127	0.1057	0.3902	3h96	B	137	0.3162	0.2794
1xod	A	114	0.1698	0.4717	2r4i	B	120	0.2750	0.5417	1i5n	B	128	0.7040	0.0000	3lb2	B	137	0.5735	0.0000
2bq4	A	114	0.2105	0.0526	2was	B	120	0.3529	0.2689	1oa8	A	128	0.1406	0.4141	3m7o	A	137	0.0000	0.5766
2gsc	C	114	0.7658	0.0000	3bb9	F	120	0.2167	0.4750	2bkm	B	128	0.6406	0.0000	3n70	F	137	0.3578	0.0734
2iay	A	114	0.1404	0.1842	3lyg	A	120	0.3000	0.3000	2fyg	A	128	0.2422	0.0938	3ohe	A	137	0.2868	0.1471
2r78	A	114	0.3070	0.2368	3lyi	A	120	0.3333	0.1917	3ct6	B	128	0.4297	0.1797	2r1a	A	138	0.0000	0.3712
3agn	A	114	0.1404	0.2632	3m3g	A	120	0.1417	0.2917	3rob	D	128	0.3047	0.3672	3f5o	H	138	0.2174	0.3841
3aps	A	114	0.3772	0.2544	3q2b	A	120	0.3500	0.2917	1ixl	A	129	0.0853	0.4419	3iho	A	138	0.5870	0.0000
3mwz	A	114	0.0789	0.4211	1oi0	A	121	0.2110	0.3486	2cx7	A	129	0.3488	0.2636	3jyb	B	138	0.0580	0.3551
3rtl	A	114	0.0000	0.5789	1oko	A	121	0.0000	0.4298	2ox8	A	129	0.1550	0.2636	3oj0	A	138	0.3188	0.1812
3slz	B	114	0.0614	0.4298	1ue7	B	121	0.1000	0.4500	3dns	B	129	0.2016	0.3101	3zsj	A	138	0.0000	0.6087
3ult	A	114	0.0000	0.4825	1zgz	A	121	0.3884	0.2066	3f6c	B	129	0.3721	0.2171	1hdk	A	139	0.0000	0.5324
1f86	B	115	0.0609	0.4870	2f9h	A	121	0.0992	0.3223	3ksp	A	129	0.1860	0.4651	1jkg	A	139	0.2518	0.4317
1h8u	A	115	0.2000	0.3304	2rh3	A	121	0.4959	0.0992	3kyj	A	129	0.8062	0.0000	1shx	B	139	0.0725	0.3406
1nqj	B	115	0.1053	0.5614	2x4j	A	121	0.0000	0.5091	3lr2	B	129	0.7364	0.0000	1th8	A	139	0.3383	0.3083
1pqh	B	115	0.0957	0.4609	2x5p	A	121	0.0000	0.5238	3siq	D	129	0.2449	0.1122	2ej8	A	139	0.2537	0.4403
1th8	B	115	0.4261	0.3217	3da5	A	121	0.3306	0.2066	2q22	C	130	0.2769	0.2000	2hfn	F	139	0.3525	0.2878
2opc	A	115	0.0000	0.3391	1ijy	A	122	0.3279	0.0328	2xr6	A	130	0.1846	0.3154	2jda	A	139	0.0000	0.5612
2vpk	A	115	0.5739	0.1043	1jif	B	122	0.1803	0.3361	2zqo	B	130	0.0000	0.3923	2oik	D	139	0.2590	0.1942
2z0b	A	115	0.0000	0.4821	1lwb	A	122	0.6148	0.0000	3cu5	A	130	0.3840	0.1840	2wy4	A	139	0.7770	0.0000
3h8u	B	115	0.0348	0.4000	2oiz	H	122	0.0328	0.2541	2r8u	A	131	0.4962	0.0000	3kf6	A	139	0.2555	0.3869
3mnm	A	115	0.0000	0.4336	2pn0	C	122	0.2213	0.2213	3fz2	A	131	0.2093	0.3876	3p6d	A	139	0.1223	0.5468
3o1c	A	115	0.2435	0.2174	2q00	B	122	0.7377	0.0000	3lyu	F	131	0.3664	0.1908	1exz	B	140	0.4000	0.0571
3pp2	A	115	0.1327	0.4159	2y78	A	122	0.0492	0.3361	3m1i	B	131	0.1450	0.4504	1q4u	A	140	0.1786	0.4143
3soj	B	115	0.2522	0.2957	3lbe	A	122	0.1475	0.4836	3nph	B	131	0.6107	0.0000	1q8c	A	140	0.6241	0.0000
1fo0	B	116	0.0000	0.5446	3n53	B	122	0.4271	0.2396	1gr3	A	132	0.0000	0.5530	1wmz	D	140	0.1643	0.3429
1jx7	A	116	0.2845	0.1897	3nkl	A	122	0.3033	0.2049	1y7r	A	132	0.2424	0.1742	1x9f	D	140	0.7286	0.0000

3d9n	B	140	0.7143	0.0000	1dzk	B	148	0.0878	0.4932	2a25	A	158	0.1611	0.4765	3eti	A	168	0.3631	0.2262
3ivv	A	140	0.0357	0.5000	1f08	A	148	0.3446	0.2838	2vvp	E	158	0.4747	0.1519	1mgt	A	169	0.3846	0.1657
3mdp	A	140	0.2424	0.2348	1q1f	A	148	0.6757	0.0000	2wtg	A	158	0.6646	0.0000	2pq5	A	169	0.4379	0.1243
3r5z	A	140	0.2929	0.3357	1v96	A	148	0.5850	0.1293	3ily	A	158	0.4276	0.1172	3qjg	L	170	0.3353	0.2471
3u3g	A	140	0.3714	0.3143	1xeb	F	148	0.1986	0.2329	3rof	A	158	0.3797	0.1392	1knq	B	171	0.4620	0.1579
4a34	T	140	0.3214	0.1857	2ost	D	148	0.2388	0.2910	2bjn	B	159	0.4832	0.1611	1oqv	A	171	0.3450	0.2164
2ofc	A	141	0.1064	0.5319	2qhk	A	148	0.3082	0.2397	3cgl	F	159	0.0615	0.2000	1txj	A	171	0.2848	0.2595
2y9w	D	141	0.0000	0.4412	3ef8	A	148	0.1837	0.4762	3dau	A	159	0.1950	0.3145	2x5y	A	171	0.2515	0.3216
3d89	A	141	0.0511	0.4599	3nbc	B	148	0.0000	0.3919	1i12	D	160	0.2975	0.2658	2y0o	A	171	0.0819	0.3392
3gix	A	141	0.3191	0.2340	3u80	A	148	0.4341	0.2093	1od6	A	160	0.4103	0.1987	3bwz	A	171	0.0409	0.4094
3gzr	B	141	0.2199	0.3901	1aqz	A	149	0.0979	0.2797	1svp	B	160	0.0000	0.4375	3kb2	B	171	0.4702	0.1429
3o5y	A	141	0.2979	0.2766	1gtz	B	149	0.3826	0.1611	1ww7	A	160	0.0250	0.5437	3rt2	A	171	0.2515	0.3801
3ztp	A	141	0.4468	0.1986	1l2h	A	149	0.0000	0.4589	3dh1	C	160	0.4125	0.1750	2b7j	C	172	0.3145	0.2013
1j3a	A	142	0.3769	0.1308	2aj7	B	149	0.0405	0.3108	3fyn	A	160	0.2810	0.2288	3by4	A	172	0.3663	0.1977
1m4r	A	142	0.6127	0.0000	2ecu	A	149	0.1477	0.4295	3o22	A	160	0.1582	0.4620	3kop	E	172	0.1205	0.3795
1sjw	A	142	0.2676	0.3451	2glz	A	149	0.4094	0.1745	3on9	B	160	0.0000	0.4688	3q62	A	172	0.1813	0.3392
1t82	D	142	0.2426	0.3750	2okv	D	149	0.2349	0.3758	3pnr	B	160	0.0000	0.5154	1lqv	B	173	0.3237	0.3642
2p39	A	142	0.0000	0.3169	2wj9	A	149	0.4610	0.1702	3v46	A	160	0.2313	0.2062	3hm2	H	173	0.3176	0.2000
2qiy	B	142	0.2444	0.4593	3s9d	A	149	0.6899	0.0000	4ds2	B	160	0.3248	0.1592	3tmp	E	173	0.4353	0.1294
2vn5	A	142	0.0282	0.5493	1x46	A	150	0.6867	0.0000	1p6o	B	161	0.4161	0.1801	1h4a	X	174	0.0000	0.4740
2wz8	A	142	0.0000	0.5074	3op6	B	150	0.2270	0.1348	1vi4	A	161	0.1429	0.2547	1yfu	A	174	0.1322	0.3506
3f7s	A	142	0.1831	0.3380	1ouw	D	151	0.0000	0.6358	1wu3	I	161	0.6584	0.0000	2aca	B	174	0.1724	0.3736
3le0	A	142	0.0000	0.4366	2vog	A	151	0.7397	0.0000	3g8w	A	161	0.2919	0.3043	3cne	D	174	0.1860	0.2267
3m7k	A	142	0.3803	0.0845	3emu	A	151	0.4483	0.1862	3mvc	B	161	0.7115	0.0000	1iaz	A	175	0.1086	0.4743
3qa9	A	142	0.3380	0.1549	3gm5	A	151	0.2517	0.3709	3n4j	A	161	0.3168	0.1553	2oh1	A	175	0.2586	0.2471
1aoh	A	143	0.0280	0.5524	3iis	M	151	0.7219	0.0000	1mtl	A	163	0.2975	0.1646	3mci	C	175	0.3943	0.1771
1hl6	B	143	0.3165	0.3525	3jud	A	151	0.0690	0.3241	2fpw	B	163	0.2761	0.1534	3ryk	B	175	0.0629	0.4343
1mzg	B	143	0.4545	0.0559	3l4r	A	151	0.0662	0.4901	3lmb	A	163	0.2883	0.2761	3d3o	A	176	0.3429	0.1771
2ig6	A	143	0.2028	0.3986	3nng	B	151	0.0530	0.4768	3vjz	B	163	0.7117	0.0000	3jtw	A	176	0.1818	0.2784
2x46	A	143	0.1119	0.3636	3pt8	A	151	0.7351	0.0000	1ryl	B	164	0.5232	0.1457	3vaa	B	176	0.5511	0.1250
2xfa	B	143	0.3217	0.2797	3qqq	B	151	0.6733	0.0000	1sbq	B	164	0.2744	0.1890	3bwv	A	177	0.2899	0.1065
3d6x	A	143	0.1314	0.3212	1ntv	A	152	0.3289	0.3289	2wuh	A	164	0.0000	0.3681	3h05	B	177	0.4096	0.1566
3iu6	A	143	0.5524	0.0000	1xm5	A	152	0.4474	0.1974	3cm3	A	164	0.4146	0.1098	3llu	A	177	0.3672	0.2034
3skj	F	143	0.0575	0.2989	3kff	A	152	0.0855	0.4868	3e8m	A	164	0.3293	0.1341	1usc	B	178	0.0899	0.3933
1ax8	A	144	0.6260	0.0000	3pr6	A	152	0.4041	0.1712	3fet	D	164	0.1951	0.3232	2ia1	A	178	0.4269	0.2047
1b93	C	144	0.3750	0.2361	3v4h	A	152	0.0741	0.5185	3mea	A	164	0.0000	0.3620	2wbn	A	178	0.2921	0.2584
1eyh	A	144	0.6944	0.0000	1hzt	A	153	0.2614	0.3007	1bgc	A	165	0.7044	0.0000	3gxb	B	178	0.3390	0.2203
1qdd	A	144	0.1597	0.2847	2flh	B	153	0.2484	0.4444	2f5j	A	165	0.6456	0.0380	3mw4	A	178	0.0337	0.5056
1w1d	A	144	0.2098	0.4196	3bt5	A	153	0.7895	0.0000	2ggz	B	165	0.5697	0.0364	3q46	A	178	0.1685	0.2697
3ht1	A	144	0.0284	0.4326	3zuc	A	153	0.0000	0.5033	2xz4	A	165	0.2667	0.1818	2acf	D	179	0.3464	0.1788
3jrv	A	144	0.5245	0.0000	1nqu	D	154	0.4675	0.2208	3jrn	A	165	0.4803	0.1645	2igi	B	179	0.4302	0.1788
3o2r	B	144	0.6667	0.0000	1rfz	D	154	0.6623	0.0000	3tiw	A	165	0.1366	0.4037	2iu1	A	179	0.6742	0.0000
3obq	A	144	0.2766	0.2695	2xhf	B	154	0.3052	0.2273	1alu	A	166	0.7025	0.0000	3mol	B	179	0.2443	0.3409
1v7m	X	145	0.5862	0.0552	2xst	A	154	0.1067	0.4467	1oxk	E	166	0.6329	0.0000	3qsz	B	179	0.1620	0.3966
2hd9	A	145	0.1655	0.3310	3a57	A	154	0.0519	0.5584	2yg2	A	166	0.1037	0.4390	1ej0	A	180	0.3722	0.2778
2xom	A	145	0.0000	0.5724	3fiq	A	154	0.1104	0.4545	3cb0	A	166	0.1566	0.4458	1xiy	A	180	0.2543	0.2601
3cbn	A	145	0.0552	0.3241	3gmx	A	154	0.1623	0.3052	3lfj	B	166	0.3614	0.2470	2i74	A	180	0.0333	0.4944
3kgz	B	145	0.0897	0.4483	3h2d	B	154	0.2532	0.2338	3m7a	B	166	0.1071	0.3214	3dal	B	180	0.1243	0.3018
3qfg	A	145	0.0828	0.3724	3piw	A	154	0.6234	0.0000	3nje	A	166	0.1720	0.4076	3qzr	A	180	0.0722	0.5056
3qm9	A	145	0.6690	0.0000	1w94	A	155	0.1613	0.2903	3p7x	A	166	0.2349	0.2651	3r8j	A	180	0.2056	0.3111
1w4s	A	146	0.0274	0.4726	2b06	A	155	0.1457	0.3841	3q72	A	166	0.3475	0.2837	3r9f	B	180	0.3111	0.3556
2a0j	A	146	0.4178	0.2192	3q64	A	155	0.2516	0.3871	3s6m	A	166	0.1220	0.2927	1jh6	A	181	0.3039	0.3370
2anx	A	146	0.5274	0.0685	3sao	A	155	0.1192	0.4503	4a02	A	166	0.0241	0.4277	1m4i	A	181	0.2320	0.3536
2hvw	C	146	0.3288	0.2740	1dyo	A	156	0.0000	0.4167	1ej2	A	167	0.4251	0.1317	1oru	A	181	0.1492	0.2818
2prx	B	146	0.1750	0.2917	1f8y	B	156	0.3462	0.1218	2nrk	A	167	0.4182	0.1636	2orw	B	181	0.2343	0.2571
3g46	A	146	0.6986	0.0000	1juq	D	156	0.5897	0.0000	2y6h	A	167	0.0000	0.4910	3gqq	F	181	0.0303	0.3818
3ijm	A	146	0.2808	0.2329	2hti	A	156	0.1250	0.4062	3c5c	B	167	0.3653	0.2575	3lvy	D	181	0.6333	0.0000
1cxq	A	147	0.4514	0.1875	3ebr	A	156	0.0513	0.3462	3h5j	B	167	0.2754	0.2036	3pn3	B	181	0.2762	0.2320
1r0u	A	147	0.0000	0.7762	3fuy	C	156	0.2308	0.2436	3ha2	A	167	0.3012	0.1747	2j2j	D	182	0.0000	0.4066
1wpu	B	147	0.3265	0.2857	1wwz	B	157	0.3312	0.3567	3he1	C	167	0.0940	0.4497	2r6v	A	182	0.0879	0.3132
1wwi	A	147	0.6531	0.0272	1yb0	A	157	0.3822	0.1529	3hnm	D	167	0.0000	0.3593	3efy	B	182	0.3736	0.1648
1zzw	A	147	0.4150	0.1565	2qv8	A	157	0.1586	0.3241	3mmh	A	167	0.3892	0.2455	3f95	A	182	0.0227	0.5341
2azw	A	147	0.2177	0.4218	3alu	A	157	0.1401	0.3121	3pyi	A	167	0.1329	0.5175	1kaq	F	183	0.3432	0.1479
2gmy	A	147	0.6939	0.0000	3i7d	A	157	0.0513	0.3077	4acj	A	167	0.1198	0.3174	1yre	A	183	0.3060	0.3552
2hbg	A	147	0.7143	0.0000	3no8	B	157	0.0000	0.6115	1nwa	A	168	0.2381	0.2381	2fhp	A	183	0.2514	0.2514
2huh	A	147	0.0000	0.5068	3pnx	F	157	0.3077	0.0769	1yn9	A	168	0.3810	0.1131	3hxi	A	183	0.2896	0.2787
2q9k	A	147	0.1156	0.3741	3qhp	B	157	0.3922	0.1765	2j3w	B	168	0.5030	0.1677	3n79	A	183	0.3989	0.3552
2zs0	C	147	0.6531	0.0000	1bow	A	158	0.1781	0.3904	2xol	A	168	0.6964	0.0000	3s8k	A	183	0.0000	0.3497

2gmw	B	184	0.3315	0.1522	3lqb	A	198	0.2020	0.1616	3ea6	A	215	0.1767	0.3814	3oru	A	231	0.0693	0.2078
2isb	A	184	0.1467	0.3043	3shg	A	198	0.5101	0.0505	3mq2	A	215	0.3023	0.2884	1rep	C	232	0.3889	0.2500
3g14	A	184	0.3466	0.1307	1z8h	B	199	0.4774	0.1106	3qi5	A	215	0.1490	0.2692	2fsq	A	232	0.2500	0.1121
3r3r	A	184	0.1359	0.3533	2j8q	B	199	0.2245	0.3622	3sib	A	215	0.5441	0.0196	2pk9	B	232	0.4861	0.0278
3u5r	H	184	0.2500	0.2120	2wz1	A	199	0.2980	0.3333	3cp7	A	216	0.0972	0.3565	3ne8	A	232	0.4053	0.1674
1cr5	A	185	0.1564	0.4022	1r45	D	200	0.3100	0.2300	3dr5	A	216	0.3704	0.1898	3s81	A	232	0.4741	0.1250
1pi1	A	185	0.5135	0.0216	2gz4	A	200	0.5200	0.0300	3qsq	A	216	0.0000	0.4722	3u3l	C	232	0.2338	0.1472
1r8n	A	185	0.0000	0.3135	2x12	A	200	0.0000	0.4300	2bo9	B	217	0.2396	0.4240	1l2u	A	233	0.4667	0.1689
1uww	B	185	0.0000	0.3575	3aia	B	200	0.3050	0.2450	2i9d	C	217	0.3077	0.1436	2vsh	B	233	0.3587	0.2332
2bba	A	185	0.0000	0.4486	3kl2	L	200	0.3850	0.1600	3flp	E	217	0.0369	0.4470	3mst	A	233	0.2661	0.1030
2ift	A	185	0.2542	0.3051	1kzl	A	202	0.1089	0.4158	3u2k	A	217	0.3351	0.2938	3uaw	A	233	0.3391	0.2747
2ns6	A	185	0.3152	0.1957	1qcs	A	202	0.1224	0.3214	2g6y	B	218	0.0185	0.5417	2fno	B	234	0.5556	0.0598
2sas	A	185	0.5676	0.0216	3abd	B	202	0.3163	0.2908	2i3d	A	218	0.2523	0.2018	2ri0	B	234	0.2906	0.1880
3o12	A	185	0.0000	0.4919	3mxo	A	202	0.2804	0.1958	2wf7	A	218	0.4450	0.1055	3cu2	B	234	0.3162	0.1923
1m3s	A	186	0.4560	0.1209	3r2q	A	202	0.5248	0.0644	3kys	C	218	0.1127	0.4930	1qwz	A	235	0.2170	0.2426
1tc1	B	186	0.2312	0.3011	2pnl	J	203	0.1823	0.2660	3nqa	B	218	0.4587	0.1789	1s5p	A	235	0.3480	0.2070
1zr3	A	186	0.3978	0.2097	3m0f	A	203	0.5665	0.0591	2cb9	A	219	0.3380	0.1737	2e10	B	235	0.2500	0.2902
2b0a	A	186	0.0806	0.3226	2ofw	C	204	0.4265	0.1324	3lgi	C	219	0.1157	0.3889	2fzv	A	235	0.3362	0.1404
3gzy	B	186	0.1882	0.3441	3bbd	A	204	0.2206	0.1667	2abw	A	220	0.2581	0.3226	3l77	A	235	0.4213	0.1532
3p6b	B	186	0.0000	0.4624	3tyt	A	204	0.2206	0.2500	3rpp	A	220	0.5023	0.1198	3no6	A	235	0.5915	0.0000
3v4k	A	186	0.3548	0.1935	2p4f	A	205	0.3906	0.2500	3tfw	B	220	0.3790	0.1826	3mbr	X	236	0.0000	0.5339
1m06	G	187	0.0000	0.4813	2pgc	E	205	0.2780	0.3024	4die	C	220	0.4533	0.1495	1ufo	D	237	0.2954	0.2025
1fvk	A	188	0.5053	0.1064	2vtw	F	205	0.0195	0.4390	1f3a	B	221	0.5430	0.0724	2bla	A	237	0.2569	0.2982
1x82	A	188	0.0851	0.3298	3mh9	C	205	0.1268	0.4829	3kg4	A	221	0.1237	0.3817	3b5o	A	237	0.6494	0.0000
2h29	A	188	0.3936	0.1702	3otm	A	205	0.1073	0.3268	3pkv	A	221	0.1357	0.3529	3dl1	A	237	0.4387	0.0613
3dcz	A	188	0.2775	0.2486	2p6w	A	206	0.3107	0.1796	1dfm	B	222	0.2569	0.3073	3nua	B	237	0.2911	0.3544
1ny7	1	189	0.0423	0.4233	2xtm	B	206	0.4000	0.1850	3f2k	B	222	0.4946	0.1183	3sc0	A	237	0.3080	0.1561
2raf	C	189	0.2698	0.1799	3o0p	A	206	0.1429	0.2857	3f5c	B	222	0.6793	0.0000	1tq5	A	238	0.0000	0.4829
2w8m	A	189	0.3452	0.2500	3qr5	B	206	0.0000	0.4268	3kmh	B	222	0.1222	0.3167	3seb	A	238	0.1345	0.3655
3se2	A	189	0.2299	0.2941	3rzn	A	206	0.5874	0.0000	3oqi	A	222	0.5631	0.1396	3t8b	B	238	0.3564	0.2178
1ou0	A	190	0.3526	0.1211	1jy5	A	207	0.2500	0.2108	1tqj	C	223	0.3105	0.1963	1pq9	D	239	0.7175	0.0314
2r2a	A	190	0.0423	0.2169	2as9	A	207	0.0676	0.3768	3fvv	A	223	0.4753	0.1839	2yln	A	240	0.3583	0.2417
3a2z	A	190	0.1000	0.3526	2p8j	A	207	0.2464	0.3188	3ifw	A	223	0.3543	0.1973	1b5e	A	241	0.3610	0.2158
3qzx	A	190	0.6053	0.0000	3da8	B	207	0.3447	0.2718	2fea	B	224	0.3616	0.1205	1ce7	A	241	0.3527	0.2324
1rre	E	191	0.2723	0.2199	3zqu	A	207	0.4106	0.1401	2ocz	A	224	0.2511	0.1689	3lw6	A	241	0.2075	0.2324
2qg6	A	191	0.2967	0.1484	1imj	A	208	0.3077	0.2163	3buu	A	224	0.0759	0.4062	3q6x	B	241	0.2158	0.2739
2x5n	A	192	0.3579	0.2105	2bu3	B	208	0.2976	0.2000	3fmf	D	224	0.4118	0.1946	3s5b	A	241	0.1561	0.2954
3b21	A	192	0.3542	0.1250	3ess	A	208	0.2550	0.2350	3jxo	A	224	0.0952	0.3571	1deu	A	242	0.2355	0.2397
3msx	B	192	0.6250	0.0000	3ig2	B	208	0.2071	0.2626	3qxh	A	224	0.4018	0.1473	3gne	B	242	0.0000	0.4545
3o2q	E	192	0.4583	0.1875	3ne0	A	208	0.2115	0.1923	1mun	A	225	0.5778	0.0000	3pa8	B	242	0.2521	0.2769
2i7d	A	193	0.3109	0.1347	3ts3	A	208	0.0882	0.3039	3c3y	A	225	0.4133	0.1956	3tdn	B	242	0.3544	0.3038
2pth	A	193	0.3886	0.2124	1vjn	A	209	0.1066	0.2893	3oii	A	225	0.2627	0.2396	1jzt	B	243	0.2922	0.1481
2xbl	C	193	0.5855	0.1140	2bdq	B	209	0.3158	0.1292	2yc3	A	226	0.3136	0.2182	2eix	A	243	0.2058	0.3539
3ngw	A	193	0.3575	0.1503	3n0u	A	209	0.7033	0.0000	3ltx	D	226	0.6593	0.0398	2a5z	C	244	0.0329	0.4280
1gqp	A	194	0.1099	0.4396	3t0h	A	209	0.2981	0.2308	3tk9	A	226	0.0708	0.3540	3bh2	A	244	0.1189	0.5246
1mf7	A	194	0.3505	0.2113	1axd	B	210	0.5359	0.0766	3zud	A	227	0.0485	0.2996	3f1l	B	244	0.4057	0.1270
1nxm	A	194	0.0412	0.3608	1ro2	A	210	0.2952	0.2190	4a3x	A	227	0.0176	0.3612	1h7e	A	245	0.3306	0.2327
1pp0	A	194	0.2708	0.4010	3kzx	A	210	0.4078	0.1699	1vgw	E	228	0.2977	0.2372	1qq5	B	245	0.4571	0.1102
2bbr	A	194	0.6053	0.0000	3ntv	A	210	0.3095	0.1714	2e2r	A	228	0.6432	0.0308	3fle	B	245	0.2049	0.1967
2car	B	194	0.3402	0.2423	3q7c	A	210	0.3238	0.1476	3p94	A	228	0.3480	0.0980	3lkk	A	245	0.3433	0.2060
3kh1	B	194	0.5474	0.0421	3tu8	A	210	0.1095	0.4190	1jyk	A	229	0.2096	0.2489	3rqz	C	245	0.2776	0.1755
3lqk	A	194	0.1856	0.1237	3u9h	B	210	0.1952	0.2857	2anu	A	229	0.2978	0.2622	1jw9	B	247	0.3817	0.1867
1itv	A	195	0.0410	0.4462	2gef	A	211	0.1608	0.3317	2p7i	B	229	0.3100	0.2271	3tx2	A	247	0.3036	0.1822
1svi	A	195	0.3388	0.2350	3giu	A	211	0.2607	0.2038	3qu5	B	229	0.4323	0.1048	2h00	C	248	0.3835	0.2330
3bhw	A	195	0.6324	0.0000	1eyq	B	212	0.3821	0.2783	3sm4	C	229	0.4585	0.1354	2j5i	A	248	0.4737	0.1700
1ioo	A	196	0.3316	0.1888	1me4	A	212	0.2115	0.1971	1dqp	B	230	0.3130	0.2130	3m0z	A	248	0.3105	0.1210
1t4w	A	196	0.1582	0.3878	2d39	B	213	0.1188	0.2921	1fsg	A	230	0.2348	0.2826	1u83	A	249	0.3656	0.1630
2in5	B	196	0.0370	0.4550	3bl5	E	213	0.4400	0.1200	2a7k	A	230	0.5174	0.1565	2ghc	X	249	0.4900	0.0161
3e3u	A	196	0.2143	0.2194	3evz	A	213	0.3384	0.2374	2h5c	A	230	0.0447	0.4413	2nxv	B	249	0.2289	0.2731
3lgb	B	196	0.5181	0.0000	3gze	D	213	0.1351	0.2865	2wur	A	230	0.0176	0.4978	2wk1	A	249	0.3909	0.1481
3nv0	A	196	0.2806	0.2653	1h0h	L	214	0.2430	0.2150	3lhi	A	230	0.3304	0.1696	3m3p	A	249	0.2702	0.1976
3r5g	B	196	0.3878	0.3418	1lvb	A	214	0.0187	0.3318	3s9c	A	230	0.0398	0.3230	1fpn	2	250	0.0360	0.3480
1zjr	A	197	0.4010	0.1878	1qca	A	214	0.2830	0.2877	1fx4	A	231	0.3723	0.2294	1isi	A	250	0.3240	0.1280
3imm	A	197	0.0000	0.3096	2v6k	A	214	0.5327	0.0654	2f6u	B	231	0.3680	0.1775	1p1x	A	250	0.4800	0.1440
3qb8	B	197	0.3756	0.2386	3o0a	B	214	0.2944	0.2290	2ixd	B	231	0.3913	0.1696	1sg4	C	250	0.5080	0.1520
3qpa	A	197	0.3198	0.1421	3see	A	214	0.0187	0.2897	3eei	A	231	0.3593	0.3030	3kkz	B	250	0.3360	0.2240
2cvd	A	198	0.5253	0.0960	1q6o	B	215	0.4419	0.2000	3kzp	B	231	0.3290	0.1948	3p2u	B	250	0.1880	0.2840
3jwi	A	198	0.3622	0.2143	2q0s	A	215	0.3628	0.1302	3ned	A	231	0.0437	0.5415	3ddo	B	251	0.2776	0.2980

3tc7	A	251	0.3904	0.1474	3sx2	H	269	0.4360	0.1520	3r24	A	292	0.2295	0.2192	2atm	A	324	0.3796	0.1481
1kcf	B	252	0.5043	0.1509	3u9q	A	269	0.6395	0.0504	3tc8	B	292	0.3322	0.1781	1o7j	C	325	0.2831	0.2031
1zmt	A	252	0.4246	0.1190	4a8j	F	269	0.2738	0.2091	3thr	D	292	0.2832	0.2762	2c29	F	326	0.4110	0.1595
3frh	A	252	0.4587	0.1901	1w98	B	270	0.5667	0.0000	3uzp	A	292	0.3425	0.1815	2dvt	C	326	0.4198	0.1204
3lkm	A	252	0.2249	0.2731	2a1l	A	270	0.2963	0.2593	3eat	X	293	0.1786	0.3179	3epw	A	326	0.4264	0.1595
3o1n	B	252	0.3968	0.1903	2zxk	A	270	0.3596	0.2022	3lye	A	293	0.4930	0.1294	3cq0	B	328	0.5525	0.1142
3p8a	B	252	0.1349	0.3968	3md7	A	270	0.1889	0.2815	1inl	A	295	0.2937	0.2972	3n3m	A	328	0.4543	0.1463
3q7z	A	252	0.3373	0.2421	1kqp	A	271	0.5166	0.0627	2hc1	A	295	0.3172	0.1931	2gn4	A	329	0.3374	0.1945
3bs4	A	253	0.3725	0.2591	1pn4	C	271	0.1807	0.3534	2oid	A	295	0.3321	0.1306	2vuw	A	329	0.2948	0.2128
3qsd	A	253	0.1937	0.1897	3lwx	A	271	0.3065	0.2462	3d7r	B	295	0.3129	0.0816	3mdy	A	329	0.3520	0.1526
3u62	A	253	0.3755	0.1897	3ppq	A	271	0.3985	0.1845	3ib7	A	295	0.2373	0.2237	3hrq	B	332	0.2422	0.3913
1sby	B	254	0.3976	0.1890	3p3c	A	272	0.2206	0.3346	2wm3	A	296	0.4291	0.1622	1y7t	B	333	0.4159	0.1835
2b7u	A	254	0.3452	0.1905	1sq4	A	273	0.0000	0.3247	2x7f	E	296	0.3619	0.1634	3rpw	A	333	0.3694	0.1261
2o55	A	254	0.2598	0.1614	2v3g	A	273	0.3553	0.1868	3cg7	A	296	0.3142	0.1588	2bw4	A	334	0.0716	0.3731
3hlx	A	254	0.6535	0.0000	2wj6	D	273	0.4615	0.1538	3odt	B	296	0.0000	0.5642	2vyn	D	334	0.2455	0.2575
3lho	A	254	0.2874	0.2402	3kw8	A	273	0.0403	0.3993	3pi6	A	297	0.3737	0.1852	3fbg	B	335	0.2814	0.2455
3pl0	A	254	0.1581	0.2806	3l80	A	273	0.3187	0.1465	1gz8	A	298	0.3185	0.1541	3v5a	A	335	0.2716	0.1910
1jt2	A	255	0.3137	0.2000	2uwa	A	274	0.0657	0.3869	2q0i	A	298	0.3356	0.2315	1yg9	A	336	0.0920	0.4417
3bf7	A	255	0.3804	0.1412	2v9l	A	274	0.3759	0.1679	2wkj	C	298	0.4966	0.1040	3ed1	C	336	0.3762	0.1716
2vws	A	256	0.4258	0.1562	2w1v	B	274	0.2555	0.2847	3ceg	A	298	0.3379	0.0990	3mz0	A	336	0.3274	0.2173
3amn	B	256	0.0469	0.5781	3o8q	A	274	0.3801	0.2214	3s40	D	298	0.2239	0.3060	1gxr	B	337	0.0000	0.5152
1xg5	C	257	0.4630	0.1440	4adu	B	274	0.4234	0.1314	3g5t	A	299	0.2642	0.1906	1jx6	A	338	0.4615	0.1686
2x61	A	257	0.3239	0.1417	2bji	B	275	0.3577	0.2409	1qhw	A	300	0.2000	0.2433	1n7h	B	340	0.4328	0.1612
3kv1	A	257	0.3228	0.1457	3ngx	B	275	0.3927	0.1527	3d02	A	300	0.3645	0.1839	1a99	D	341	0.4370	0.1672
3r0v	A	257	0.3463	0.1051	2w4j	A	276	0.3225	0.1775	2o4h	A	301	0.2292	0.2425	3rpd	B	341	0.3519	0.1408
1b0u	A	258	0.3969	0.2529	3fdj	A	276	0.3007	0.2428	3b59	F	301	0.1167	0.3200	2x5x	A	342	0.1842	0.1374
3ujc	A	258	0.4729	0.2054	3no0	C	276	0.0145	0.4674	3s25	A	301	0.0000	0.4662	3qve	C	343	0.1154	0.3615
1i7e	A	259	0.1245	0.3610	3rq5	A	276	0.3948	0.1181	1uf5	B	303	0.2409	0.2706	3rtx	A	343	0.2424	0.3030
3l0v	A	259	0.2891	0.1680	1pv2	F	277	0.2624	0.1749	1xto	A	303	0.0960	0.1755	3elf	A	345	0.4835	0.1171
3no3	A	259	0.2815	0.1933	2qik	A	277	0.1852	0.3667	2fp8	B	303	0.0495	0.4752	1ek6	A	346	0.3873	0.1734
3s83	A	259	0.3984	0.1797	3aay	A	277	0.3040	0.1355	3uuw	D	303	0.3630	0.2706	4dem	F	346	0.7139	0.0116
2a40	B	260	0.2692	0.2577	4dgq	C	277	0.4188	0.1516	3ewm	B	304	0.3038	0.2287	1guq	A	347	0.2594	0.1960
2i5i	B	260	0.2962	0.1269	1bkp	B	278	0.2734	0.2374	3ijd	B	304	0.4878	0.1498	2ox0	B	348	0.2299	0.2443
3epb	A	260	0.1929	0.3425	2y3c	A	278	0.1403	0.1906	1h1n	A	305	0.3355	0.1776	2zfi	A	349	0.3051	0.2689
3g91	A	260	0.2385	0.2462	3awu	A	278	0.3273	0.0144	1i1k	C	305	0.2475	0.2676	1jix	A	351	0.3077	0.1766
3hna	B	260	0.0615	0.2385	3e3m	D	278	0.4207	0.1882	3bny	D	305	0.7114	0.0000	2cf5	A	352	0.2386	0.2642
3oab	C	260	0.7078	0.0000	2gfq	C	279	0.2222	0.2437	1aq0	B	306	0.3399	0.1667	2poc	D	352	0.4516	0.1701
3oig	A	260	0.3808	0.1462	1vl0	A	280	0.3821	0.1679	1zl0	A	306	0.2908	0.2157	3amr	A	352	0.0199	0.4716
3vhv	A	260	0.6680	0.0508	2vla	A	280	0.5321	0.0571	2iiz	A	306	0.2092	0.1830	3db2	A	352	0.2968	0.2478
1ef8	A	261	0.4436	0.1595	3nuq	A	280	0.4296	0.0926	3sov	A	306	0.0000	0.4837	3fqg	A	352	0.2310	0.2848
1uwc	A	261	0.3065	0.2452	3npk	A	281	0.6767	0.0150	3cz8	A	308	0.3069	0.2673	2c0h	A	353	0.2946	0.1785
3c26	A	261	0.2364	0.3023	3rd7	A	281	0.1812	0.3696	2aeb	B	309	0.3689	0.1327	3a72	A	353	0.0227	0.4363
2gnp	A	262	0.3855	0.1374	3rl5	A	281	0.1964	0.2000	2xwv	A	309	0.4498	0.1942	3giy	A	353	0.3966	0.1275
2x9z	A	262	0.0267	0.4160	3fcx	A	282	0.3358	0.2435	3lo8	A	309	0.2244	0.2574	1fg7	A	354	0.3333	0.1582
3lum	D	262	0.1374	0.1527	3oen	A	282	0.3452	0.2460	2yxt	A	310	0.3758	0.1961	3kx6	D	355	0.4448	0.1453
3n0r	A	262	0.4903	0.0927	2c3a	A	283	0.0909	0.2925	3oaj	B	310	0.1516	0.3258	3gmi	A	356	0.5028	0.1180
3rpc	D	262	0.1718	0.1985	3rlg	A	283	0.3201	0.1691	1jr7	A	311	0.2443	0.2671	3m5q	A	357	0.3109	0.0336
1zrh	A	263	0.3916	0.1635	3acx	A	284	0.7218	0.0000	1xvx	A	311	0.3698	0.1543	3pwk	A	357	0.2857	0.2101
2qjv	B	263	0.0152	0.3270	3bf5	A	284	0.2465	0.2077	2q5r	C	311	0.3410	0.2361	2hds	B	358	0.3464	0.2067
3ijw	B	263	0.3080	0.1521	3qit	A	284	0.3759	0.2021	3oa2	D	311	0.3500	0.2267	3aqi	B	359	0.4457	0.1337
3ryd	A	263	0.3156	0.2319	1uoc	A	286	0.3887	0.1283	2e6f	A	312	0.3526	0.1987	3ojn	D	359	0.1504	0.3120
3u2u	B	263	0.3118	0.1901	2qrv	A	286	0.2637	0.1612	2aa3	A	313	0.4059	0.1848	3qc2	B	359	0.0336	0.3277
3mbk	B	264	0.2917	0.1061	3gdc	C	286	0.0804	0.4580	1e7s	A	314	0.4268	0.1561	3zwf	A	359	0.1901	0.3118
1xdn	A	265	0.1887	0.2792	3olj	A	286	0.6678	0.0000	3o4p	A	314	0.0000	0.4873	1w23	A	360	0.3556	0.1667
1y5m	B	265	0.4830	0.1434	1pzx	A	287	0.4101	0.2518	2hcr	B	315	0.3139	0.2492	2oui	D	360	0.2528	0.2278
3kws	B	265	0.3321	0.1094	2jl1	A	287	0.4425	0.1463	3h4x	A	315	0.2667	0.1619	3sg0	A	361	0.3823	0.1579
2dwu	C	266	0.4624	0.1767	2yb1	A	287	0.3789	0.1404	3tqe	A	316	0.4557	0.0696	3bpt	A	362	0.3840	0.0912
2qvp	A	266	0.2857	0.2406	3b4u	A	287	0.4669	0.1289	1ckn	B	317	0.1767	0.3912	3f0h	A	362	0.3241	0.1330
3ajd	A	266	0.2980	0.2745	1oi7	A	288	0.3838	0.2288	1qlw	B	317	0.2492	0.1861	3okf	B	362	0.4589	0.1841
3ctp	A	266	0.3496	0.1880	2uy2	A	288	0.3090	0.1806	2qmq	A	317	0.4173	0.1978	3rf7	A	362	0.3417	0.0944
3en0	C	266	0.2452	0.2797	3czq	A	288	0.3056	0.1250	1q35	A	318	0.4025	0.1604	3s9j	A	362	0.0249	0.3425
2eb4	B	267	0.2481	0.2895	3rd5	A	288	0.3717	0.1599	1ypf	A	318	0.3243	0.1453	2jer	H	365	0.2384	0.2356
2x9g	A	267	0.4348	0.1462	3ee4	A	289	0.6886	0.0000	4a8t	A	318	0.4277	0.1321	3os4	B	365	0.4329	0.1315
3iof	A	267	0.2952	0.2467	3qyj	B	290	0.4310	0.1724	3s7o	A	319	0.2880	0.2658	1mg7	B	367	0.3390	0.2740
3ois	D	268	0.2463	0.1716	3a5f	B	291	0.4536	0.1203	3qn1	B	321	0.2919	0.2852	1urs	A	367	0.4208	0.1940
3p8k	A	268	0.2388	0.2724	3cij	B	291	0.3162	0.2199	1wp5	A	322	0.0714	0.3354	3vmk	B	368	0.3995	0.1766
1ekq	A	269	0.5294	0.1412	3nre	A	291	0.0000	0.4467	3pt5	A	323	0.2243	0.1402	3alj	A	369	0.3171	0.2791
1q5r	N	269	0.3413	0.2381	1o58	B	292	0.3881	0.1503	1gve	B	324	0.4106	0.1424	3iu0	A	369	0.3268	0.1296

3g15 B 370 0.5000 0.1194	1toj A 405 0.4495 0.1389	3mm1 A 439 0.2232 0.1845	3pxl A 499 0.0501 0.3507
3kiz B 371 0.2453 0.1995	3ndi A 405 0.3309 0.2444	3byj A 440 0.0886 0.3886	2e4t A 509 0.2574 0.2731
4dnu A 372 0.0000 0.3844	3mqd A 407 0.3464 0.1818	3ryc B 442 0.4282 0.1759	2j04 B 511 0.0632 0.3621
2gdq B 373 0.3539 0.2091	1uuq A 410 0.3878 0.1341	2xfg A 446 0.4641 0.0291	3fot A 511 0.3524 0.2087
2v52 B 373 0.3729 0.2044	2yhg A 411 0.2628 0.2044	1qzq B 448 0.1686 0.1982	3sgg A 512 0.3223 0.1680
2jhf A 374 0.2353 0.2299	3v7p A 412 0.3683 0.2317	3se8 G 449 0.1734 0.3642	3v2u D 519 0.4272 0.1650
3led B 375 0.2693 0.1867	3szy A 413 0.3148 0.1985	1x9d A 453 0.4690 0.1062	1p1j B 524 0.3153 0.2224
1mdo A 376 0.3388 0.1257	1so2 D 415 0.5450 0.0109	2ij2 A 453 0.4812 0.1109	2y1k A 527 0.3061 0.1502
2efj A 377 0.4375 0.1733	2hzy A 416 0.1827 0.2740	2xhg A 457 0.3370 0.1991	1hxa A 528 0.6402 0.0000
3b7f A 378 0.0216 0.2749	1w9h A 417 0.3566 0.2369	1gkp E 458 0.3035 0.2096	3c2u B 537 0.0074 0.4432
2ip1 A 379 0.5054 0.0914	3a4r B 417 0.2105 0.2763	3gju A 458 0.3559 0.1332	3drf A 539 0.2319 0.2115
2p02 A 380 0.2816 0.2553	1ejd A 419 0.3055 0.2482	2qzu A 465 0.2284 0.1250	3h4t A 539 0.4322 0.1228
1y42 X 381 0.5391 0.0809	3bon A 422 0.2967 0.1699	3rqt A 465 0.2452 0.2366	3c9f A 546 0.2533 0.2158
3k5i B 381 0.3050 0.2626	1ra0 A 423 0.3262 0.2104	3vny A 466 0.2511 0.2918	1w9m A 553 0.4358 0.1049
1nof A 383 0.2193 0.2898	3qt9 A 424 0.4458 0.1085	3bnj A 471 0.4522 0.0594	3gsz B 563 0.4472 0.1234
3ozy A 386 0.3782 0.1684	2y27 A 425 0.2783 0.1840	3pz7 A 471 0.0930 0.3721	2ad7 A 571 0.0403 0.3538
3sf6 A 387 0.5220 0.1499	3nvs A 426 0.2864 0.2207	3kal A 475 0.3885 0.2548	2bhu A 589 0.2461 0.2306
1bi5 A 389 0.3856 0.2082	3p1v B 426 0.1111 0.2667	3u7q A 477 0.4319 0.1300	3moe A 624 0.2423 0.2197
1sq9 A 393 0.0105 0.4987	2xf3 A 427 0.3892 0.2146	3mk1 A 481 0.2703 0.1538	2w91 A 636 0.1512 0.3370
2ord B 393 0.3384 0.1730	3tkt A 427 0.5000 0.1238	2jc9 A 486 0.3547 0.1624	1h41 B 708 0.3746 0.1083
4dzi C 396 0.3496 0.1183	2npi B 428 0.2243 0.2453	2xfr A 487 0.3265 0.1047	2yfo A 719 0.1933 0.3018
1q0q A 398 0.4598 0.1482	2vwr A 430 0.1579 0.3789	3od3 A 488 0.0902 0.3730	1h16 A 759 0.4453 0.1186
2w8t A 398 0.3879 0.1814	3bzn A 430 0.3233 0.3512	1b3o A 490 0.3453 0.1564	1n62 E 796 0.2776 0.2513
3nc6 A 399 0.4840 0.0931	3qfh B 430 0.2934 0.1784	1w1o A 495 0.3091 0.2178	2gj4 A 824 0.4609 0.1466
3o3m A 399 0.5338 0.1003	3qww A 430 0.4651 0.1209	2ckw A 496 0.4180 0.1578	
3m7v B 401 0.3142 0.1945	2ptz A 431 0.3805 0.1694	2nt0 D 497 0.2394 0.2696	
1ht6 A 404 0.2723 0.1658	2vdu D 435 0.0503 0.4735	2vif A 498 0.2424 0.2273	
3aow C 404 0.4381 0.1460	3dnt A 436 0.2758 0.1631	3ive A 498 0.2444 0.2202	

Referências Bibliográficas

[1] W. Kabsch e C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. Programa disponível em `http://swift.cmbi.ru.nl/gv/dssp/`.

[2] K.A. Dill e J.L. MacCallum. The Protein-Folding Problem, 50 Years On. *Science*, 338(6110):1042–1046, November 2012.

[3] J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, e H. Wyckoff. A three-dimensional model of the myoglobin molecule obtained by X-Ray analysis. *Nature*, 181(4610):662–666, 1958.

[4] C.J. Epstein, R.F. Goldberger, e C.B. Anfinsen. The genetic control of tertiary protein structure: studies with model systems. In *Cold Spring Harbor symposia on quantitative biology*, volume 28, páginas 439–449. Cold Spring Harbor Laboratory Press, 1963.

[5] C. Levinthal. How to fold graciously. In P. et al. Debrunner, editor, *Mossbauer Spectroscopy in Biological Systems*, páginas 22–24, Urbana, Illinois, 1969. University of Illinois Press.

[6] K.A. Dill, S.B. Ozkan, M.S. Shell, e T.R. Weikl. The protein folding problem. *Annual review of biophysics*, 37:289–316, January 2008.

[7] M.H.J. Cordes, A.R. Davidson, e R.T. Sauer. Sequence space, folding and protein design. *Current Opinion in Structural Biology*, 6(1):3–10, 1996.

[8] O. Schueler e H. Margalit. Conservation of salt bridges in protein families. *Journal of Molecular Biology*, 248(1):125–135, 1995.

[9] J. Chen e W.E. Stites. Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry*, 40(50):15280–15289, 2001.

[10] Kauzmann W. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, 14:1–63, 1959.

[11] K.A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[12] K.A. Dill. Polymer principles and protein folding. *Protein Science*, 8(06):1166–1180, 1999.

[13] M.B. Swindells, M.W. MacArthur, e J.M. Thornton. Intrinsic φ, ψ propensities of amino acids, derived from the coil regions of known structures. *Nature Structural & Molecular Biology*, 2(7):596–603, 1995.

[14] G.E. Crooks e S.E. Brenner. Protein secondary structure: entropy, correlations and prediction. *Bioinformatics*, page 1321, 2004.

[15] A.D. Solis e S. Rackovsky. On the use of secondary structure in protein structure prediction: a bioinformatic analysis. *Polymer*, 45(2):525–546, 2004.

[16] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, e H.S. Chan. Principles of protein folding–a perspective from simple exact models. *Protein Science*, 4(4):561, 1995.

[17] W.A. Lim e R.T. Sauer. The role of internal packing interactions in determining the structure and stability of a protein. *Journal of Molecular Biology*, 219(2):359–376, 1991.

[18] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, M.H. Hecht, et al. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262(5140):1680–1702, 1993.

[19] M.H. Hecht, A. Das, A. Go, L.H. Bradley, e Y. Wei. De novo proteins from designed combinatorial libraries. *Protein Science*, 13(7):1711–1723, 2009.

[20] C.E. Schafmeister, S.L. LaPorte, L.J.W. Miercke, e R.M. Stroud. A designed four helix bundle protein with native-like structure. *Nature Structural & Molecular Biology*, 4(12):1039–1046, 1997.

[21] A.F. Pereira de Araújo, A.L.C. Gomes, A.A. Bursztyn, e E.I. Shakhnovich. Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 70(3):971–983, 2008.

[22] A.F. Pereira de Araujo e J.N. Onuchic. A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. *Proceedings of the National Academy of Sciences*, 106(45):19001, 2009.

[23] J. Moult, K. Fidelis, A. Kryshtafovych, e A. Tramontano. Critical assessment of methods of protein structure prediction (CASP)–round IX. *Proteins: Structure, Function, and Bioinformatics*, 79 Suppl 10:1–5, January 2011.

[24] D. Baker e A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.

[25] A. Sali e T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234:779–815, 1993.

[26] T. Defay e F.E. Cohen. Evaluation of current techniques for ab initio protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 23(3):431–445, 1995.

[27] J. Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3):285–9, June 2005.

[28] L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao, e N.V. Grishin. CASP9 assessment of free modeling target predictions. *Proteins: Structure, Function, and Bioinformatics*, 79 Suppl 10:59–73, January 2011.

[29] K.T. Simons, R. Bonneau, I. Ruczinski, e D. Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):171–176, 1999.

[30] K.T. Simons, C. Kooperberg, E. Huang, e D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.

[31] K.T. Simons, I. Ruczinski, C. Kooperberg, B.A. Fox, C. Bystroff, e D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Structure, Function, and Bioinformatics*, 34(1):82–95, 1999.

[32] P. Bradley, K.M.S. Misura, e D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

[33] S. Wu, J. Skolnick, e Y. Zhang. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC biology*, 5(1):17, 2007.

[34] A. Roy, A. Kucukural, e Y. Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 2010.

[35] Y. Zhang. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins: Structure, Function, and Bioinformatics*, 77 Suppl 9(August):100–13, January 2009.

[36] S. Wu e Y. Zhang. Muster: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics*, 72(2):547–556, 2008.

[37] Y. Zhang, A. Kolinski, e J. Skolnick. Touchstone ii: a new approach to ab initio protein structure prediction. *Biophysical Journal*, 85(2):1145–1164, 2003.

[38] Y. Zhang e J. Skolnick. Spicker: a clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, 25(6):865–871, 2004.

[39] D.E. Shaw, R.O. Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, e K.J. Bowers. Millisecond-scale molecular dynamics simulations on Anton. In *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*, páginas 1–11. IEEE, 2009.

[40] S. Piana, K. Lindorff-Larsen, e D.E. Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100(9):L47–L49, 2011.

[41] K. Lindorff-Larsen, S. Piana, R.O. Dror, e D.E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[42] A. Dickson e C.L. Brooks III. Native states of fast-folding proteins are kinetic traps. *Journal of the American Chemical Society*, 135(12):4729–4734, 2013.

[43] J.R. Rocha, M.G. van der Linden, D.C. Ferreira, P.H. Azevêdo, e A.F.P. de Araújo. Information-theoretic analysis and prediction of protein atomic burials: On the search for an informational intermediate between sequence and structure. *Bioinformatics*, 2012.

[44] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[45] S. Griep e U. Hobohm. PDBselect 1992–2009 and PDBfilter-select. *Nucleic Acids Research*, 38(suppl 1):D318–D319, 2010.

[46] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, e M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535–542, 1977.

[47] A.L.C. Gomes, J.R. de Rezende, A.F. Pereira de Araújo, e E.I. Shakhnovich. Description of atomic burials in compact globular proteins by fermi-dirac probability distributions. *Proteins: Structure, Function, and Bioinformatics*, 66(2):304–320, 2007.

[48] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[49] R.R. Rocha. Análise informacional dos enterramentos atômicos em proteínas globulares. Dissertação de Mestrado, Universidade de Brasília, 2012.

[50] B. Efron e R. Tibshirani. *An introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall, New York, 1993.

[51] C.E. Shannon. A mathematical theory of communication. *The Bell Technical Journal*, 27(4):379–423, 1948.

[52] T.M. Cover e J.A. Thomas. *Elements of information theory*. Wiley-interscience, 2006.

[53] M.A Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, Thompson J.D., Gibson T.J., e Higgins D.G. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007. Programa disponível em `http://www.clustal.org/clustal2/`.

[54] R.A. Abagyan e S. Batalov. Do aligned sequences share the same fold? *Journal of Molecular Biology*, 273(1):355–368, 1997.

[55] L.S. Boutemy, S.R.F. King, J. Win, R.K. Hughes, T.A. Clarke, T.M.A. Blumenschein, S. Kamoun, e M.J. Banfield. Structures of *Phytophthora* RXLR effector proteins a conserved but adaptable fold underpins functional diversity. *Journal of Biological Chemistry*, 286(41):35834–35842, 2011.

[56] S. Thoms, K.E.A. Max, M. Wunderlich, T. Jacso, H. Lilie, B. Reif, U. Heinemann, e F.X. Schmid. Dimer formation of a stabilized Gβ1 variant: a structural and energetic analysis. *Journal of Molecular Biology*, 391(5):918–932, 2009.

[57] U. Mueller, D. Perl, F.X. Schmid, e U. Heinemann. Thermal stability and atomic-resolution crystal structure of the *Bacillus caldolyticus* cold shock protein. *Journal of Molecular Biology*, 297(4):975–988, 2000.

[58] P.C. Whitford, O. Miyashita, Y. Levy, e J.N. Onuchic. Conformational transitions of adenylate kinase: switching by cracking. *Journal of Molecular Biology*, 366(5):1661–1671, 2007.

[59] A. Leach. Chapter 4: Empirical force field models: Molecular mechanics. In *Molecular Modelling: Principles and Applications (2nd Edition)*, páginas 165–247. Prentice Hall, 2 edition, April 2001.

[60] P.H. Hünenberger. Thermostat algorithms for molecular dynamics simulations. *Adv. Polym. Sci*, 173:105–149, 2005.

[61] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A.R.H.J. DiNola, e J.R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81:3684, 1984.

[62] Schrödinger, L.L.C. The PyMOL molecular graphics system, version 1.3r1. Disponível em `http://www.pymol.org/`, 2010.

[63] Zhao, Y. Brief introduction to the thermostats. Disponível em `http://www.math.ucsd.edu/~y1zhao/ResearchNotes/ResearchNote007Thermostat.pdf`, 2011.

[64] A.D. McLachlan. Rapid comparison of protein structures. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 38(6):871–873, 1982.

[65] Martin, A.C.R. and Porter, C.T. Profit, version 3.1. Disponível em `http://www.bioinf.org.uk/software/profit/`, 2009.

[66] P.N. Dodds e J.P. Rathjen. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nature Reviews Genetics*, 11(8):539–548, 2010.

[67] R.H.Y. Jiang, S. Tripathy, F. Govers, e B.M. Tyler. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proceedings of the National Academy of Sciences*, 105(12):4874–4879, 2008.

[68] L. Björck e G. Kronvall. Purification and some properties of streptococcal protein G, a novel IgG-binding reagent. *The Journal of Immunology*, 133(2):969–974, 1984.

[69] S. Kmiecik e A. Kolinski. Folding pathway of the B1 domain of protein G explored by multiscale modeling. *Biophysical Journal*, 94(3):726–736, 2008.

[70] G. Horn, R. Hofweber, W. Kremer, e H.R. Kalbitzer. Structure and function of bacterial cold shock proteins. *Cellular and Molecular Life Sciences*, 64(12):1457–1470, 2007.

[71] Tomas M.C. e Joy A.T. *Elements of Information Theory*, chapter 2. Wiley-Interscience, 2006.

BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 21 2012, pages 2755–2762
doi:10.1093/bioinformatics/bts512

Structural bioinformatics

Advance Access publication August 24, 2012

Information-theoretic analysis and prediction of protein atomic burials: on the search for an informational intermediate between sequence and structure

Juliana R. Rocha†, Marx G. van der Linden†, Diogo C. Ferreira, Paulo H. Azevêdo and Antônio F. Pereira de Araújo*

Laboratório de Biologia Teórica e Computacional, Departamento de Biologia Celular, Universidade de Brasília, Brasília-DF 70910-900, Brazil

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: It has been recently suggested that atomic burials, as expressed by molecular central distances, contain sufficient information to determine the tertiary structure of small globular proteins. A possible approach to structural determination from sequence could therefore involve a sequence-to-burial intermediate prediction step whose accuracy, however, is theoretically limited by the mutual information between these two variables. We use a non-redundant set of globular protein structures to estimate the mutual information between local amino acid sequence and atomic burials. Discretizing central distances of C_α or C_β atoms in equiprobable burial levels, we estimate relevant mutual information measures that are compared with actual predictions obtained from a Naive Bayesian Classifier (NBC) and a Hidden Markov Model (HMM).

Results: Mutual information density for 20 amino acids and two or three burial levels were estimated to be roughly 15% of the unconditional burial entropy density. Lower estimates for the mutual information between local amino acid sequence and burial of a single residue indicated an increase in mutual information with the number of burial levels up to at least five or six levels. Prediction schemes were found to efficiently extract the available burial information from local sequence. Lower estimates for the mutual information involving single burials are consistently approached by predictions from the NBC and actually surpassed by predictions from the HMM. Near-optimal prediction for the HMM is indicated by the agreement between its density of prediction information and the corresponding density of mutual information between input and output representations.

Availability: The dataset of protein structures and the prediction implementations are available at http://www.btc.unb.br/ (in 'Software').

Contact: aaraujo@unb.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 5, 2012; revised on July 31, 2012; accepted on August 13, 2012

1 INTRODUCTION

It has been a common statement in biology that amino acid sequences contain sufficient information to determine protein tertiary structures. Fulfilment of the implied possibility of structure prediction from sequence is actually considered one of the most important unsolved problems of molecular biophysics, as reviewed by different groups (Dill *et al.*, 2008; Onuchic and Wolynes, 2004; Shakhnovich, 2006). Such an intrinsically informational assertion, however, has only more recently been extensively investigated within the context of Shannon's information theory. Although informational concepts have been used in algorithms for secondary structure prediction from local sequence since the 70s (Garnier *et al.*, 1978), for example, the limit imposed on prediction by the mutual information between these two quantities was estimated only a few years ago (Crooks and Brenner, 2004). Incidentally, an informational analysis of backbone dihedral angles has also exposed the unfeasibility of tertiary structure determination from an even perfect three-state secondary structure prediction (Solis and Rackovsky, 2004). The recurrent utilization of statistical potentials in computational biology has also been interpreted explicitly in informational terms (Solis and Rackovsky, 2007). A particularly relevant example is the analysis of pairwise contact potentials, which revealed a surprisingly modest mutual information between contact partners (Cline *et al.*, 2002; Crooks *et al.*, 2004). General distance constraints have also been investigated, at least in the context of minimalist protein models (Sullivan *et al.*, 2003).

Contrasting with secondary structure, atomic burials appear to encode sufficient information for structural determination. Contrasting with pairwise contacts, they have a much better chance of being adequately estimated from sequence information. Monte Carlo simulations of geometrically realistic protein models using native burial information, as expressed by atomic distances from the molecular center, have successfully recovered the tertiary structure of small globular proteins (Pereira de Araújo *et al.*, 2008). A simple computational experiment combining Molecular Dynamics of similar models with discretized burial levels has additionally provided an upper bound for the amount of required burial information. It actually turned out to be comparable to, and therefore encodable by, the information (entropy) of local protein sequences (Pereira de Araújo and Onuchic, 2009). The observed discriminatory difference between

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

burial and secondary structure representations does not arise therefore from a trivial difference in precision. A very precise representation of all backbone dihedral angles can clearly encode tertiary structures, even using a small amount of information, or number of letters, in α-helical regions and possibly β-strands, but requiring a large, sequence-incompatible, number of letters in intervening loops. The distinction appears to be more basic and related to different types of information encoded in the two local representations. While secondary structure is a local representation of purely local structure, burials include global structural information in a local representation, as is evident from the fact that the whole tertiary structure is required for determination of burials, but not secondary structure, of any short fragment of amino acids.

The possibility of structural determination from sequence-dependent burial information, when combined to appropriate sequence-independent constraints, is consistent with the perceptible previous success in native fold recognition from the arrangement of hydrophobic and polar residues (Huang *et al.*, 1995). It has also been further supported recently by a purely analytical model which was able to recover native-like burial traces from sequence hydrophobicity information combined to simple constraints on chain connectivity and overall globular size (England, 2011). A potential approach to tertiary structure prediction could therefore involve a sequence-to-burial intermediate prediction step. It must be noted that theoretical encodability, as provided by entropy compatibility, is necessary but not sufficient to demonstrate actual encoding. The accuracy of any burial prediction from sequence must be further limited by the observed correlation between burials and sequences, as conveniently quantified by the mutual information between these two quantities. In this study, we estimate the mutual information between burials and local amino acid sequence in globular proteins. The resulting fraction of sequence entropy actually involved in burial encoding provides theoretical limits to which prediction algorithms should be compared. We additionally investigate the efficiency of simple statistical prediction schemes, namely, a Naive Bayesian Classifier (NBC) and a Hidden Markov Model (HMM), in extracting the available burial information from local sequence.

2 METHODS

In this study, we estimated probabilities from frequencies observed in a dataset of representative globular structures derived from PDBSELECT (Hobohm and Sander, 1994). From the list made available in November 2009, we selected structures determined by X-ray crystallography with resolution better than 2.5 Å and excluded chains not satisfying the globularity criterion given by the expected relation between radius of gyration and the number of residues, $R_g \leq 2.9N_r^{1/3}$ Å (Gomes *et al.*, 2007). Membrane proteins were also excluded, simply by removing PDB files containing the word 'MEMBRANE'. The resulting collection, from now on simply referred to as the databank, is composed of 1499 chains, with a total of ~263 000 residues. Statistical errors on computed probabilities and entropies were estimated, and systematic biases corrected for, by a bootstrap procedure using 50 randomly generated replicas of the databank (Crooks and Brenner, 2004; Efron and Tibshirani, 1993). In addition to the complete alphabet of 20 amino acid identities, we have also used the reduced alphabets HP and HPN. Hydrophobic and polar residues were grouped in the HP alphabet as H = {A, C, F, G, I, L, M, V, W, Y} and P = {D, E, H, K, N, P, Q, R, S, T}, respectively. In HPN a third, 'neutral', class includes residues from both HP groups, N = {A, G, H, S, T}. Burials, b, were obtained from the atomic distances from the molecular center, r, of C_α or C_β atoms, normalized by the radius of gyration, R_g, or $b = r/R_g$, and grouped in approximately equiprobable burial levels, resulting in a collection of burial alphabets $\{\chi L\}$, where χ is either α or β, representing the atomic type for which burials are defined, and L is the number of burial layers. Cutoff burial values for different burial levels were obtained from the estimated burial distribution obtained by Gomes *et al.* (2007). We usually use superscripts to indicate block size and integer subscripts to indicate position within the block, with '0' representing the central block position by convention. If necessary, however, we also indicate particular alphabets as subscripts in our notation, such as $H(Q^N_{\mathrm{HP}})$, $h(B_{\beta 5})$, $I(Q^N_{20}; B^N_{\alpha 2})$.

N-block entropies for residue identities, $H(Q^N)$, and burials, $H(B^N)$, were computed according to Shannon's basic equation

$$H(X^N) = -\sum_{x^N} p(x^N)\log_2 p(x^N),$$

where the sum is over all blocks of N adjacent letters, x^N, either identities or burials, and probabilities are estimated from corresponding frequencies in the databank. A linear dependence of the estimated entropy on block size in the range $m < N < m'$,

$$H(X^N) = Nh(X) + E_X. \tag{1}$$

is consistent with a Markovian process of order m, where $h(X)$ is the entropy density and E_X is the N-independent excess entropy, which indicates the uncertainty resolved by local correlations. Deviation from linearity for $N < m$ arises from these local correlations between letters while for $N > m'$ frequencies in the databank become poor estimates for actual probabilities and the estimated entropy converges to an alphabet-independent value that depends on the overall size of the databank, a situation we refer to as 'saturation'. Estimates for $h(X)$ and E_X can therefore be obtained from the observed dependence of $H(X^N)$ on N if the order of the underlying Markov process is sufficiently small and the dataset is sufficiently large so that $m \ll m'$ and the linear region can be clearly identified.

For the mutual information between blocks of identities and burials, Q^N and B^N, a limiting linear behavior is also expected, or

$$I(Q^N; B^N) = H(Q^N) - H(Q^N|B^N) = Ni(Q; B) + E_{Q;B}. \tag{2}$$

and an estimate for the corresponding mutual information density, $i(Q; B)$, a quantity of much interest that imposes an upper limit on any possible prediction of the local sequence of burials from the local sequence of identities, could again be obtained from N-block entropy estimates. In this case, however, because the number of different blocks increases more sharply with block size, saturation should occur at a much shorter block length. We use therefore an approximation,

$$i(Q; B) \approx \lim_{N\to\infty} I(Q_0; B^N) \equiv I(Q_0; B^\infty). \tag{3}$$

that is valid when the letters in one of the sequences are statistically independent both unconditionally and conditionally to the other sequence, as it turns out to be the case for identities with respect to burials. The density of mutual information is estimated accordingly by extrapolation of the dependence on N of $I(Q_0; B^N)$, the mutual information between N-blocks of burials, B^N, and the identity of the central residue in the block, Q_0,

$$I(Q_0; B^N) = H(Q_0) - H(Q_0|B^N). \tag{4}$$

where $H(Q_0)$ is the single identity entropy, obtained with probabilities estimated directly from corresponding frequencies, and $H(Q_0|B^N)$ is the conditional entropy of central residue identity conditional to burial block. This procedure was used by Crooks and Brenner (2004) to estimate the mutual information density between sequences of amino acid residues and corresponding sequences of secondary structure assignments.

Underlying conditional probabilities were obtained from corresponding frequencies, or $p(Q_0|B^N) = n(Q_0, B^N)/n(B^N)$, only for the HP alphabet, since statistics turned out to be sufficient. For the other alphabets conditional probabilities were estimated as

$$p(Q_0|B^N) = \frac{n(Q_0, B^N) + (20 \times p(Q_0|B_0))}{n(B^N) + 20}, \quad (5)$$

using 20 'pseudo-counts' with prior probability $p(Q_0|B_0)$ in an attempt to minimize artifacts from low-frequency events. Due to pseudo-counts, the estimated mutual information turns out to be increasingly smaller than its actual value as N becomes large. While the actual mutual information must increase monotonically with N, its estimate will decrease for large N, providing again a simple signature of databank saturation. For the HP alphabet, pseudo-counts were not used and saturation manifests itself as an abrupt increase in estimated mutual information causing an upward inflection in the estimated curve. Data points were fitted, before saturation, to a single exponential $f(x) = a - b\exp(-x/c)$, with limiting behavior provided by adjusted parameter a, or to a symmetrically inflected sigmoid $f(x) = \frac{a}{(1-\exp(-b(x-c)))} + d$, with limiting behavior provided by $a+d$. Fitting to an asymmetric Gompertz function provided similar estimates but with larger errors, reflecting the larger number of adjustable parameters (not shown).

In addition to $I(Q_0; B^\infty)$, we are also interested in the converse quantity, $I(Q^\infty; B_0)$, since it provides a limit for the prediction of individual burial values given the local sequence of identities. Saturation might again become a problem for large alphabets of identities, in which case it is useful to consider the following lower bound:

$$\begin{aligned}\sum_{i=1}^{N} I(Q_i; B_0) &= \sum_{i=1}^{N} [H(Q_i) - H(Q_i|B_0)] \le H(Q^N) - H(Q^N|B_0) \\ &= I(Q^N; B_0),\end{aligned} \quad (6)$$

with limiting behavior

$$I(Q^\infty; B_0)^- \equiv \lim_{N\to\infty} \sum_{i=1}^{N} I(Q_i; B_0) \le I(Q^\infty; B_0) \quad (7)$$

Each of the N 'positional' mutual information terms between Q_i and B_0 is computed from the same number of possible combinations, independently of N. The results for the tractable HP alphabet and two burial levels, shown in the Supplementary Information, indicate that Equation (3) is indeed a good approximation while a strict inequality is expected in Equation (7).

In order to compare our mutual information estimates with actual predictions, we implemented two simple statistical schemes for predicting discrete atomic burial levels from amino acid sequence in globular proteins: a NBC and a HMM. Both methods are supervised learning algorithms, i.e. they employ a learning step, in which they gather data from a training set to generate some statistical model, followed by a prediction step, in which they use the model to predict new data. We have used the same dataset of structures as for the informational analysis, now randomly divided in training and testing subsets. Statistical errors and biases were again estimated by bootstrapping resampling with 50 replicas. While the NBC estimates the probability for different burial levels of a given residue simply from a local 'window' of identities in the primary sequence, neglecting most correlations between adjacent residues, the HMM considers explicitly the correlations between 'fragments' of hidden variables, including burials, which are modeled as producing the observed primary sequence. Both algorithms are described in detail in the Supplementary Information, as well as the procedures to obtain the corresponding prediction information, I_p, and prediction information densities, i_p, to be compared with the mutual information estimates $I(Q^\infty; B)^-$ and $i(Q; B)$, respectively.

3 RESULTS

Figure 1 illustrates the statistical behavior of local sequences of C_β burials, as determined from central distances normalized by radius of gyration. N-block entropy is shown as a function of block size N. Different curves correspond to different alphabets, ranging from two to five equally probable burial levels. Deviation from linearity for large N results from saturation of the databank as all curves converge to the same alphabet-independent saturated limit behavior. Deviation from linearity for small N and, more perceptively, a positive intercept with the ordinate axis reflect the expected local correlations between adjacent burial levels. These results suggest a low-order markovicity, with m not higher than 2 or 3. Analogous results for C_α burials, shown in the Supplementary Information, indicate a qualitatively similar behavior. For identities, on the other hand, as also shown in the Supplementary Information, it is apparent that $H(Q^N)$ increases linearly from the origin for all alphabets, being consistent with zero-order markovicity, $m = 0$, or equivalently, statistical independence between amino acid identities along the sequence. Accordingly, as shown in Table 1, residue entropy density $h(Q)$ is very close to the single letter entropy, $H(Q^1)$, increasing from essentially 1 for HP sequences, $h(Q_{HP}) \approx H(Q^1_{HP}) \approx 1$ bit/residue, to $h(Q_{20}) \approx H(Q^1_{20}) \approx 4.18$ bits/residue for 20 amino acid letters while mutual information between adjacent identities is close to zero. Entropy densities of correlated burials, however, are significantly lower than corresponding single burial entropies, with a positive mutual information between adjacent burials, such as $h(B_{\alpha 2}) \approx 0.62 < H(B_{\alpha 2}) \approx 1$ bit/residue and $I(B_i; B_{i+1}) \approx 0.34$ bit for two C_α burial levels. C_β burials consistently display larger entropy densities, such as $h(B_{\beta 2}) \approx 0.73$ and $h(B_{\beta 3}) \approx 1.1$ bits/residue for two and three burial levels, respectively, to be compared with $h(B_{\alpha 2}) \approx 0.62$ and $h(B_{\alpha 3}) \approx 0.95$ bit/residue for C_α burials.

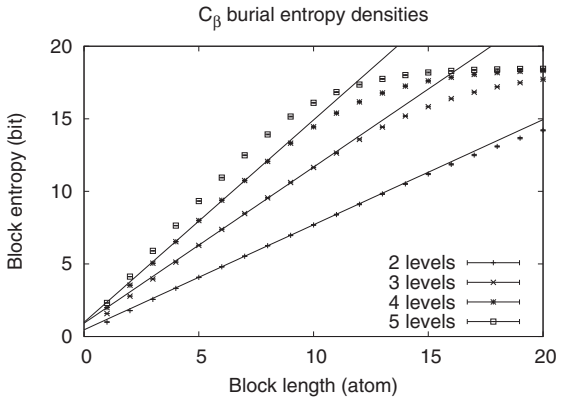


Fig. 1. N-block sequence entropy estimates as a function of block size N for different alphabets of C_β burial levels. Both the entropy density (inclination) and excess entropy (intersect with the ordinates) are obtained from straight lines fitted to the linear region, which is clearly identified for $L = 2$ and $L = 3$. Deviation from linearity for small N is indicative of local correlations while deviation at large N is due to databank saturation. Analogous results for amino acid identities and C_α burials are shown in the Supplementary Information

Table 1. Single sequence analysis

	$H(X)$	$I(X_i; X_{i+1})$	$h(X)$	E_X
HP	1.00000(9)	0.00072(8)	0.9969(2)	0.0096(7)
HPN	1.5806(4)	0.0009(1)	1.5734(7)	0.018(3)
20	4.185(2)	0.005(7)	4.176(4)	0.010(5)
$\alpha 2$	0.99974(6)	0.342(3)	0.619(1)	0.513(9)
$\alpha 3$	1.5796(3)	0.574(3)	0.933(4)	0.91(2)
$\beta 2$	0.9988(1)	0.211(3)	0.724(2)	0.46(2)
$\beta 3$	1.5804(2)	0.377(4)	1.075(6)	0.92(4)

Letter entropy, $H(X)$, and mutual information between adjacent letters, $I(X_i; X_{i+1})$, are in bits. Entropy density $h(X)$, in bits/letter, and corresponding excess entropy E_X, in bits, were obtained from data fits shown in Figure 1 or in the Supplementary Information. Each line corresponds to a different alphabet of amino acid identities or burials, as indicated in the first column. Error in the last significant digit is shown in parentheses.

The dependence on N of the estimates for mutual information between N-blocks of burials and central residue identities, $I(Q_0; B^N)$, is shown in Figure 2 for two and three levels of C_β burials. Analogous results for C_α burials are shown in the Supplementary Information. Mutual information density, $i(Q; B) \approx I(Q_0; B^\infty)$, was obtained by extrapolation from exponential or sigmoidal fits to the points before saturation, as indicated by solid lines and shown in Table 2. Mutual information density is always larger for C_β burials when compared with C_α burials with the same alphabet combination, such as $i(Q_{20}; B_{\alpha 2}) \approx 0.09 < i(Q_{20}; B_{\beta 2}) \approx 1.13$ bits/residue. As could be anticipated, it tends to increase with alphabet size either of amino acid identities or burials such as, in the case of C_β atoms, from $i(Q_{HP}; B_{\beta 2}) \approx 0.07$ bit/residue for the HP alphabet and $L = 2$ burial layers, to $i(Q_{20}; B_{\beta 3}) \approx 0.18$ bit/residue, for 20 amino acid letters and $L = 3$ layers. Databank saturation prevented reliable density estimates for $L > 3$.

Positional mutual information values, $I(Q_i; B_0)$, are shown in Figure 3a for 20 amino acid letters and different numbers of burial levels of C_β atoms. Positional mutual information is essentially 0 for burial and identity pairs separated by more than 15 residues. We therefore use the sum $\sum_{i=1}^{N} I(Q_i; B_0)$ with $N = 31$ as a reasonable approximation of $I(Q^\infty; B_0)^- \equiv \sum_{i=1}^{\infty} I(Q_i; B_0)$ which, as indicated in the Supplementary Information, is expected to be a lower bound for $I(Q^\infty; B_0)$. We were also able to explore the effect of many burial levels on $I(Q^\infty; B_0)^-$. As shown in Figure 3b, $I(Q^\infty; B_0)^-$ for C_β increases significantly from two layers to five layers, approximately from 0.13 to 0.18 bit, but only slightly for additional layers with asymptotic limit close to 0.2 bit. Qualitatively similar results were obtained for C_α atoms but mutual information between single burials and local sequence tends again to be smaller in this case when compared with C_β atoms, although the difference is smaller than for mutual information density, as also seen in Table 2. We also show for comparison in the same table the mutual information between single letters, $I(Q; B)$.

The performance of two-layer C_β burial prediction is summarized in Figure 4. Analogous results for C_α burials are shown in the Supplementary Information. Prediction accuracy, A (a,b), and prediction information, I_p (c,d), as determined by

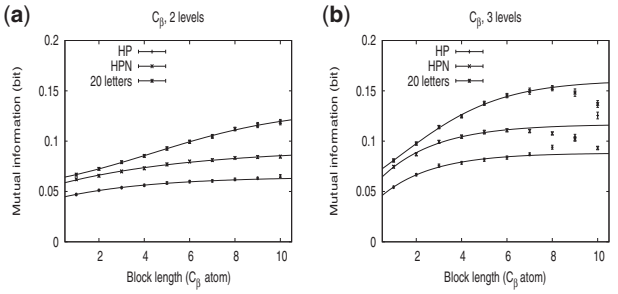


Fig. 2. Estimates for the mutual information, $I(Q_0; B^N)$, between a single central amino acid identity, Q_0, and N-blocks of burials, B^N, as a function of block size N, for two (**a**) and three (**b**) levels of C_β burials. Different sets of points correspond to different alphabets of amino acid identities. Lines represent exponential or sigmoidal fits to the data before saturation from which limiting values $i(Q; B) \approx I(Q_0; B^\infty)$ are obtained. Saturation for $L = 2$ occurs at $N \approx 11$ and is not perceived in the displayed range, while for $L = 3$ it occurs $N \approx 8$, as observed in (b). Analogous results for C_α burials are shown in the Supplementary Information

Table 2. Inter-sequence analysis

L		$I(Q; B)$		$i(Q; B)$		$I(Q^\infty; B_0)^-$	
		C_α	C_β	C_α	C_β	C_α	C_β
2	HP	0.0297(6)	0.0472(9)	0.050(3)	0.068(2)	0.059(3)	0.070(3)
	HPN	0.0420(9)	0.062(1)	0.068(3)	0.092(2)	0.089(7)	0.100(4)
	20	0.046(1)	0.067(1)	0.091(7)	0.13(1)	0.113(5)	0.124(5)
3	HP	0.0357(9)	0.054(1)	0.066(4)	0.088(2)	0.075(4)	0.086(4)
	HPN	0.051(1)	0.075(1)	0.091(4)	0.117(3)	0.114(5)	0.125(5)
	20	0.0570(9)	0.081(2)	0.130(6)	0.176(6)	0.149(7)	0.159(6)

Mutual information between single letters, $I(Q; B)$ in bits, mutual information density, $i(Q; B)$ in bits/pair, as obtained in Figure 2 and Supplementary Information, and the lower estimate for the mutual information between single burial and local sequence of identities, $I(Q^\infty; B_0)^-$, as obtained in Figure 3, for C_α and C_β atoms are shown for different combinations of identity alphabet and number of burial layers, as indicated in the first two columns. Error in the last significant digit is shown in parentheses.

Equations (S9) and (S10) of the Supplementary Information, are plotted as a function of window size for the NBC (a,c) and as a function of fragment size for the HMM (b,d). In addition to the complete alphabet of 20 amino acids, tests were also performed using the HP and HPN-reduced alphabets. For the NBC, we report results for the simpler variation provided by Equation (S4) of the Supplementary Information, NBC1 (non-shaded symbols), and also for the variation using positional probabilities conditional to central residue identity, as provided by Equation (S5) of the Supplementary Information, NBC2 (shaded symbols). Both accuracy and information increase significantly as the window grows from one to nine residues, but not perceptibly for longer windows. Overall performance is higher for C_β than for C_α atoms. For 20 amino acids, accuracy increases from ~61% to above 65% for C_α atoms and from around 63% to above 66% for C_β. These few percentage points in accuracy improvement actually correspond to around 100% increase in prediction information, from around 4 to above 10 centibits and

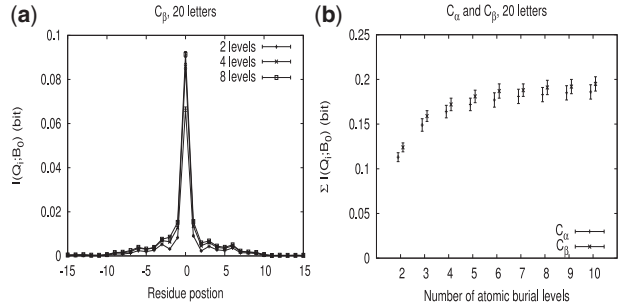


Fig. 3. Positional mutual information $I(Q_i; B_0)$ between amino acid identity at position i, Q_i, within the N-block of identities Q^N, and central C_β burial, B_0, for 20 amino acid letters and various numbers of burial levels (**a**) and limiting behavior for the sum of positional mutual information terms, obtained with fixed block size $N = 31$, as a function of the number of burial levels for C_α and C_β atoms (**b**)

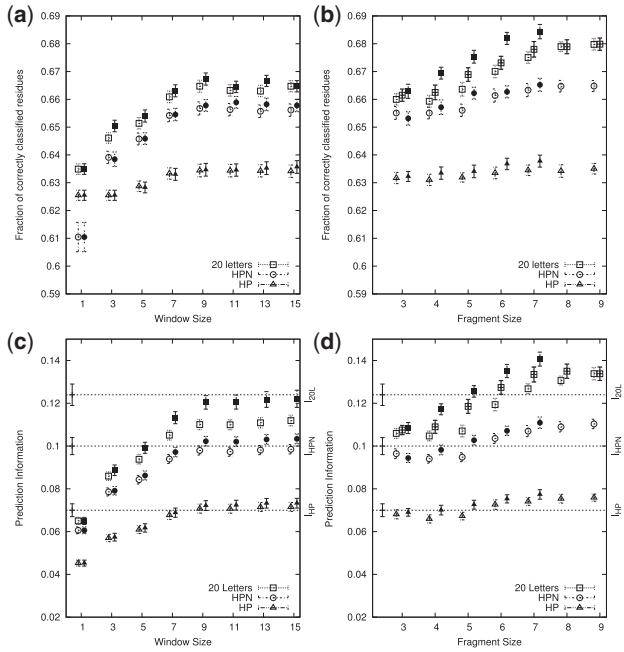


Fig. 4. Prediction accuracy A (**a**,**b**) and prediction information I_p (**c**,**d**) for two levels of C_β burials with different identity alphabets. Plots in the first column (a,c) show results for NBC predictions; the second column (b,d) refers to the HMM results. The NBC method is bounded, within error, to the limits established by corresponding $I(Q^\infty; B_0)^-$ estimates (dotted horizontal lines), while the same limits are surpassed by the HMM method (d). In all plots, unshaded symbols represent the simplest version of each algorithm (NBC1 or HMM with nothing but burial levels encoded into the hidden variables) and shaded symbols represent improved versions (NBC2 or HMM with secondary structures). For HMM, half-shaded symbols represent the version that used side-chain orientations

from around 6 to above 11 centibits for C_α and C_β, respectively, for NBC1. Further improvement provided by NBC2, although hardly perceptible in the accuracy measure, is consistently observed for prediction information, accounting for more than 1 centibit of additional information for 20 amino acids while sampling error is of the order of millibits. For the HP and HPN alphabets, both NBC1 and NBC2 predictions agree, within sampling error, with the corresponding lower limits provided by $I(Q^\infty; B_0)^-$ while for 20 amino acids this is the case for NBC2.

For the HMM, tested fragment lengths ranged from 3 to 9, but some configurations could not be tested due to hardware constraints related to computer memory usage with many hidden variables. It is clear in the plots of Figure 4b and d that the fragment length has a direct correlation with the quality of results for HMM prediction, especially when the full 20-letter alphabet is used to represent amino acid sequences. The connections between burial levels and secondary structures (shaded symbols) and between burial levels and two possible side chain orientations (obtained from the comparison between C_β and C_α burials and represented as half-shaded symbols) were also investigated by incorporating the corresponding hidden variables into the HMM states. Both approaches were successful in improving the prediction of burial levels, and the usage of secondary structures was slightly more effective than that of side-chain orientations. Incidentally, it was found that not only the prediction accuracy of burial levels but also that of secondary structures is improved when both features are considered together (data not shown). Our most accurate results for burial prediction were around 67.5 and 68.5% of correctly classified residues, respectively, for C_α and C_β. Corresponding prediction information values of ~0.13 and 0.14 bit are higher than the lower limits provided by $I(Q^\infty; B_0)^-$, as was consistently observed for the HMM algorithm, particularly with the configurations that employed additional descriptors to the hidden variables and fragment sizes of at least six to seven residues. As with the NBC, prediction of C_β was generally better than that of C_α.

Since the HMM algorithm works with relative probabilities of fragments of burial levels, it is meaningful to estimate the *density* of prediction information, i_p, according to Equation (S12) of the Supplementary Information, i.e. the amount of new prediction information discovered for each new residue once the previous burials have already been established. Figure 5 shows $h_N(B|B(Q))$ (a) Equation (S14) of the Supplementary Information, for the various HMM prediction schemes for C_β burials, as well as corresponding values of $h_N(B)$, Equation (S13) of the Supplementary Information, computed from block entropies shown in Figure 1. The difference between these quantities is the estimate for the prediction information density, i_p, Equation (S12) of the Supplementary Information, which is shown in (b). Our results can be compared with the corresponding estimates for the mutual information density between sequences and burials, $i(B; Q)$, from Table 2, also displayed in (b) as dotted horizontal lines, which should act as effective upper limits on prediction quality. Analogous results for C_α burials are shown in the Supplementary Information. Since i_p for $N \geq 7$ agrees within sampling error with $i(B; Q)$, it is suggested that our best overall results for two burial levels are extracting virtually all of the burial information that is available in local sequences.

Figure 6 compares the prediction information achieved when the NBC and HMM methods are applied to predict discrete C_β burials into more than two layers. Analogous results for C_α are shown in the Supplementary Information. In all cases, it is clear that the quality of prediction is improved when the number of

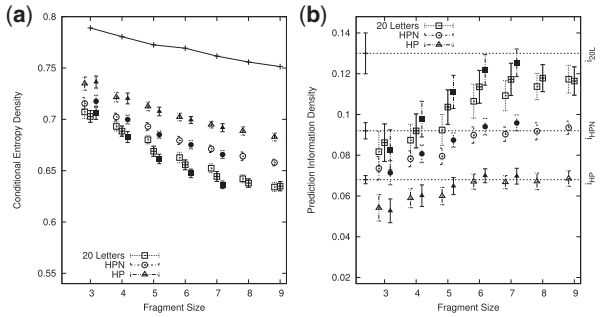


Fig. 5. For HMM results, the *density* of prediction information, i_p, can be calculated as the difference between an N-dependent estimate for the entropy density of burial levels, $h_N(B)$, Equation (S13) of the Supplementary Information (shown as a solid line in **a**), and an analogous estimate for the entropy density conditional to prediction, $h_N(B|B(Q))$, Equation (S14) of the Supplementary Information (shown as points in a). Resulting differences are plotted in (**b**) in comparison to the upper limit provided by the observed existing mutual information density between burials and sequences, $i(B; Q)$ (horizontal dashed lines). The results for C_β predictions are shown here. Analogous results for C_α predictions are shown in the Supplementary Information. Point symbols are encoded similarly to Figure 4

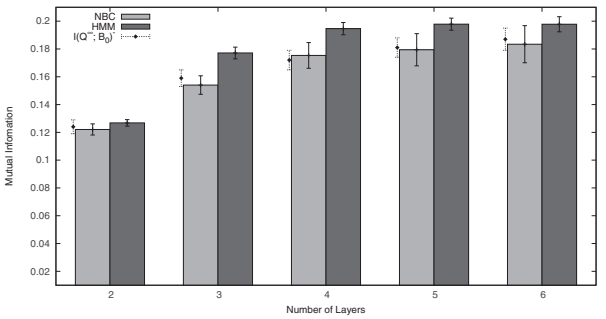


Fig. 6. As the number of discrete burial layers increases, the quality of prediction, as measured by the prediction information, I_p, also improves, at least up to four to five layers. The results are shown for NBC2 and HMM C_β predictions. Analogous results for C_α are shown in the Supplementary Information. Window size of 15 and fragment size of 7 were used for NBC and HMM, respectively. HMM predictions were performed with no additional descriptors to the hidden variables. Dotted error bars represent the estimated lower bounds for the mutual information between single burial and sequence of identities, $I(Q^\infty; B_0)^-$

layers is increased up to a number of 4. The rise in quality for five or six layers, however, is less significant, suggesting an upper limit for the number of layers into which it is useful to split a protein for burial-level prediction. As already observed for two burial layers, prediction information tends to be larger for C_β when compared with C_α atoms. Furthermore, $I(Q^\infty; B_0)^-$ values are also approached by NBC and surpassed by HMM predictions.

4 DISCUSSION

In this study, we estimate by extrapolation, neglecting long range correlations, the mutual information density between local sequence of amino acid identities and corresponding burials, $i(Q; B) \approx I(Q_0; B^\infty)$. It must be noted that the underlying probability distributions, estimated from local block statistics, are much simpler than distributions of whole amino acid sequences and tertiary structures. In particular, they are consistent with markovicity and a linear dependence of entropy, and mutual information, on block length, as shown in Figure 1. Meaningful densities of entropy and mutual information can be estimated for this simplified statistical scheme with different reduced alphabets. Additionally, and most importantly, resulting estimates for $i(Q; B)$ provide upper limits for the quality of prediction associating local sequences of burials and identities, a clearly attemptable task with established learning algorithms. Prediction of single burial values from local sequence, on the other hand, should be limited simply by the mutual information between local sequence and single burial, $I(Q^\infty; B_0)$, which is difficult to estimate for 20 amino acid letters due to databank saturation. We provide therefore a lower bound, $I(Q^\infty; B_0)^- < I(Q^\infty; B_0)$, further neglecting local correlations between amino acid identities conditional to single central burial. For the tractable HP alphabet, the difference between $I(Q^\infty; B_0)$ and $I(Q^\infty; B_0)^-$ is a single centibit, as shown in the Supplementary Information.

Single sequence statistical behavior, as summarized in Table 1, is qualitatively similar to what was previously observed for secondary structure by Crooks and Brenner (2004). While amino acid identities in local sequences appear to be statistically independent, short-range correlations are detected for the one-dimensional structural descriptor, either secondary structure or burial. Correlations between burials are stronger for C_α than for C_β atoms, as evidenced by smaller entropy density and larger mutual information between adjacent letters in the first case. This observation is likely to be at least partly associated to a longer distance along the sequence between adjacent C_β when compared with C_α atoms. As shown in Table 2, local sequence appears to be more informative about C_β than C_α burials, as indicated by larger values of $I(Q^\infty; B_0)^-$ and $i(Q; B)$ in the first case. Nevertheless, the proportional contribution to mutual information from local sequence beyond single residue identity appears to be larger for C_α when compared with C_β, as suggested by larger values for $I(Q^\infty; B_0)^-/I(Q; B)$ for the backbone atom.

Our estimates for the mutual information density, $i(Q; B)$, indicate that the uncertainty about burials that is resolvable from local sequence, already considering the reduction provided by sequence-independent burial local correlations, can be as small as 9 centibits/residue, as for two levels of C_α burials, and also at least as large as 18 centibits/residue, observed for three levels of C_β burials. These values are comparable to estimates involving secondary structure (16 centibits/residue; Crooks and Brenner, 2004), and are around 15% of the corresponding burial entropy density. Estimates for $i(Q; B)$ tend to be larger than for corresponding estimates for $I(Q^\infty; B_0)^-$, particularly for C_α atoms, in which case the difference is consistently between 1 and 2 centibits. It is suggested, therefore, that a couple of centibits of extra burial information might be extracted from sequences, in this case, when local burial correlations are accounted for. The effect on C_β atoms is smaller, again indicating a milder dependence of burial behavior from the side-chain atom on adjacent residues, either through their identities or burials.

Presently investigated burial levels, defined by equiprobable layers of central distances, display some qualitative similarity with burial levels defined from accessible surface areas, as reported by Crooks *et al.* (2004). Oscillations in positional mutual information observed in Figure 3, reflecting secondary structure exposure periodicity, are also observed in analogous plots involving burials in that previous investigation, although not for identities or secondary structure assignments. Notably, however, single amino acid identities appear to be more informative about accessible surfaces than about central distances. While single residue mutual information between identity and two bins of burials reported by Crooks *et al.* (2004), is 0.15 bit, our presently estimated value for $I(Q_{20}; B_{\beta 2})$ is only 0.07 bit, or about half of the corresponding density, $i(Q_{20}; B_{\beta 2})$, as shown in Table 2. Correlations between adjacent central distances, on the other hand, appear to be larger, as shown by larger values of $I(B_i; B_{i+1})$ in Table 1 when compared with values reported in that previous investigation.

These discrepancies might be partly associated to different procedures for determination of burial levels. While levels of accessible surfaces were explicitly determined from mutual information maximization, our levels of central distances simply maximize unconditional uncertainty. It is possible, nevertheless, that intrinsic physical differences between the two measures are also involved. Although correlated in globular proteins (Pereira de Araújo *et al.*, 2008), it is apparent that accessible surface area should be affected more directly by residue hydrophobicity while being somewhat less dependent on adjacent residues. It is not presently clear how much information could be expected from actual predictions of accessible areas from local sequence, since mutual information densities have not been reported. Weaker correlations when compared with central distances, however, are indicative of a less pronounced increase in prediction information with additional local environment beyond single residue. In any case, even if eventually more predictable than central distances, it remains to be shown if accessible areas can be as efficient in tertiary structure determination.

Our prediction results indicate that most of the burial information shared by local sequences is easily captured by simple statistical prediction schemes based on HMM or, to a lesser extent, NBC. Interestingly, $I(Q^{\infty}; B_0)^{-}$ is approached by the NBC algorithm, which neglects most identity correlations conditional to single burials, and actually surpassed by the HMM algorithm, which appropriately accounts for such correlations. Furthermore, near-optimal prediction for HMM algorithms is indicated by the corresponding mutual information density approaching our present estimate for $i(Q; B)$. From the results with reduced identity alphabets, it is apparent that only about half of the burial information extractable from local sequence using all 20 amino acid letters is still extractable when the HP-reduced alphabet is used instead. The significant improvement provided by the HPN alphabet, with just a single additional letter, indicates however that judiciously chosen reduced alphabets might still be useful in actual prediction, particularly in situations in which the size of the training set might become a limiting factor. In the opposite situation, when the training set is sufficiently large, prediction could be improved by increasing the number of burial levels, as indicated by Figure 6, or by including more hidden variables in the HMM.

In any case, independently of the size of the databank, burial prediction information is unavoidably restricted within a small fraction of the unconditional burial uncertainty, as provided by the density of mutual information between identities and burials, $i(Q; B)$. Even considering the possibility of judicious partitioning of the databank, such as according to chain size or structural class, the basic situation is unlikely to change significantly. As has been previously noted (Crooks and Brenner, 2004), a small amount of mutual information between local sequence and structural descriptors, when compared with the descriptor entropy density, indicates that local structure, as reflected in secondary structure or burials, must be largely determined by non-local information. It is useful, however, to distinguish between sequence-dependent and sequence-independent non-local information. After all, a large amount of structure-determining information is provided by sequence-independent constraints, analogous to grammatical rules of human languages (Pereira de Araújo and Onuchic, 2009). The information to be obtained from sequences, corresponding in the same analogy to the actual literature codified in written texts, should actually be much smaller. The distinction between sequence-dependent and sequence independent information is already apparent locally. The uncertainty of 1 bit for two burial levels of a single C_α atom, for example, diminishes to 0.6 bit due to sequence-independent local information, or a reduction of 0.4 bit, while around 0.1 bit is resolvable by sequence-dependent local information. A particularly interesting possibility, from the predictor's perspective, would correspond to sufficient sequence-dependent information for tertiary structure determination being exclusively local, while non-local information would be sequence-independent.

A large amount of sequence-independent non-local structural information is actually inferred from the small expected total number of protein shapes, Ω_s, which has been estimated by different groups to be in the order of several thousands (Chotia, 1992; Govindarajan *et al.*, 1999; Koonin *et al.*, 2002; Zhang and DeLisi, 1998). If Ω_s is assumed to be 10 000, for example, the corresponding entropy would be limited from above by $\log_2 \Omega_s$ and could not be more than around 13 bits per structure, or only 0.05 bit/residue for a putative typical length of 260 residues (0.1 bit/residue for 130 residues). This would be the uncertainty about whole structures, and therefore burials, to be resolved from sequence. The large remaining single burial uncertainty, e.g. $\approx (1 - 005 = 095)$ bits/residue for two C_α burial levels, must therefore be resolvable by sequence-independent information, both local (≈ 0.4 bits/residue, as discussed above) and non-local (≈ 0.55 bits/residue, as a consequence). Note that even if the total effective number of structures turns out to be larger or smaller by up to two orders of magnitude, the estimated amount of sequence-dependent structural information could not change by more than a couple of centibits/residue. It is interesting that an independent argument, based the thermodynamic stability of globular proteins, provided a compatible entropy estimate, $\approx$ 10–30 bits per macromolecule (Crooks *et al.*, 2004).

This small amount of sequence-dependent information (literature), when compared with the large amount of sequence-independent constraints (grammar), is an unavoidable consequence of a modest total number of structures when compared with possible sequences. It is also clearly consistent with the sound elusiveness of possible solutions for the problem

of *ab initio* protein structure prediction, contrasting to significant success in homology modeling. Note that the entropy of whole amino acid sequences must indeed be much larger than structural entropy since many sequences fold to each single structure (Koehl and Levitt, 2002; Larson *et al.*, 2002), although smaller, and less trivial, than estimated from local statistics. Long-range sequence correlations have been detected (Pande *et al.*, 1994) and must produce deviations from markovicity, contributing not only to reduce the entropy but also to destroy its linear dependence on chain length. Crucially, in any case, the presently reported small information for burial predictions can still turn out to be sufficient for structural determination when combined to appropriate sequence-independent constraints.

5 CONCLUSION

Knowledge about atomic burial levels has been previously shown to be both sufficient for structural determination of small globular proteins and entropically compatible with amino acid sequences. Our present results, however, indicate that only a fraction around 15%, at least for C_α and C_β atoms, of burial uncertainty is resolvable by local amino acid sequence. On the bright side, most of this sequence-dependent burial information is easily extractable by simple prediction schemes, such as the presently implemented NBC and HMM. Most importantly, these predictions provide parameters for future folding simulations completely independent of knowledge about the native structure. The possibility of structural prediction of globular proteins from amino acid sequence using atomic burials as informational intermediates, including a possible combined improvement of sequence-independent constraints and burial prediction schemes, can now be investigated directly.

Funding: This research was supported by the Conselho Nacional de Pesquisa (CNPq), grant 478121/2011-3. J.R.R. received a graduate stipend from the Cordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES). M.G.V.L. received a graduate stipend from CNPq. D.C.F. and P.H.A. received undergraduate research stipends (IC) from CNPq. A.F.P.A. received a research stipend (PQ) from CNPq.

Conflict of Interest: none declared.

REFERENCES

Chotia,C. (1992) One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Cline,M. *et al.* (2002) Information-theoretic dissection of pairwise contact potentials. *Proteins*, **49**, 7–14.

Crooks,G.E. and Brenner,S.E. (2004) Protein structure prediction: entropy, correlations and prediction. *Bioinformatics*, **20**, 1603–1611.

Crooks,G. *et al.* (2004) Measurements of protein sequence–structure correlations. *Proteins*, **57**, 804–810.

Dill,K.A. *et al.* (2008) The protein folding problem. *Annu. Rev. Biophys.*, **37**, 289–316.

Efron,B. and Tibshirani,R. (1993) *An Introduction to the Bootstrap Monographs on Statistics and Applied Probability*. Chapman & Hall, New York.

England,J. (2011) Allostery in protein domains reflects a balance of steric and hydrophobic effects. *Structure*, **19**, 967–975.

Garnier,J. *et al.* (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.

Gomes,A.L.C. *et al.* (2007) Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. *Proteins*, **66**, 304–320.

Govindarajan,S. *et al.* (1999) Estimating the total number of protein folds. *Proteins*, **35**, 408–414.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Huang,E.S. *et al.* (1995) Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.*, **252**, 709–720.

Koehl,P. and Levitt,M. (2002) Protein topology and stability define the space of allowed sequences. *Proc. Natl Acad. Sci. USA*, **99**, 1280–1285.

Koonin,E.V. *et al.* (2002) The structure of protein universe and genome evolution. *Nature*, **420**, 218–223.

Larson,S.M. *et al.* (2002) Throughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci.*, **11**, 2804–2813.

Onuchic,J.N. and Wolynes,P.G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.

Pande,V.S. *et al.* (1994) Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? *Proc. Natl Acad. Sci. USA*, **91**, 12972–12975.

Pereira de Araújo,A.F. and Onuchic,J.N. (2009) A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. *Proc. Natl Acad. Sci. USA*, **106**, 19001–19004.

Pereira de Araújo,A.F. *et al.* (2008) Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. *Proteins*, **70**, 971–983.

Shakhnovich,E. (2006) Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.*, **106**, 1559–1588.

Solis,A. and Rackovsky,S. (2004) On the use of secondary structure in protein structure prediction: a bioinformatic analysis. *Polymer*, **45**, 525–546.

Solis,A.D. and Rackovsky,S. (2007) Property-based sequence representations do not adequately encode local protein folding information. *Proteins*, **67**, 785–788.

Sullivan,D.C. *et al.* (2003) Information content of molecular structures. *Biophys. J.*, **85**, 174–190.

Zhang,C. and DeLisi,C. (1998) Estimating the total number of protein folds. *J. Mol. Biol.*, **284**, 1301–1305.

BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–5

SUPPLEMENTARY INFORMATION for Information-theoretic analysis and prediction of protein atomic burials: On the search for an informational intermediate between sequence and structure

Juliana R. Rocha*, Marx G. van der Linden*, Diogo C. Ferreira, Paulo H. Azevêdo and Antônio F. Pereira de Araújo†

Laboratório de Biologia Teórica e Computacional, Departamento de Biologia Celular, Universidade de Brasília, Brasília-DF 70910-900, Brazil

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 STATISTICAL SCHEMES FOR BURIAL PREDICTION

1.1 Naive Bayesian Classifier (NBC)

Given a protein sequence, $Q = \{q_1, \cdots, q_N\}$, corresponding discrete burials, $B = \{b_1, \cdots, b_N\}$, and the "alphabets" $\mathcal{Q} = \{\chi_1, \cdots, \chi_{L_\mathcal{Q}}\}$ and $\mathcal{B} = \{\beta_1, \cdots, \beta_{L_\mathcal{B}}\}$ of residue identities and burial levels, $q_i \in \mathcal{Q}$ and $b_i \in \mathcal{B}$, it is possible to estimate for each position i the probability for the different burial levels $\beta \in \mathcal{B}$, conditional to the sequence of a local window of $2w+1$ amino acids centered at position i, $p(b_i = \beta|Q_w)$, or $p(\beta|Q_w)$ for short, with $Q_w = \{q_{i-w}, \cdots, q_i, \cdots, q_{i+w}\}$ and $\sum_\beta p(\beta|Q_w) = 1$ (Fig. 1-a). From Bayes' rule, we have

$$p(\beta|Q_w) = \frac{p(Q_w|\beta)p(\beta)}{p(Q_w)} = \frac{p(\{q_{i-w}, \cdots, q_{i+w}\}|\beta)p(\beta)}{p(\{q_{i-w}, \cdots, q_{i+w}\})}. \quad (1)$$

In a Naive Bayesian Classifier (NBC) it is assumed that residue identities in the window can be considered statistically independent both unconditionally, *i.e.*

$$p(\{q_{i-w}, \cdots, q_{i+w}\}) = \prod_j p(q_{i+j}), \quad (2)$$

and also conditionally to the burial level of the central residue, *i.e.*

$$p(\{q_{i-w}, \cdots, q_{i+w}\}|\beta) = \prod_j p(q_{i+j}|\beta) = \prod_j \frac{p(q_{i+j}, \beta)}{p(\beta)}, \quad (3)$$

where the index j indicates the position in the window, from $-w$ to w. Equation 1 then becomes:

$$\begin{aligned} p(\beta|\{q_{i-w}, \cdots, q_{i+w}\}) &= p(\beta)\prod_j \left(\frac{p(q_{i+j}, \beta)}{p(q_{i+j})p(\beta)}\right) \\ &= p(\beta)\prod_j \left(\frac{p(\chi_j, \beta|j)}{p(\chi_j)p(\beta)}\right), \end{aligned} \quad (4)$$

where $p(\beta)$ is the unconditional burial probability of the central residue, estimated from the the frequency of residues with burial level β in the data bank independently of residue identity or sequence position (when burial levels are equiprobable $p(\beta) = (1/L_\mathcal{B})$ for all β). Note in the denominator of the above equation that the unconditional probabilities of residue identities at position j, $\chi_j = q_{i+j}$, are assumed to be independent of position, or $p(q_{i+j}) = p(\chi_j, j) = p(\chi_j)p(j)$, while the joint probabilities between identities and burials, appearing in the numerator, depends explicitly on the position j, or $p(q_{i+j}, \beta) = p(\chi_j, \beta, j)$. In this way, the number of parameters to be computed from corresponding frequencies in the data bank in order to estimate the conditional burial probability using equation 4 is $L_\mathcal{B} \times L_\mathcal{Q} \times (2w+1)$, for the $\frac{p((\chi,\beta)|j)}{p(\chi)p(\beta)}$ values.

The interpretation of Eq. 4 is straightforward. Sequence-independent probability for burial level β at the center of the window either decreases or increases as knowledge of residue identities, $q_{i+j} = \chi$, at different window positions j are taken into consideration, depending on whether the joint probabilities conditional to j, $p((\chi, \beta)|j)$ are smaller or larger than expected under the assumption of statistical independence, $p(\chi)p(\beta)$. If the probabilities are expressed in negative logarithmic scale the product of factors, either larger or smaller than unity, becomes a sum of mutual information terms, correspondingly either positive or negative, and the resulting scheme becomes similar the classical GOR algorithm for secondary structure prediction (Garnier *et al.*, 1978). Positive qualities of the NBC are its simplicity and flexibility. The predicted atomic burial might correspond, depending on the parameters being used, to any atom of the central residue. It might be considered as independent of residue identity, like C_α in general, or very specific, like $C_\beta 2$ of isoleucine. In this last case positional probabilities in Eq. 4 are necessarily conditioned to central residue identity, or

$$p(\beta|\{q_{i-w}, \cdots, q_{i+w}\}) = p(\beta)\prod_j \left(\frac{p(\chi_j, \beta|j, \chi_0)}{p(\chi_j)p(\beta)}\right). \quad (5)$$

*These authors contributed equally to this work

†to whom correspondence should be addressed

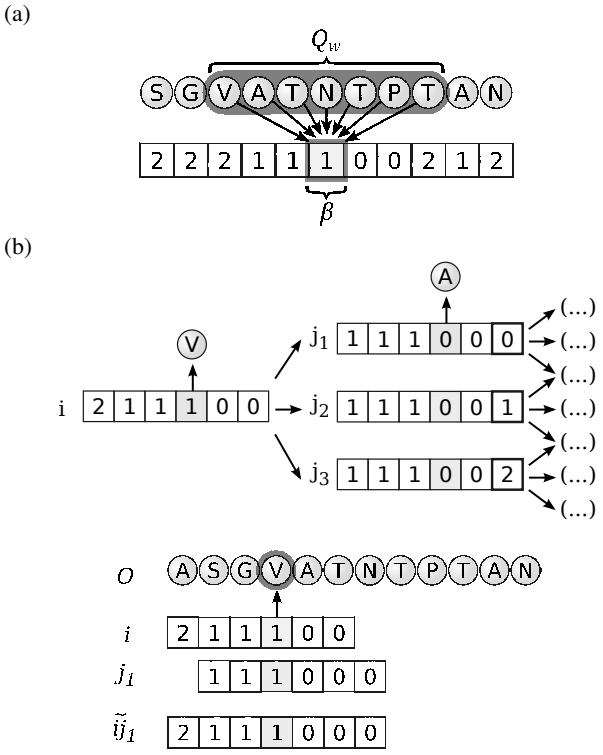


Fig. 1. (a) In the Naive Bayesian Classifier algorithm (NBC), $p(\beta|Q_w)$ defines the probability of having the burial β associated to the central residue in a window of residues Q_w. (b) In the Hidden Markov Model (HMM), each state corresponds to a fragment of $f-1$ hidden variables and has only L_H possible successors (in this example, $f=7$ and $L_H=3$). The transition probability of one state i to another state j is related to $\widetilde{ij}$, the fragment of size f that encompasses both states (Eq. 6)

In the general case this additional conditioning is optional but might be beneficial since identity correlations conditional to single burial are now partially accounted for.

1.2 Hidden Markov Model (HMM)

A discrete, first-order, Markov process is a system comprised of a set of N states and a fixed matrix of transition probabilities $A = \{a_{ij}\}$ with $i,j \leq N$. At any time, the system can be described as being in one of the possible states, i. The system undergoes a change of state at regularly spaced discrete times, with a_{ij} describing the probability of reaching state j immediately after state i. In a Hidden Markov Model (HMM), the states themselves are not observable events, but they define probabilistic functions for "emission" of the observables. The definition of an HMM includes an alphabet of M observable symbols and a probability distribution $B = \{b_j(k)\}$, where $b_j(k)$ is the probability of emitting symbol k when in state j, with $1 \leq j \leq N$ and $1 \leq k \leq M$. One of the basic problems investigated in the context of HMMs is to find the sequence of hidden states that best explains a sequence of observable variables, which is elegantly solved by the forward-backward algorithm (Rabiner, 1989).

We have implemented an HMM for discrete burial prediction inspired by the one proposed by Crooks and Brenner, 2004, for secondary structure prediction. The alphabet of observables $\mathcal{Q} = \{\chi_1, \cdots, \chi_{L_\mathcal{Q}}\}$ consists of amino acid residue identities. In the most simple HP representation we have $M = L_{\mathcal{Q}_{\rm HP}} = 2$ while for 20 amino acids we naturally have $M = L_{\mathcal{Q}_{20}} = 20$. We also define an additional alphabet of general "hidden" variables, $\mathcal{H} = \{\eta_1, \cdots \eta_{L_\mathcal{H}}\}$, comprised of $L_\mathcal{H}$ symbols, which might simply correspond to different burial levels or, alternatively, to more sophisticated descriptors such as burial level combined to secondary structure. Hidden variables must be well defined for all residues. Hidden states in our model are biunivocally mapped to blocks of $f-1$ hidden variables for adjacent residues along the primary sequence. The total number of hidden states is therefore $T_S = (L_\mathcal{H})^{f-1}$, where f (typically around $5-9$) is the size of a fragment containing two overlapped sequences, mapped to states i and j, which is mnemonically represented by $\widetilde{ij}$. Transition probabilities between hidden states must reflect therefore the overlap between corresponding sequences of hidden variables. For example, with $L_\mathcal{H} = 3$, $f = 7$, a state that maps to the burial sequence $i \leftrightarrow [211100]$ has only three possible successors: $j_1 \leftrightarrow [2111000]$, $j_2 \leftrightarrow [2111001]$ and $j_3 \leftrightarrow [2111002]$ (Fig. 1-b). The transition probability matrix is accordingly very sparse, with $a_{ij} = 0$ whenever sequences mapped to i and j do not overlap and

$$a_{ij} = \frac{p(\widetilde{ij})}{p(i)p(j)} \tag{6}$$

for overlapping sequences.

Emission probabilities might also be conveniently considered as dependent on the fragments $\widetilde{ij}$, $B = \{b_{\widetilde{ij}}(k)\}$, where $b_{\widetilde{ij}}(k)$ is the probability of emitting observable symbol k when moving from state i to state j, or

$$b_{\widetilde{ij}}(k) = p(k|\widetilde{ij}) = \frac{p(k,\widetilde{ij})}{p(\widetilde{ij})}. \tag{7}$$

Probabilities on the right side of equations 6 and 7 are estimated from frequencies observed in the training set of representative examples, either using simple counts exclusively or, in the case $p(k|\ \widetilde{ij})$, in combination with pseudocounts to correct for bias from poor sampling of large fragments. Once $A = \{a_{ij}\}$ and $B = \{b_{ij}(k)\}$ have been determined in the training step, prediction is performed by the standard forward-backward algorithm Rabiner (1989), which generates the probabilities $\gamma_t(i)$ of being in any state i when emitting the observed symbol at position t along the sequence. Finally, the probability $P_t(\eta)$ of finding the hidden variable η at position t is obtained by summing over all hidden states containing η at the central position.

$$P_t(\eta) = \sum_{i=1}^{Ts} \gamma_t(i)\delta_\eta(i), \tag{8}$$

where $\delta_\eta(i)$ equals either one, if the central hidden variable in hidden state i equals η, or zero, if this is not the case.

In our tests, in addition to the more straightforward approach in which the only possible values for the hidden variables are the burial levels of the corresponding residues, we also performed predictions with arrangements that employed different combinations of burial levels and additional descriptors of structure to configure the hidden variables. In this case, these extra descriptors were supplied to the

algorithm only in the learning step, alongside with the burial levels, and the prediction step was used to infer all properties at the same time. For example, to simultaneously predict 2-layer burials and 3 possibilities of secondary structure (helix, sheet, loop), an alphabet of 6 hidden variables was used. In a similar arrangement, 4 hidden variables were used to represent 2 layers of C_α burial, plus the information of whether C_β is more or less buried than C_α.

1.3 Evaluation Parameters

Prediction schemes were initially evaluated by their accuracy, computed as the ratio between the number of correct atomic classifications, n_c, and total number of atoms, n_t.

$$A = \frac{n_c}{n_t}. \tag{9}$$

Accuracy is the simplest evaluation parameter for discrete schemes and might be used globally, for atoms in the testing set without distinction between proteins, or locally for each individual protein. Additionally, it might be used for arbitrary sets of atomic types, such as "C_α", "backbone", "side-chain", "Isoleucine $C_\beta 2$", etc. It might provide a meaningful comparison, or at least a reasonable ordering in prediction quality, between schemes using the same input and output representations. Direct interpretation becomes problematic, however, when this is not the case. An accuracy $A = 50\%$ clearly corresponds to a bad prediction for two equiprobable output burial levels, no better than ignoring the input and choosing the output randomly, but it might be significant for three burial levels, in which case random prediction would correspond to $A = 33\%$. Additionally, there is no obvious upper limit indicating optimal performance.

Inspired by the previous analysis of secondary structure prediction provided by Crooks and Brenner, 2004, we have also computed mean log-likelihoods to estimate the mutual information between observed burials and their prediction from sequence,

$$I_p = I(B; B(Q)) = H(B) - H(B|B(Q)), \tag{10}$$

which we call simply "prediction information". The unconditional burial entropy $H(B)$ is simply $\log_2 L$, where L is the number of equiprobable burial layers. The entropy of observed burials conditional to their predictions, $H(B|B(Q))$, is estimated by their mean log odds according to predicted probabilities,

$$H(B|B(Q)) = -\frac{1}{M}\sum_{i=1}^{M} \log_2 p(b_i), \tag{11}$$

where M is the total number of residues in the data bank, labeled by i, with observed burial b_i, and $p(b_i)$ is the predicted probability of this observed burial. For the HMM algorithm, which provides probabilities for whole fragments of correlated burials, it is useful to additionally consider the following approximation for the density of prediction information,

$$\begin{aligned} i_p &= i(B; B(Q)) \\ &= \lim_{N\to\infty} \frac{I(B^N; B^N(Q))}{N} \\ &\approx \frac{I(B^N; B^N(Q)) - I(B^1; B^1(Q))}{N-1} \\ &= h_N(B) - h_N(B|B(Q)), \end{aligned} \tag{12}$$

with

$$h_N(B) = \frac{H(B^N) - H(B^1)}{N-1} \tag{13}$$

and

$$h_N(B|B(Q)) = \frac{H(B^N|B^N(Q)) - H(B^1|B^1(Q))}{N-1}. \tag{14}$$

$h_N(B)$ is computed from the entropy of burial blocks and $h_N(B|B(Q))$ is computed from log odds of burial fragments according to predicted probabilities. Due to the data processing inequality (Cover and Thomas, 2006), prediction information is conveniently bounded by the mutual information between observed burial and amino acid sequence, or $I_p \leq I(B; Q^\infty)$. Similarly, prediction information density must be bounded by the density of mutual information between input and output representations, $i_p \leq i(B; Q)$.

2 ADDITIONAL RESULTS

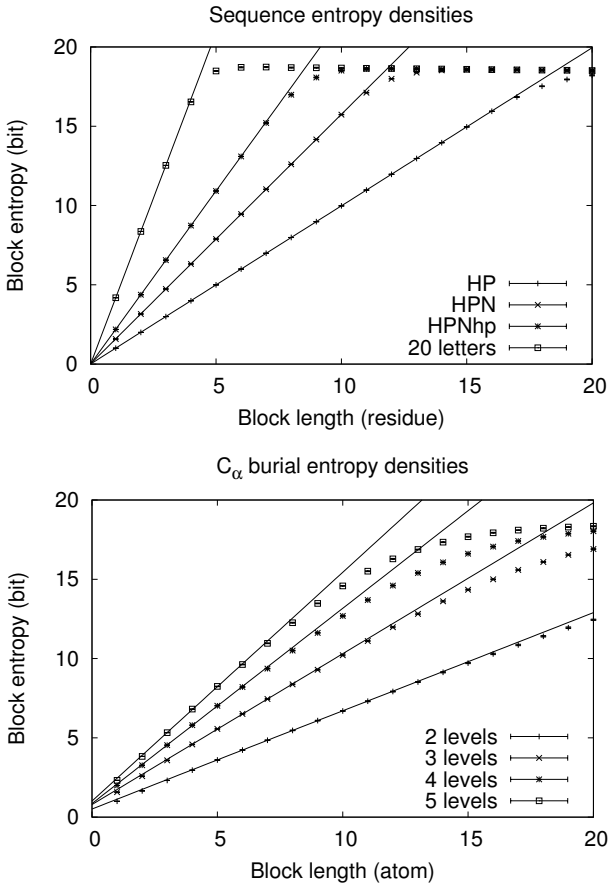


Fig. 2. Entropy for identities and C_α burials. N-block sequence entropy estimates as a function of block size N for different alphabets of C_α burial levels. Straight lines represent linear fits to the data from which the entropy density (inclination) and excess entropy (intersect with the ordinates) are obtained.

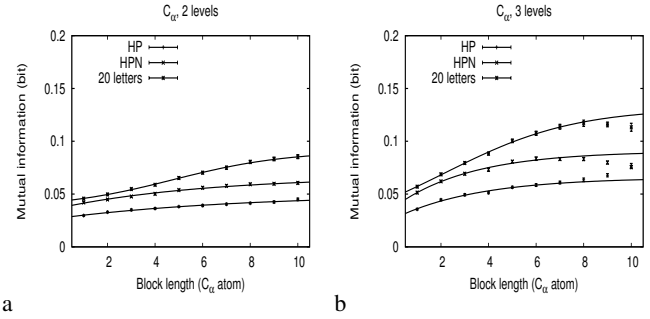


Fig. 3. Mutual information for C_α burials. Estimates for the mutual information, $I(Q_0; B^N)$, between a single central amino acid identity, Q_0, and N-blocks of burials, B^N, as a function of block size N, for 2 (a) and 3 (b) levels of C_α burials. Different sets of points correspond to different alphabets of amino acid identities. Lines represent exponential or sigmoidal fits to the data before saturation from which limiting values $i(Q;B) \approx I(Q_0; B^\infty)$ are obtained. Saturation for $L = 2$ occurs at $N \approx 11$ and is not perceived in the displayed range while for $L = 3$ it occurs $N \approx 8$, as observed in (b).

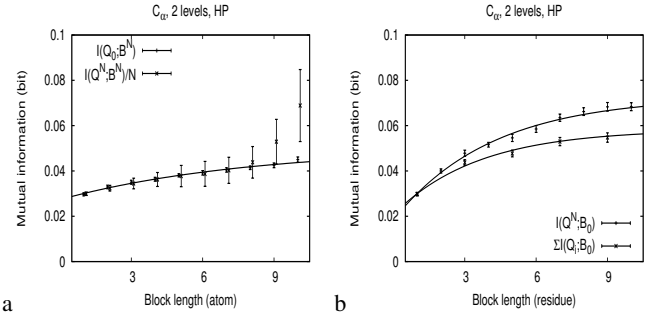


Fig. 4. Approximations in mutual information estimates. Comparison between $I(Q_0; B^N)$ and $I(Q^N; N^N)/N$, for the HP alphabet and 2 levels of C_α burials, reveals virtually coincident values before saturation (a), suggesting that Eq. 3 of the main article is a good approximation. Comparison between $I(Q^N; B_0)$ and $\sum I(Q_i; B_0)$, on the other hand, reveals discrepancies even for small blocks (b), indicating that a strict inequality should be assumed in Eq. 7 of the main article. Curves represent single exponential fits to the data before saturation. The difference in the extrapolated limits, $I(Q^\infty; B_0)$ and $I(Q^\infty; B_0)^-$, is close to $(0.07 - 0.06) = 0.01$ bit, well above sampling error. As expected, saturation for $I(Q^N; B^N)/N$ occurs at smaller N, as indicated by a steep increase at $N \approx 8$ for the present data set, when compared to saturation of $I(Q_0; B^N)$ which occurs at $N \approx 11$. It is indicated, therefore, that amino acid identities in local sequences, although appropriately approximated as independent both unconditionally and conditionally to burial sequence, deviate perceptively from statistical independence upon conditioning to a single residue burial, B_0.

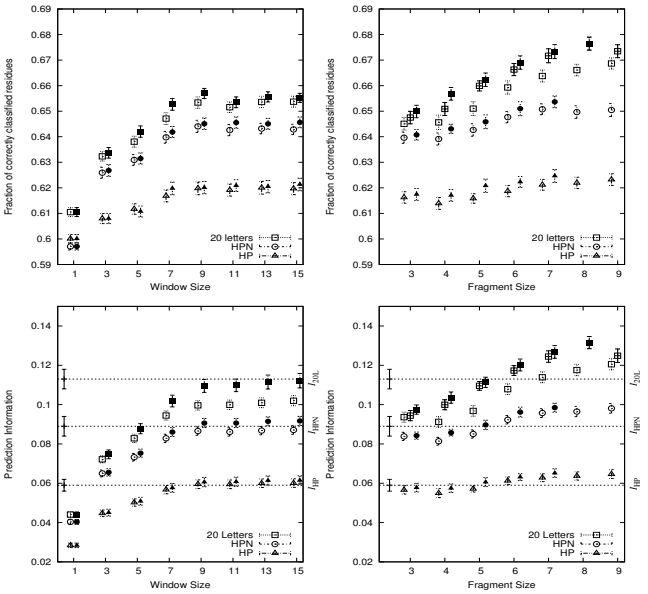


Fig. 5. Accuracy and prediction information for C_α burials. Prediction accuracy A (a,b) and prediction information I_p (c,d) for two levels of C_α burials with different identity alphabets. Plots in the first column (a,c) show results for NBC predictions; the second column (b,d) refers to the HMM results. The NBC method is bounded, within error, to the limits established by corresponding $I(Q^\infty; B_0)^-$ estimates (dotted horizontal lines), while the same limits are surpassed by the HMM method (d). In all plots, unshaded symbols represent the simplest version of each algorithm (NBC1 or HMM with nothing but burial levels encoded into the hidden variables) and shaded symbols represent improved versions (NBC2 or HMM with secondary structures). For HMM, half-shaded symbols represent the version that used sidechain orientations.

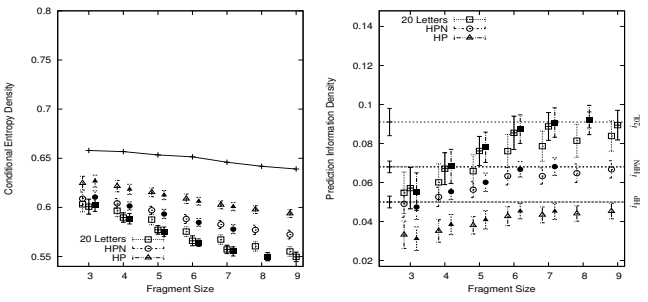


Fig. 6. Prediction information density for C_α burials. For HMM results, the *density* of prediction information, i_p, can be calculated as the difference between an N-dependent estimate for the entropy density of burial levels, $h_N(B)$, Eq. 13, and an analogous estimate for the entropy density conditional to prediction, $h_N(B|B(Q))$, Eq. 14 (shown as points in a). Resulting differences are plotted in (b) in comparison to the upper limit provided by the observed existing mutual information density between burials and sequences, $i(B;Q)$ (horizontal dashed lines). Point symbols are encoded similarly to Fig. 5.

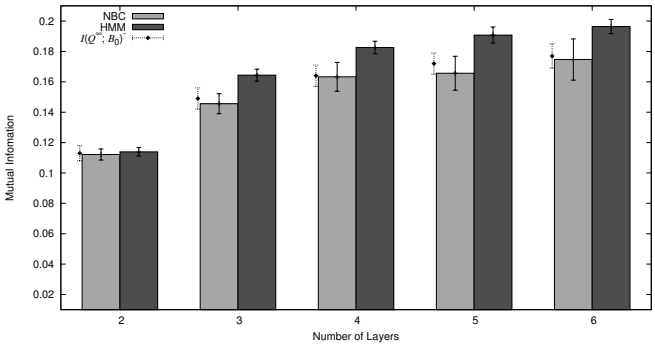


Fig. 7. Dependence of prediction information on the number of burial levels for C_α atoms. As the number of discrete burial layers increases, the quality of prediction, as measured by the prediction information, I_p, also improves, at least up to 4-5 layers. Window size of 15 and fragment size of 7 were used for NBC and HMM, respectively. HMM predictions were performed with no additional descriptors to the hidden variables. Dotted error bars represent the estimated lower bounds for the mutual information between single burial and sequence of identities, $I(Q^\infty; B_0)^-$.

REFERENCES

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, chapter 2. Wiley-Interscience.

Crooks, G. E. and Brenner, S. E. (2004). Protein structure prediction: entropy, correlations and prediction. *Bioinformatics*, **20**, 1603–1611.

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J.Mol.Biol.*, **120**, 97–120.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Ab initio protein folding simulations using atomic burials as informational intermediates between sequence and structure

Marx Gomes van der Linden,[1] Diogo César Ferreira,[1] Leandro Cristante de Oliveira,[1,2] José N. Onuchic,[3] and Antônio F. Pereira de Araújo[1]*

[1] Departamento de Biologia Celular, Laboratório de Biofísica Teórica e Computacional, Universidade de Brasília, Brasília-DF 70910-900, Brazil

[2] Departamento de Física, IBILCE, Universidade Estadual Paulista (UNESP), São José do Rio Preto-SP, 15054-000, Brazil

[3] Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005

The three-dimensional structure of proteins is determined by their linear amino acid sequences but decipherment of the underlying protein folding code has remained elusive. Recent studies have suggested that burials, as expressed by atomic distances to the molecular center, are sufficiently informative for structural determination while potentially obtainable from sequences. Here we provide direct evidence for this distinctive role of burials in the folding code, demonstrating that burial propensities estimated from local sequence can indeed be used to fold globular proteins in ab initio simulations. We have used a statistical scheme based on a Hidden Markov Model (HMM) to classify all heavy atoms of a protein into a small number of burial atomic types depending on sequence context. Molecular dynamics simulations were then performed with a potential that forces all atoms of each type towards their predicted burial level, while simple geometric constraints were imposed on covalent structure and hydrogen bond formation. The correct folded conformation was obtained and distinguished in simulations that started from extended chains for a selection of structures comprising all three folding classes and high burial prediction quality. These results demonstrate that atomic burials can act as informational intermediates between sequence and structure, providing a new conceptual framework for improving structural prediction and understanding the fundamentals of protein folding.

Proteins 2013; 00:000–000.

Key words: protein folding; structure prediction; computer simulation; hydrophobic potential; atomic burial.

INTRODUCTION

General understanding of protein folding has improved significantly during the last decades thanks to theoretical advances framed in terms of energy landscapes, which emphasize the diversity of microscopic routes between unfolded and folded states underlying the observable macroscopic folding behavior.[1–5] Results from the Critical Assessment of Structural Prediction (CASP) experiments[6] have also shown significant improvement in structural prediction, relying strongly on powerful computational resources and an efficient use of information about previously known structures, either in the form of templates for template-based high resolution modeling or in the parametrization of heuristic potentials for free modeling, which still remains more challenging.[7] As another recent encouraging development, computationally intensive "brute force" simulations of fast folding domains have arrived at the native structure using physically inspired semi-empirical potentials.[8–10] However, important these findings might be, particularly considering, on one hand, the significant fraction of the conformational space that has already been mapped[11] and, on the other hand, our continuously growing computational capabilities,[12] it is noteworthy that no simple set of rules associating arbitrary sequences to structures has emerged. The eventual discovery of such rules would give much insight into the actual encoding of native conformations in amino acid sequences and could eventually provide, as a corollary, a general prediction scheme

Grant sponsor: Conselho Nacional de Pesquisa (CNPq); grant number: 478121/2011-3; Grant sponsor: Center for Theoretical Biological Physics sponsored by the NSF; grant numbers: PHY-1308264 and NSF-MCB-1214457; Grant sponsor: Cancer Prevention and Research Institute of Texas (to J.N.O.)..
*Correspondence to: Antônio F. Pereira de Araújo, Departamento de Biologia Celular, Laboratório de Biologia Teórica, Universidade de Brasília, Brasília-DF 70910-900, Brazil. E-mail:aaraujo@unb.br
Received 4 September 2013; Revised 8 November 2013; Accepted 19 November 2013
Published online 00 Month 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24483

capable of dealing with new structures while avoiding time-consuming and error-generating unnecessary details.

In this direction, we have previously shown that the native conformation of globular proteins can be obtained from a modest amount of information about native atomic burials, as expressed by distances to the molecular center, when appropriately combined to simple geometrical constraints.[13,14] Simulations of small globular proteins beginning from randomly generated extended conformations arrived at native-like conformations when atoms were pushed towards their native burials with constraints enforcing covalent geometry and formation of hydrogen bonds for buried putative donors and acceptors, independently of partner. Notably, no pairwise attractive contact interactions nor torsional bias around single dihedrals were required to distinguish the native topology with its correct secondary structure. Furthermore, discretization of provided native burials in a small number of layers, with all atoms in each layer subjected to the same burial force, demonstrated that burial information could be rather imprecise, corresponding to an estimated informational entropy comparable to the entropy of protein sequences.[14] Recent estimates for the mutual information between sequences of amino acids and corresponding burials have also suggested that around 15% of atomic burial uncertainty, for C_α or C_β atoms, could be resolved by local amino acid sequence. Additionally, burial predictions from sequence using simple statistical schemes such as Naive Bayesian Classifiers (NBC) and, particularly, Hidden Markov Models (HMM), were found to successfully extract most of this available sequence-dependent burial information. Extracted information was shown to increase with the number of burial layers, up to at least four layers, and when the dependence of burial on side chain orientation was taken into consideration.[15]

A natural hypothesis consistent with our previous results is that the required burial information for structural determination could be obtained directly from sequence, indicating a possible general mechanism for informational transfer between sequence and structure.[13] In a free analogy with human communication, burials would correspond to the language in which tertiary structures are encoded in the amino acid script. Decoding a particular message would require reading burials from sequence, as we read phonemes from written text, and combining this sequence-dependent information to sequence-independent constraints, analogous to grammatical rules of human languages that associate meaning to a sequence of sounds.[14] A clear separation between sequence-dependent and sequence-independent information, or literature and grammar, has practical implications for the design of folding simulations. Clearly, direct reading from sequence, possibly using statistical learning algorithms, should be attempted only for sequence-dependent information, that is, atomic burials. Additionally, in case of conflict between sequence-dependent and sequence-independent signals, the latter should prevail. Simulations are actually intended to guide the chain to conformations that maximize the compatibility with necessarily inaccurate sequence-dependent information while still satisfying required sequence-independent constraints.

Here we investigate this hypothesis directly. We combine discrete burial predictions from sequence, obtained by a statistical scheme based on a previously described HMM,[15] to ab initio molecular dynamics simulations forcing each atom toward its predicted burial layer with geometric constraints imposed on covalent structure and hydrogen bond formation. We have obtained and distinguished correctly folded conformations for a selected group of globular proteins, comprising all three structural classes and good burial prediction quality, with correct burial assignment into four burial layers for $\approx 56\%$ of the atoms. Sequence-dependent burial constraints in the absence of hydrogen bonds easily guide the chain to an ensemble of layered conformations, with most atoms in their predicted layers but displaying a heterogeneous distribution of RMSD values to the native structure, uncorrelated with burial energy. As sequence-independent hydrogen bonds are gradually enforced within this layered ensemble, the chain is either guided towards native-like conformations, as indicated by a decrease in RMSD values and their correlation with hydrogen bond energy, or adopts protein-unlike conformations detectable by an abnormally low fraction of standard secondary structure. Even though the complete range of applicability for the present approach must still be investigated, including its expected dependence on burial prediction quality and on possible improvement of sequence-independent constraints, our present results already demonstrate that atomic burials can indeed act as informational intermediates between sequence and structure, providing the first direct evidence for our basic hypothesis of a distinctive role of burials in the folding code.

METHODS

We use the term "atomic burial" in the present study to denote the distance of an atom in the native structure of a protein to the structural geometrical center, that is, its "central distance,"

$$r=|\vec{r}|=\sqrt{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}, \qquad (1)$$

where $\vec{r}$ is the "central vector" connecting the geometrical center (x_0, y_0, z_0), whose coordinates are the averages over all atoms in the structure, to the atomic position (x, y, z).[13,14] We divide the structure of N atoms in a small number, L, of concentric "layers" which are used

ID: padmavathym Time: 20:02 I Path: N:/3b2/PROT/Vol00000/130243/APPFile/JW-PROT130243

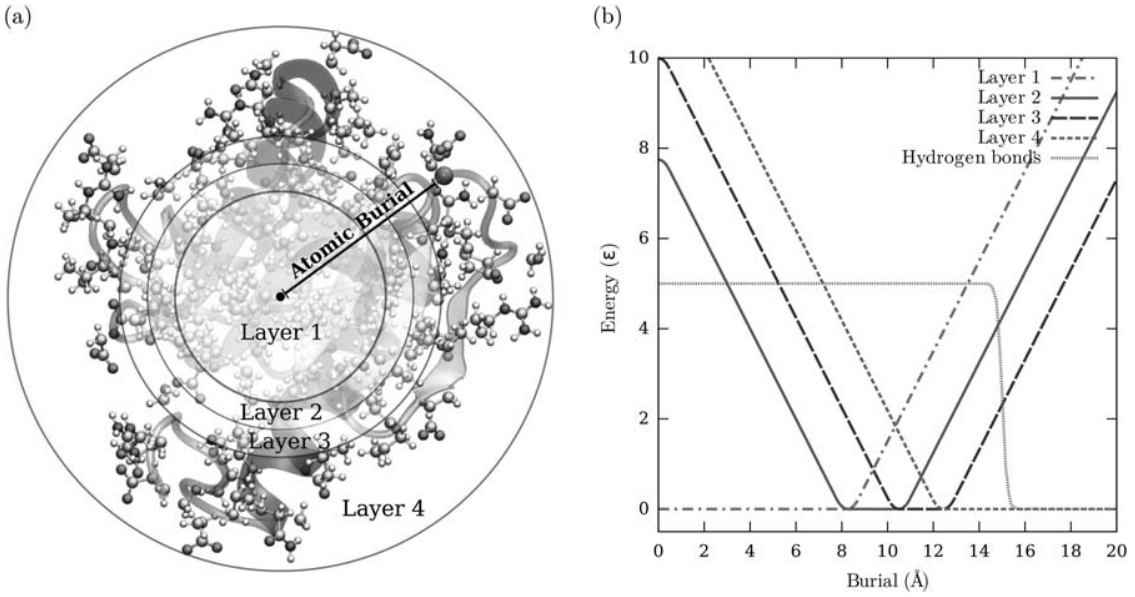


Figure 1

Methods summary. (**a**) Illustration in a specific globular protein of atomic burials and burial layers as used in the present study. Intermediate layers (layer 2 and layer 3) are thinner than the internal layer 1 or external layer 4 in order to provide the same expected number of atoms in all layers. Limits for each layer were obtained from the expected radius of gyration as a function of the number N_r of residues, $R_g = (2.7\sqrt[3]{N_r})$Å, combined to a previously estimated probability density for the number of atoms as a function of normalized central distance, r/R_g.[15] (**b**) Nonstandard terms of the potential function used in the molecular dynamics simulations. The burial potential felt by each atom depending on its predicted burial atomic type increases linearly with distance from the corresponding burial layer. As a crucial sequence-independent term, there is a strong penalty for polar backbone atoms to get near the structural center ($r_i < 15$Å for the three structures under consideration) unless forming a geometrically restrictive hydrogen bond.[14] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to classify all atoms into a correspondingly small set of burial atomic types, represented by the "alphabet" $\mathcal{B}=\{\beta_0, \beta_1, \cdots, \beta_{L-1}\}$. The two central distances limiting each layer are chosen to provide the same expected number, N/L, of atoms in all layers. Here we use $L = 4$ burial layers. See Figure 1(a).

Burial type predictions were obtained with the Hidden Markov Model (HMM) method described in Ref. 16, which was derived from a scheme originally developed for secondary structure prediction.[17] This method is a supervised learning algorithm, meaning that it works by learning statistical patterns of associations between sequences and burials in a training set of proteins with known structures, and then applying these patterns to predict burials for new sequences. In our HMM, the patterns that are inferred during the training phase consist fundamentally of statistical associations along the sequence between adjacent fragments of burial types (transition probabilities) and between amino acid identities and the burial types of surrounding residues (emission probabilities). This model is consistent with the observation that, while amino acid identities are almost statistically independent, atomic burials of adjacent residues are strongly correlated.[16] We used a training set of 278 globular structures smaller than 80 residues, derived from the PDBSELECT[18] list of March 2012, which is intended to maximize structural diversity while minimizing sequence redundancy at the level 25% identity. For burial prediction of all sequences in the training set we used alignments made with CLUSTAL[19] to remove, prior to each prediction, any eventual sequence with more than 25% identity to the sequence being predicted that could have evaded the PDBSELECT procedure. No sequence with more than 20% sequence identity to the predicted sequence was present in the training set for the proteins used in folding simulations. We have also excluded non-globular structures, identified by a large radius of gyration given the number of residues, with $R_g > 2.9\sqrt[3]{N_r}$, and membrane proteins, with pdb files containing the word "MEMBRANE."[16]

The present HMM implementation uses fragments of five adjacent residues for the HMM states of burial types and estimates the probabilities, $p_i(\beta' o)$, for each residue i to have its C_α atom at each burial layer, $\beta' \in \mathcal{B}$, combined to a relative C_β orientation, $o \in \{\downarrow, \uparrow\}$, either less ($\downarrow$) or more ($\uparrow$)distant from the center than C_α, that is, either $r_i^{C_\beta} < r_i^{C_\alpha}$ or $r_i^{C_\beta} > r_i^{C_\alpha}$, respectively. Burial layer probabilities for every heavy atom a in each residue i, $p_i^a(\beta)$, were then obtained by extrapolation using relevant conditional probabilities:

ID: padmavathym Time: 20:02 I Path: N:/3b2/PROT/Vol00000/130243/APPFile/JW-PROT130243

$$p_i^a(\beta)=\sum_{\beta' o} p_i(\beta' o)p_a(\beta|\beta' o) \quad (2)$$

where $\beta \in \mathcal{B}$ and $\sum_{\mathcal{B}} p_i^a(\beta)=1$. The explicit dependence of $p_i^a(\beta)$ on sequence position i comes from the factor obtained from the HMM, $p_i(\beta' o)$. The other factor, $p_a(\beta|\beta' o)$, is the conditional probability of atom a being at burial layer β conditional to C_α burial layer and C_β orientation, $\beta' o$. It depends on the chemical identity of atom a, including residue type, and is estimated from corresponding frequencies in the training set, independently of sequence position. The burial layer with highest probability was finally assigned as the burial type of atom a in residue i.

Molecular dynamics simulations, with all nonhydrogen atoms of the protein represented as single beads of unit mass, m, were performed as in Ref. 14, with a program adapted from a Fortran code previously used in simulations with structure-based C_α models[20] and modified afterwards to handle all atoms. Our burial all-atom model is derived from the all-atom structure-based model described in Ref. 21, with remotion of the native-dependent contact and dihedral energy terms and the addition of specific terms for hydrogen bonds and atomic burials. The resulting potential has the following functional form:

$$V=\sum_{\text{bonds}} k_d(d-d_0)^2+\sum_{\text{angles}} k_\theta(\theta-\theta_0)^2+\sum_{\text{chiral/planar}} k_\chi(\chi-\chi_0)^2 +\sum_{\text{pairs}} \epsilon_{\text{rep}}\left(\frac{\sigma_{\text{rep}}}{d}\right)^{12}+\sum_{\text{atoms}} B(r)+\sum_{\text{don/acc}} \epsilon_{\text{hb}} f(r,\Lambda), \quad (3)$$

where d stands for distance between two atoms, θ and χ for appropriate angles, r for central distance, and Λ for the total number of hydrogen bonds in which a putative donor or acceptor is involved, as explained below. The harmonic terms constrain the covalent geometry through bond distances, angles, planar dihedrals (peptide bonds and aromatic rings), and C_β chirality, with $k_d=100\epsilon\text{Å}^{-2}$, $k_\theta=20\epsilon\text{rad}^{-2}$ and $k_\chi=10\epsilon\text{rad}^{-2}$, for planar dihedrals, or $20\epsilon\text{rad}^{-2}$, for C_β chirality, where ϵ is our unit of energy, while d_0, θ_0, and χ_0 are taken from an extended conformation constructed from sequence with standard amino acid geometries using the program PyMOL.[22] The repulsive term is applied to all pairs of atoms that are separated by more than two covalent bonds and do not belong to the same planar dihedral, with $\epsilon_{\text{rep}}=1.0\epsilon$ and $\sigma_{\text{rep}}=2.5$Å, except for the repulsion between C_β carbons and backbone carbonyl oxygens, in which case a larger value $\sigma_{\text{rep}}=3$Å is used instead.

The burial term is applied to all atoms, pushing them towards their type-dependent predicted layers with a constant force of $\pm 1\epsilon\text{Å}^{-1}$. That is, $B(r)$ is zero everywhere inside the 2δ long interval $(r^*-\delta, r^*+\delta)$ and increases linearly outside this interval with slope ± 1, except for small quadratic sections required to maintain differentiability at every point, as described in Ref. 14 and shown in Figure 1(b). The burial parameters r^* and δ are fixed for each of the four layers in terms of the radius of gyration, R_g. We use $(r^*/R_g, \delta/R_g)=(0.378, 0.378), (0.859, 0.103), (1.051, 0.089)$, and $(1.57, 0.43)$ for the four burial layers, in this order. These values correspond to equal areas under the burial probability density function estimated in Ref. 15. The radius of gyration of each simulated protein is estimated from its number of residues, N_r, according to an expected dependence $R_g \approx 2.7\sqrt[3]{N_r}$ Å.

The hydrogen bond term is applied to all backbone nitrogen and oxygen atoms and is intended to penalize their internalization, by ϵ_{hb}, unless they form a single geometrically restrictive hydrogen bond, independently of partner. We use the following function in the hydrogen bond term:

$$f(r,\Lambda)=F(r)(1-\Lambda), \text{for } \Lambda \leq 0.95 \quad (4)$$

and

$$f(r,\Lambda)=0, \text{for } \Lambda > 1.05, \quad (5)$$

with an appropriate intermediate quadratic region for $0.95 < \Lambda < 1.05$, with derivative increasing linearly from -1 to 0, in order to maintain differentiability at $\Lambda=1$. The constant value of 0 for $\Lambda > 1.05$ is a modification with respect to our previous study and is intended to avoid a bias for multiple bond formation by a single putative donor or acceptor. Multiple bond formation by single atoms were not a problem for the potential with native burials but can become a complication when predicted, and necessarily inaccurate, burials are used instead. The dependence on r is still provided by a Fermi function

$$F(r)=\frac{1}{1+\exp(\beta_r(r-\mu_r))}, \quad (6)$$

which changes from 1 to 0 abruptly, as controlled by β_r, around $r=\mu_r$. Here we use $\mu_r=15$Å and $\beta_r=10\text{Å}^{-1}$. We also quantify hydrogen bond formation between a possible donor, i, and a possible acceptor, j, by a combination of Fermi functions,

$$\lambda_{ij}(h,\eta,\theta)=F(h)F(\eta)F(\theta), \quad (7)$$

which changes abruptly but continuously from 1 to 0 as any of the three controlling variables exceeds their thresholds. These three controlling variables are computed from the coordinates $\{\vec{r}_1,\ldots,\vec{r}_5\}$ of the following five atoms: the acceptor carbonyl oxygen (1), the donor nitrogen (2), the two atoms adjacent to this nitrogen (3

ID: padmavathym Time: 20:02 I Path: N:/3b2/PROT/Vol00000/130243/APPFile/JW-PROT130243

and 4), and the carbon adjacent to the acceptor oxygen (5). These coordinates define three convenient vectors: $\vec{v}_1=\vec{r}_2-\vec{r}_1$, $\vec{v}_2=\vec{r}_3+\vec{r}_4-2\vec{r}_2$ and $\vec{v}_3=\vec{r}_1-\vec{r}_5$. In terms of these vectors $h=|\vec{v}_1|$ is the norm of $\vec{v}_1$, η is the angle between $\vec{v}_1$ and $\vec{v}_2$, and θ is the angle between $\vec{v}_1$ and $\vec{v}_3$.

The total number of hydrogen bonds formed by a given possible donor, *i*, is obtained by the sum of hydrogen bond formation for all putative bonds in which it is involved,

$$\Lambda_i=\sum_j \lambda_{ij}, \tag{8}$$

and conversely for possible acceptors. We use the following hydrogen bond parameters: $\mu_h=3\text{Å}, \beta_h=100\,\text{Å}^{-1}$, $\mu_\eta=0.5\,\text{rad}$, $\beta_\eta=100\,\text{rad}^{-1}$, $\mu_\theta=0.7\text{rad}$, $\beta_\theta=100\ \text{rad}^{-1}$. The hydrogen bond energetic penalty, ϵ_{hb}, increased linearly from 0 to 5ϵ (annealed hydrogen bonds) during the simulation. Absolute temperature was maintained at $T=1\epsilon$ by a Berendsen thermostat. The energetic cost of $2\epsilon_{hb}$ for breaking a buried hydrogen bond is therefore 10 T at the end of the simulations with hydrogen bond annealing. The time step of integration in the molecular dynamics procedure is 0.005τ, where τ is the unit of time determined by our units of distance, Å, mass, *m*, and energy, ϵ, that is, $1\tau=1\text{Å}\sqrt{m/\epsilon}$.

The potential is therefore very simple and not intended, at least in its present form, to distinguish between minuntiae of different native-like conformations. Notably, no attractive van der Waals interaction is included nor any torsional potential around single bond dihedrals, eventhough favorable contact pairwise interactions and dihedral orientations are known to play a dominant role in many other potentials used in folding simulations.[23–25] Previous simulations,[14] however, have shown that the present potential, using native burial information in the burial terms with as few as just three burial layers, is sufficient not only to constrain the chain within a native-like ensemble with average RMSD from the native structure around a couple of angstroms, but also to consistently guide randomly generated initial conformations to this native-like ensemble. Additionally, due to this simple form, conformational sampling is relatively fast since the underlying energy landscape is much smoother than for detailed potentials intended to distinguish between slightly different high resolution models. In our previous study with native burials the burial term was annealed during the simulations. It turns out that the presently performed annealing of the hydrogen bond term is more efficient in avoiding kinetic trapping, particularly for proteins containing β-sheets.

RESULTS

Burial prediction results are summarized in Figure 2(a). Prediction accuracy for all heavy atoms in four burial layers is plotted for the 278 small globular structures against accuracy rank. Accuracy varies from close to 25%, which is no better than random prediction, to higher than 60%, with an average of 45% and standard deviation of 7% (horizontal lines). The average log-likelihood, $\langle LL\rangle=-(1/M)\sum p_i^a(\beta_n)\log_2 p_i^a(\beta_n)$, where $p_i^a(\beta_n)$ is the predicted probability for the observed burial layer in the native structure, β_n, and M is the number of atoms in the data set, is close 1.75 bits, resulting in a prediction information, as used in our previous study,[16] $I_p=\log_2 4-\langle LL\rangle\approx 0.25$ bits. This is higher than our previously reported values, below 0.2 bits, for four layers of C_α or C_β atoms with size-independent training and testing sets, with no separation by chain length.[16] Consistently with this last observation, Figure 2(b) shows that the presently reported prediction accuracy for all atoms in small structures is also higher than for predictions of the same structures using a size-independent training set. It should be noted that prediction improvement arises from the present compatibility, with respect to chain length, between structures in the training set and sequences to be predicted, and not because prediction for very small proteins is particularly easy. As shown in the same panel, prediction accuracy for groups of longer chains, with the number of residues N_r satisfying $81<N_r\leq 120$ or $121<N_r\leq 160$, are actually comparable to our present results for $N_r\leq 80$ when training is performed inside each group and higher than for training in the whole set including all lengths. For even longer chains we observe a slight decrease in prediction accuracy.

Maximal, average and standard deviation of pairwise identities with proteins in the training set are also shown for each predicted protein in Figure 2(a) and no correlation with prediction accuracy is apparent. Accuracy in the investigated set does not correlate either with simple structural parameters such as the fraction of α or β secondary structure (not shown). We observe, however, that prediction accuracy for each protein does increase, as expected, for more reliable atomic predictions, as measured by the probability, $p_i^a(\beta_p)$, for the predicted layer β_p. As seen in Figure 2(c), prediction accuracy for each protein tends to increase slightly for predictions with $p_i^a(\beta_p)>0.3$ and more significantly for $p_i^a(\beta_p)>0.4$ and $p_i^a(\beta_p)>0.5$. The fraction of atoms satisfying these increasing restrictive criteria in each protein necessarily decreases, however, as shown in Figure 2(d). In the present investigation we use all burial predictions on an equal basis, independently of reliability.

We selected three proteins with high burial prediction accuracy, comprising all three structural classes, for a detailed analysis using molecular dynamics folding simulations: the all-α XLR Effector AVR3A11 from *Phytophthora capsici* (RePc, PDB code 3zr8, 65 aa), a variant of the α/β protein G β2 domain from *Streptococcus* sp. (ProtGSsp, PDB code 3fil, 56 aa), and the topologically

ID: padmavathym Time: 20:02 I Path: N:/3b2/PROT/Vol00000/130243/APPFile/JW-PROT130243

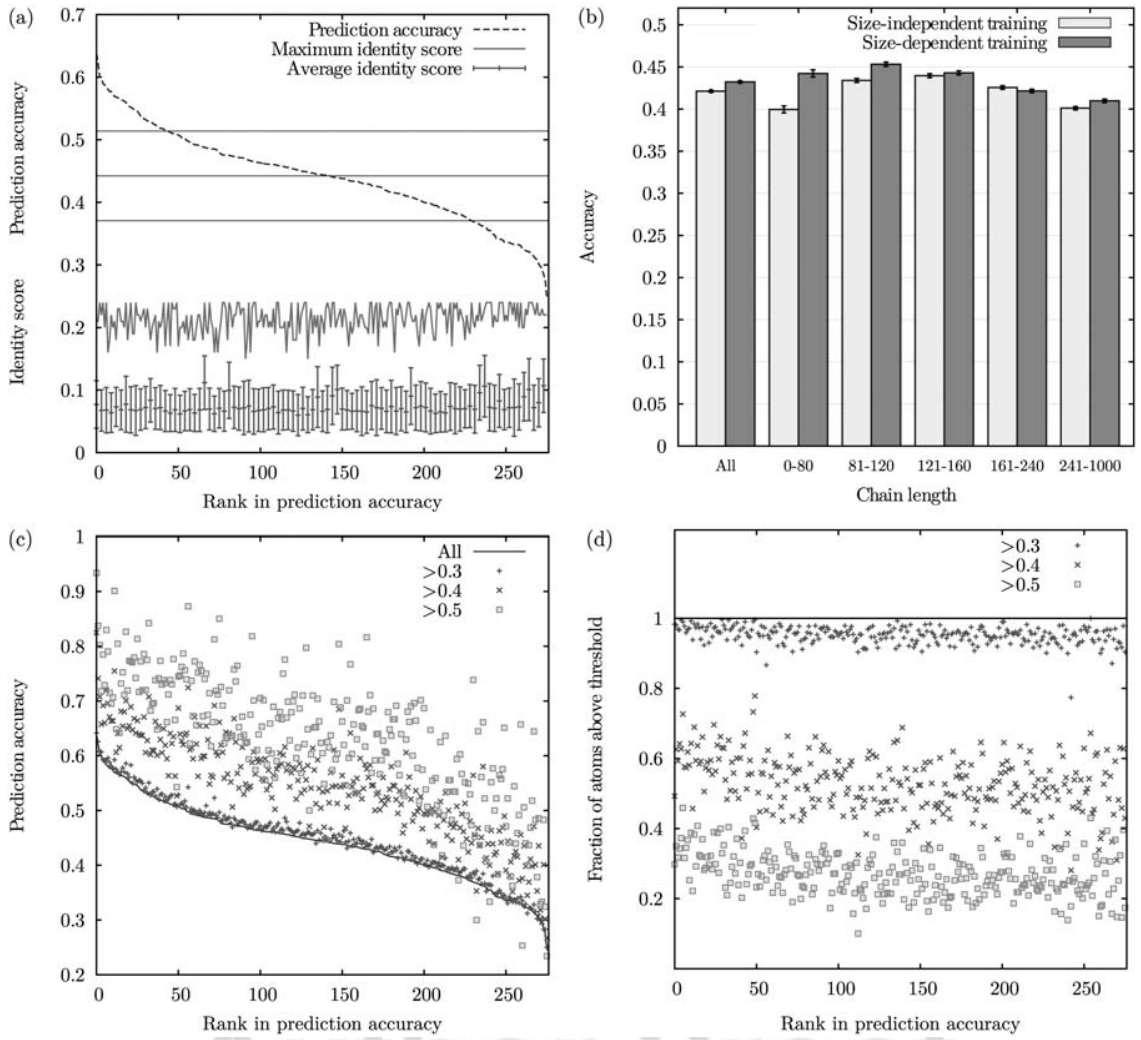


Figure 2

Burial prediction results. (**a**) The horizontal axis represents the N_s=278 proteins in the dataset ordered by prediction accuracy rank. The sigmoidal curve indicates the fraction of residues that were assigned to the correct burial layers for the respective protein with sample average and standard deviation, $\mu_s \pm \sigma_s$, indicated by horizontal lines. Identity scores between the reference protein at a given rank and all proteins in the training set were obtained with CLUSTAL[18] and their average and standard deviation are indicated by error bars while the highest scores are shown by the connected line. (**b**) Average accuracy for each group of structures as a function of chain length range for size-dependent and size-independent training sets are represented by solid bars. Error bars represent standard deviations of the mean, $\sigma_s/\sqrt{N_s}$, which are small because the number of structures in each sample is large. For $N_r < 80$, for example, we have $\sigma_s \approx 7\%$, as shown in (a), and $\sigma_s/\sqrt{N_s} \approx 7/\sqrt{278} \approx 0.05\%$. (**c**) Each set of points shows, as a function of prediction accuracy rank, the prediction accuracy exclusively for atoms whose probabilities of being in the predicted layer, $p_i^a(\beta_p)$, are above a certain threshold. (**d**) Each set of points shows the fraction of atoms satisfying the criteria used in the previous plot. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

nontrivial all-β cold shock protein of *Bacillus caldolyticus* (CsBc, PDB code 1c9o, 66 aa). Their burial predictions are shown in Figure 3(a–c), with a percentage of correctly assigned atoms around 56%. For comparison purposes, around 15% of our current protein dataset of small structures have more than 50% of their atoms correctly assigned by this procedure. The training set for each of these three selected proteins contained only sequences with less than 20% identity to them and burial prediction accuracy would not change significantly if more stringent cutoffs were used instead: less than 1% difference at 15% cutoff and within 3% difference (54%, 57%, and 59%) at 10% cutoff. Molecular dynamics simulations for these three proteins with annealed hydrogen bonds were performed as described in the Methods section. As nonstandard terms we have a constant force pushing each atom toward a single burial layer among four pre-established possibilities and an energetic penalty,

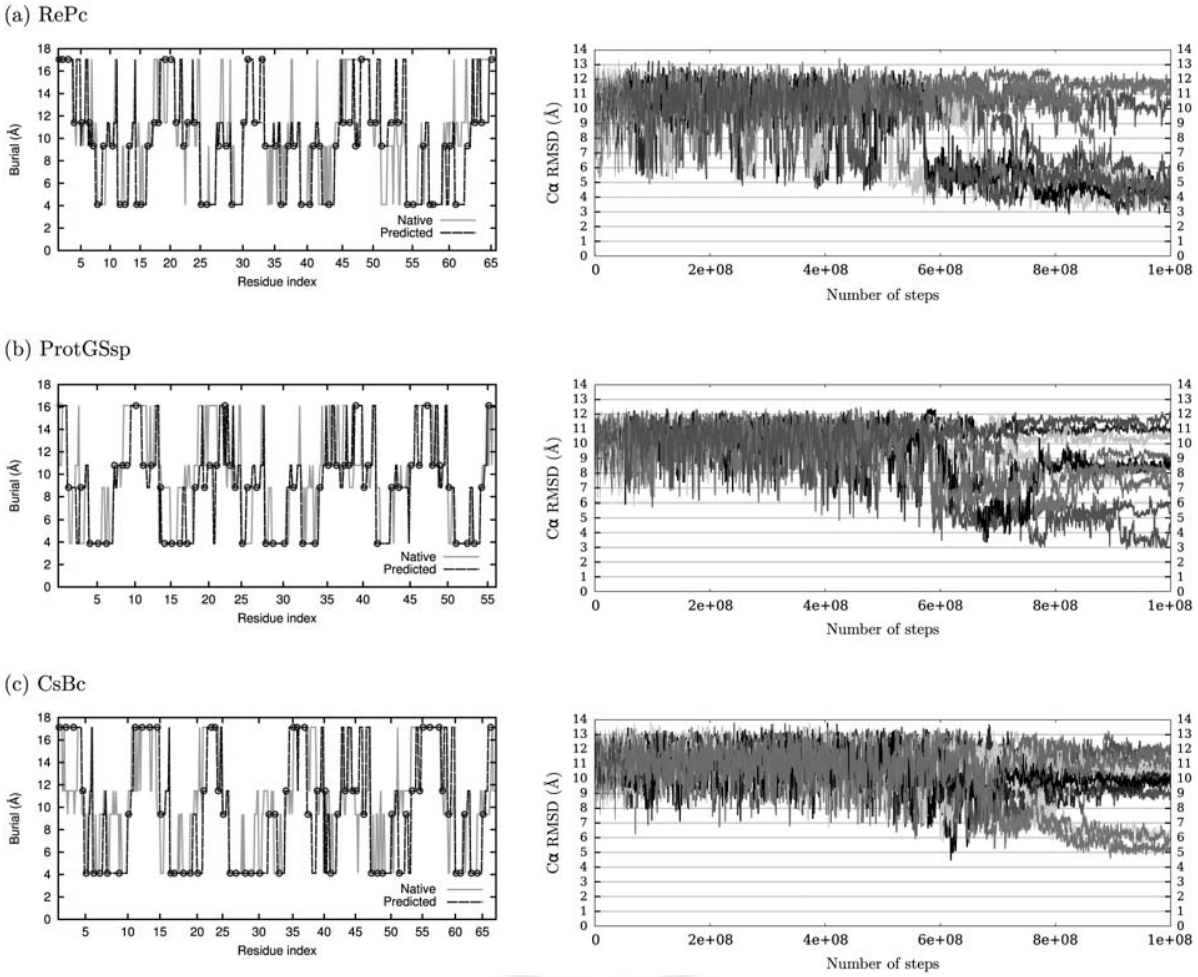


Figure 3

Burial predictions and folding trajectories. Prediction of burial layers (left). Central layer positions for each heavy atom are in shown in blue/dashed lines for predicted burials and in red/solid lines for actual burials, in terms of distance from the molecular center. Circles indicate C_α atoms. Folding simulations (right). C_α RMSD is plotted as a function of simulation time step. Each panel shows ten to eighteen independent trajectories for a single protein, using a burial potential derived from the prediction shown immediately above. Burial constraints remain constant while the energetic penalty for not forming hydrogen bonds increases linearly in time. Simulations were performed at absolute temperature (in energy units) $T=\epsilon$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

which increases linearly in time, for backbone polar atoms to get buried unless forming a geometrically restrictive hydrogen bond, independently of partner. The layer to which each atom was pushed, that is, its burial atomic type, was obtained from the burial prediction. No information about the native structure was included.

As shown in Figure 3(d–f), the C_α root mean square deviation (RMSD) from the native structure oscillates strongly during the first half of the simulations, reflecting rapid interconversion between different conformations in the absence of strong hydrogen bonds, even in the presence of burial constraints. As hydrogen bonds become stronger, oscillations become smaller and different trajectories converge to rather uniform conformational ensembles. Seven out of ten trajectories resulted in correct folding for the α-helical RePc, converging to RMSD values between 3Å and 5Å, which is consistent with trajectories beginning from the native structure under the same final conditions (not shown). We find that the average hydrogen bond (HB) energy term over the last 10% steps of each trajectory performs quite well in discriminating final trajectories with low average RMSD. As shown in Figure 4(a), the three trajectories with lowest average HB energy, between 140ϵ and 150ϵ, correspond to the lowest average RMSD values, around 4Å. The only other trajectory with average HB energy below 160ϵ displays all helices correctly formed but in an incorrect mutual disposition, symmetric to the native structure, F4

ID: padmavathym Time: 20:02 I Path: N:/3b2/PROT/Vol00000/130243/APPFile/JW-PROT130243

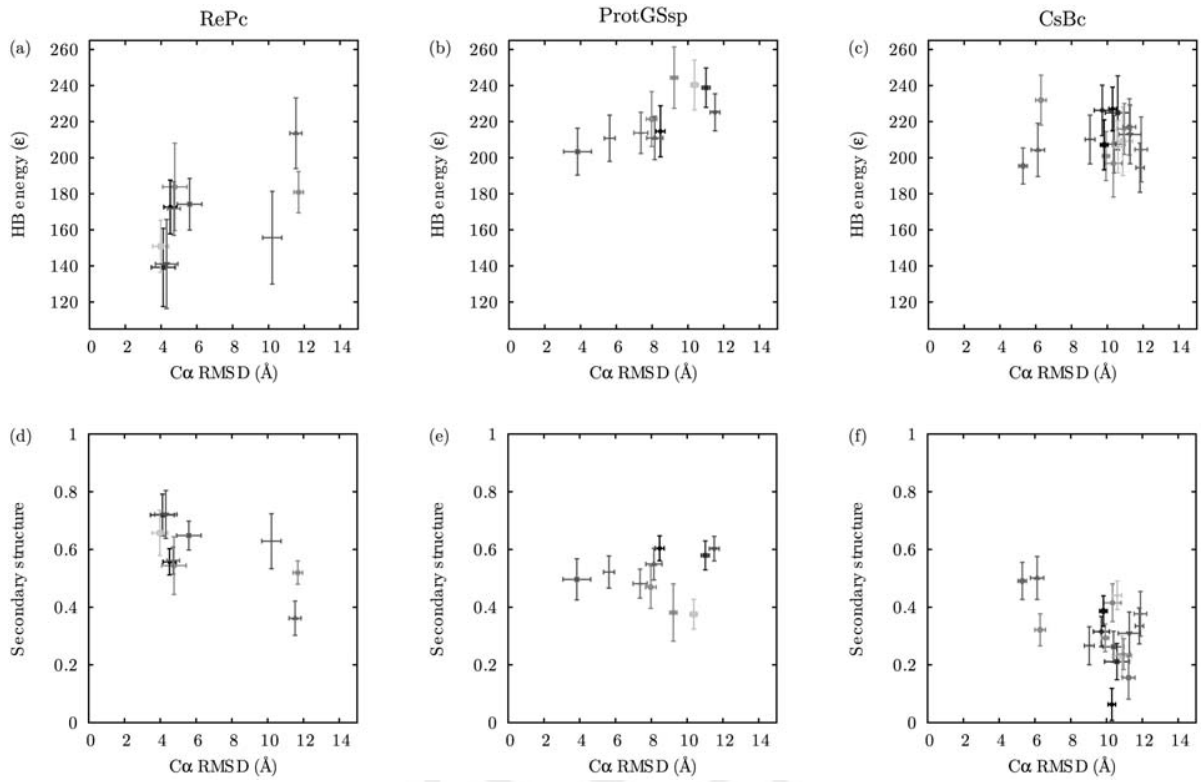


Figure 4

HB energy and ratio of secondary structure for the final portion of trajectories. Average hydrogen bond energy, (top) and average fraction of residues forming standard α-helix or β-sheet secondary structure (bottom) for the last 10% of each trajectory shown in Figure 3, plotted as a function of average C_α RMSD from the native structure, with corresponding standard deviations as error bars. The HB energy is part of the actual potential governing the trajectories while secondary structure was computed independently by the Dictionary of Protein Secondary Structure (DSSP)[27] with the program provided at the site http://swift.cmbi.ru.nl/gv/dssp. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

with RMSD around 10Å. Such "mirror" images were also found in our simulations using native burial information[13,14] and have been previously observed in other simulations of helical proteins with pairwise contact potentials.[26]

For ProtGSssp, the distribution of final RMSD for different trajectories was more heterogeneous, with two trajectories with final RMSD values between 3Å and 6Å, but average RMSD values correlates with average HB energy, with the lowest and second lowest average HB energies around 200ϵ and 210ϵ, respectively, appropriately corresponding to lowest and second lowest average RMSD around 4Å and 6Å. For the topologically complex CsBc, three trajectories out of eighteen resulted in average RMSD values between 5Å and 6Å. However, the group of three trajectories with low final RMSD values is not clearly distinguished from the remaining trajectories by their average HB energies alone. The trajectory with lowest average RMSD, close to 5Å, is actually one of the two trajectories with lowest average HB energies, both close to 195ϵ, but the average RMSD for the other low-energy trajectory is close to 12 Å. Additionally, the trajectory with second lowest RMSD, close to 6 Å corresponds to an average HB energy close to 205ϵ while many trajectories with large RMSD, between 9 Å and 12 Å, also have average HB energies between 200ϵ and 210ϵ.

We observe, however, that many of these trajectories display an abnormally low fraction of residues adopting standard α-helix or β-sheet secondary structure, below 0.4 as computed by the Dictionary of Protein Secondary Structure (DSSP),[27] and that the only two to display a fraction of secondary structure formation above 0.5 are in the group of low RMSD, as seen in Figure 4(f). We note that one trajectory RePc and two trajectories for ProtGSsp also display a low fraction of secondary structure, as seen in Figure 4(d,e), but they were correctly distinguished by the HB energy term, as should be expected. The fact that trajectories for CsBc that have a low fraction of secondary structure might still have low HB energy values might suggest some imperfection in

ID: padmavathym Time: 20:02 I Path: N:/3b2/PROT/Vol00000/130243/APPFile/JW-PROT130243

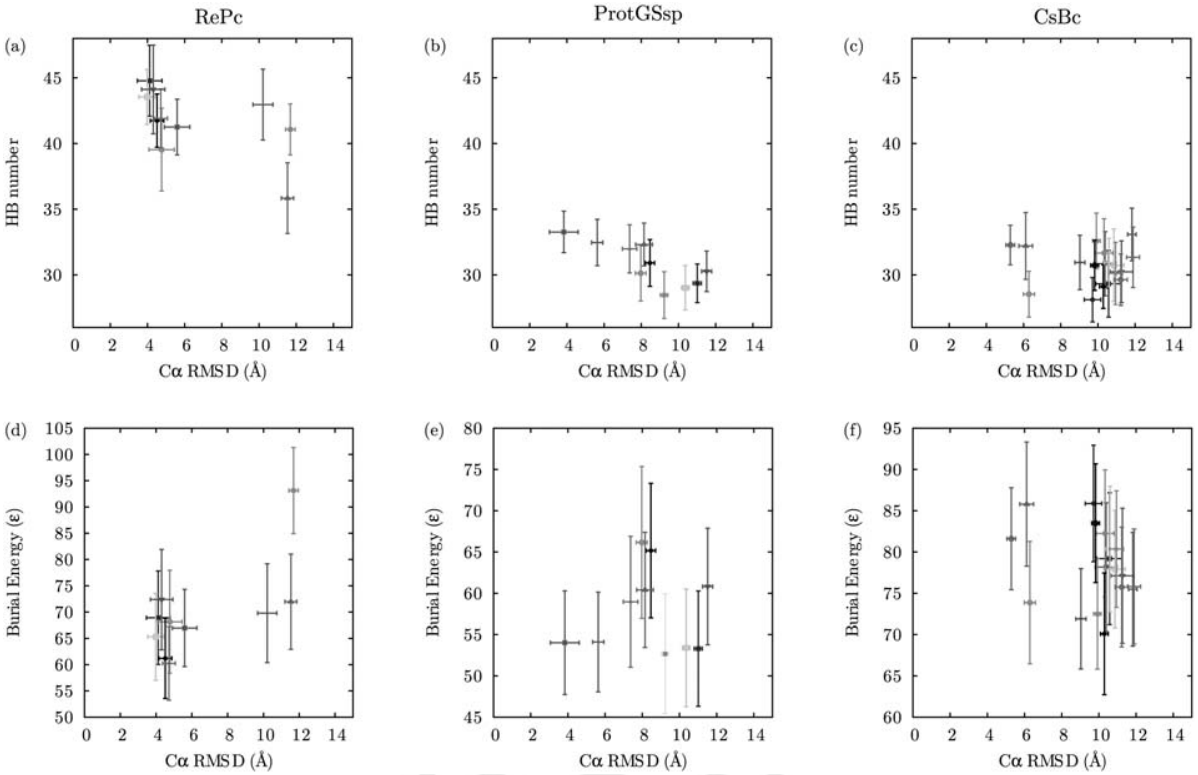


Figure 5

HB number and burial energy of the final portion of trajectories. Number of formed hydrogen bonds (top) and value of the burial term in the energy potential (bottom) for the last 10% of each trajectory shown in Figure 3, plotted as a function of average C_α RMSD from the native structure, with corresponding standard deviations as error bars. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

our current sequence-independent constraints, which happened to be more perceptible for the all-β protein, but also points out a direction for possible improvement.

Since the HB energy is inversely correlated to the number of hydrogen bonds, provided conformations are equally compact, the total number of hydrogen bonds, or HB number, should provide a similar indicator of low RMSD trajectories. Figure 5(a–c) shows that this is actually the case and that differences in final HB energy between trajectories correspond to an appropriate difference in average HB number, with the increment of a single hydrogen bond corresponding roughly to a decrease of 10ϵ in HB energy. Differently from HB energy or HB number, the burial energy is not a good indicator of low RMSD trajectories. As shown in Figure 5(d–f), no correlation is apparent between average final RMSD and average final burial energy for the trajectories of ProtGSssp and CsBc, or for the group of RePc trajectories with low average RMSD, below 6 Å. Furthermore, the range of differences in average burial energy between trajectories, typically around 10ϵ, is smaller than the range of differences in average HB energy, typically around 40ϵ.

In any case, it is a particularly encouraging result that whenever standard secondary structure does form it is consistent with the native pattern and the HB energy becomes indicative of native topology. If we eliminate final trajectories with a fraction of secondary structure below 0.4 and select from the remaining lot the one with lowest average final HB energy, a trajectory with low average final RMSD is obtained. If we now choose inside this trajectory the individual conformation with the highest number of hydrogen bonds we arrive at the structures shown in Figure 6. The agreement with their native counterparts is encouraging, particularly considering the simplicity of our scheme and the fact that no information about the native structure was used either in the folding simulations or in the selection criterion.

The interplay between sequence-dependent burial energy and sequence-independent hydrogen bond formation along the whole folding process is illustrated in Figure 7 with a single trajectory of RePc. RMSD as a function of simulation time step for this trajectory, which was already shown in Figure 3(d) inside the group of ten RePc trajectories, is now more clearly seen by itself

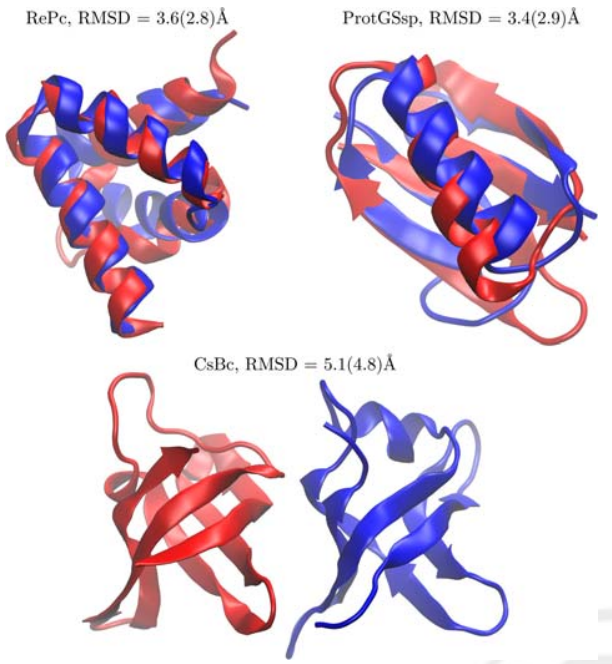


Figure 6

Representative conformations. Representative conformation obtained from the trajectories (blue) for each of the three proteins compared to the native structure (red). For each protein, we selected among the final parts of the independent trajectories the one with the lowest average HB energy, excluding beforehand any eventual trajectory that reached an average fraction of secondary structure below 0.5. The conformation with the largest number of hydrogen bonds in the selected trajectory was then adopted as our predicted structure (when more than one conformation had the same number of hydrogen bonds, we choose the one with lowest energy for the hydrogen bond term). C_α RMSD values for the selected structures are shown beside their names, being close to 0.5 Å higher than the global minimum in their trajectories, which are shown in parenthesis.

C O L O R

in Figure 7(a), while burial energy, HB energy and HB number are shown in Figure 7(b). Average HB energy initially increases, as expected, as the HB energetic penalty ϵ_{hb} is gradually augmented but only a very few hydrogen bonds are transiently formed in this initial part of the trajectory. HB number then increases rather abruptly while HB energy initially decreases and then appears to stabilize, on average, with occasional oscillations mirroring the behavior of HB number. The onset of hydrogen bond formation coincides with the decrease in RMSD fluctuations and convergence to a low RMSD conformational ensemble. The burial energy, on the other hand, oscillates around 80ϵ from the beginning of trajectory and is not greatly affected by the onset of hydrogen bond formation. Within the conformational ensemble explored in the trajectory, therefore, conformational distance from the native structure, as measured by RMSD, is not correlated to the sequence-dependent burial energy but inversely correlated to sequence-independent hydrogen bond formation, particularly for low RMSD values, as directly seen in Figure 7(c,d).

Sequence-independent hydrogen bond formation, and not the sequence-dependent burial energy term, arises therefore as an indicator of native-likeness within the conformational ensemble explored during our simulations. Note that this conformational ensemble, however, is already efficiently constrained by the sequence-dependent burial term. As observed in Figure 7(a), and also in Figure 3(a), the RePc chain transiently adopts conformations with RMSD values close to 5 Å from the native structure in the first part of the trajectory, when hydrogen bonds are still absent. Relatively low RMSD values in the absence of hydrogen bonds are also observed in the trajectories of the two other proteins, with different ranges in explored RMSD values, as seen in Figure 3(b,c). For comparison, we show in Figure 8 F8 that trajectories of RePc with all atoms being pushed in the absence of hydrogen bonds to a single burial layer, combining the four original layers, display larger average and minimum RMSD from the native structure than the initial 10% of the folding trajectories. In fact, RMSD values to the native structure of the all-α RePc are similar for the compact ensemble of RePc itself and the compact ensemble of the all-β CsBc, and vice versa.

These compact conformational ensembles, therefore, even though satisfying the sequence-dependent covalent geometry, are effectively sequence-independent in terms of conformational distance to any particular native structure. For the folding trajectories, on the other hand, sequence-dependent burial constraints result in sequence-dependent, "layered", ensembles with appropriately smaller RMSD values to the corresponding native structure. In other words, sequence-dependent burial information determines a constrained, layered, conformational ensemble but is not able to distinguish native-like conformations within this ensemble. Sequence-independent hydrogen bonds, on the other hand, become a good indicator of native-like conformations within the sequence-dependent layered ensemble. They would not be able to distinguish a unique native structure inside sequence-independent compact ensembles because all globular native structures would be compatible with the constraints in compaction and hydrogen bond formation.

DISCUSSION

We have obtained and distinguished native-like conformations in ab initio folding simulations using sequence-dependent burial predictions combined to sequence-independent geometrical constraints on covalent geometry and hydrogen bond formation. This important result demonstrates the possibility of using atomic burials as informational intermediates between sequence and

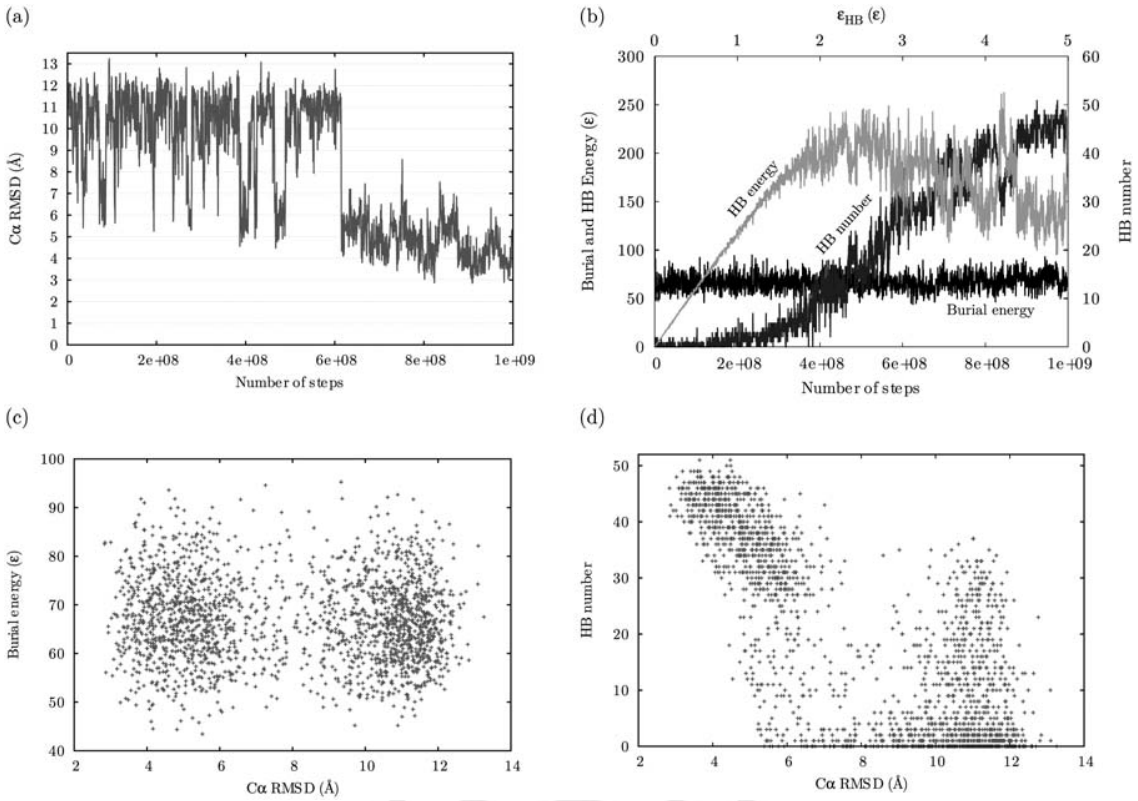


Figure 7

Analysis of a single trajectory of RePc. (**a**) Progress of the native Cα RMSD of a single trajectory of RePc as a function of simulation time step. (**b**) Burial and HB energies are shown as functions of time step in the scale indicated on the left hand vertical axis. The number of hydrogen bonds formed by the structure in each step (HB number) are shown to the scale indicated in the right hand vertical axis. (**c**) Burial energy as a function of Cα RMSD for the entire trajectory. (c) HB number as a function of Cα RMSD for the entire trajectory. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

structure in folding simulations and protein structure prediction. The resulting scenario is physically intuitive and consistent with general folding principles already known for decades, although not usually emphasized in prediction schemes or folding simulations. The tendency of different regions of the chain to be more or less exposed to the solvent depending on sequence has long been considered as a possible dominant factor in the folding code,[28] but no correspondingly simple prediction scheme has previously materialized. Even within sequence-independent terms, the preeminent role played by geometrically restrictive hydrogen bonds, unspecific with respect to secondary structure, is reminiscent of the classical articles by Linus Pauling from the early 50s[29,30] but it stands out in comparison with normally used folding potentials. The distinction between sequence-dependent and sequence-independent information is not in itself original either, since it is implicit in suggestions that the sequence selects a structure from a small set of physically viable possibilities, for example, Refs. 31 and [32], and has also been considered explicitly in previous statistical prediction schemes, for example, Ref. 33. The insight that sequence-dependent information could be formed exclusively by burials, however, is a fundamental original hypothesis underlying the present scheme. This possibility had been hinted more than a decade ago by simulations of minimalist lattice models[34] and investigated more recently in the context of atomistic simulations using native burials combined to informational analyses.[13,14,16]

It is also clear, on the other hand, that our simulated trajectories are not expected to reflect details of the actual folding process, such as the time evolution of conformational averages and corresponding fluctuations, nor, even to a lesser extent, free energy barriers associated to rate limiting steps. The statistical, sequence-dependent, burial potential provides information about expected atomic central distances in the native structure, at the very end of the folding process. It might ultimately arise from a complex interaction between physical factors unlikely to be realistically modeled, along the whole process, by a single combined effective potential. Similarly,

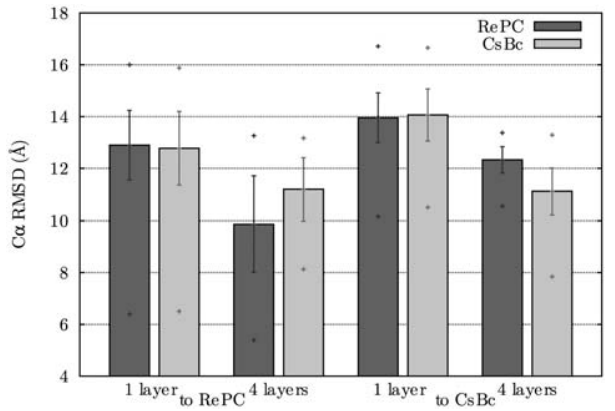


Figure 8

RMSD of one-layer and four-layer trajectories to different native structures. Simulations with a single layer, corresponding to the union of all four burial layers, were performed without the HB potential for the RePC (all-α) and CsBc (all-β) proteins, for a number of steps equivalent of 10% of the number used in the four-layer simulations described in Figure 3. The average RMSD within these trajectories with respect to both native structures were calculated and indicated in the columns labeled "1 layer" as solid bars, with standard deviation as error bars and minimal and maximal values as single points. The same comparison was also performed using the first 10% steps of two selected four-layer trajectories (when the HB potential still has a low weight). These results are shown in the columns labeled "4 layers." . [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

hydrogen bond annealing is not motivated by a putative effectively linear variation along actual folding time but is simply intended to avoid kinetic trapping, particularly for proteins containing β-sheets. Our simulations can then be seen as intended to combine correct physical ingredients in a computationally convenient order, generating an artificial path between unfolded and folded states. Realistic, path-independent, native-like conformations can still be obtained but path-dependent features might be quite unrealistic, even if possibly convenient, such as artificially fast kinetics unrelated to real folding times. Additionally, since the burial potential provides information about expected native central distances but not their expected fluctuations, simulated dynamical properties could also be unrealistic even in the native state although this possible discrepancy is not expected to affect structural prediction.

It is important, in any case, to discuss which physical factors contribute to the effective burial potential in order to understand how they can be estimated from local sequence statistics in the first place and if there are situations in which poor prediction performance could be anticipated. Side-chain hydrophobicity is expected to play an important role, since it is clearly involved in determining how different residues tend to be more or less exposed to the solvent. Accordingly, empirical distributions of atomic central distances in globular proteins were found to correlate with residue hydrophobicity.[15] Additionally, reasonable estimates for C_α central distances were obtained with an analytical polymer model that combined standard residue hydrophobicities with a simple constraint on globular size.[35] Difficulties could be readily anticipated at least in two general situations: (1) for elongated or otherwise insufficiently globular proteins, in which case solvent exposure is not expected to correlate with central distance, and (2) for constituents of macromolecular complexes possibly stabilized by hydrophobic interactions, in which case hydrophobicity is not necessarily indicative of internalization as observed in the isolated constituent protein. We presently avoid the first problem by excluding nonglobular structures. A more detailed analysis of the structures for which burials happen to be poorly assigned will be required in order to decide to what extent the second problem is affecting our current predictions.

It must also be noted that central distances might be somewhat correlated[13] but are not equivalent to other more transparent measures of solvent exposure, such as accessible surface areas, even in perfectly globular structures. Our previous observation that native-like conformations could be obtained from a sequence-compatible amount of central distance information[14] is unlikely to be valid for accessible surfaces. Intuitively, central distances appear to be more informative about the native structure, in terms of providing stronger constraints on available conformational space. Conversely, it could appear that this larger amount of information should be harder to obtain from sequence, particularly for large structures, in which case it is possible to imagine atoms with quite different central distances but equally unexposed to the solvent in terms of accessible surface. It is equally apparent, on the other hand, that chain connectivity should impose stronger correlations on central distances than on accessible surfaces, implying some extra amount of sequence-independent information in the first case. As a simple example, a chain segment that connects two regions known to be, respectively, in the most internal and most external burial layers, must necessarily cross through intermediate burial layers independently of accessible areas, or sequence. The resulting constraint might be significant, particularly for short connecting segments.

Regarding actual prediction, our previous results[16] have indicated that burials defined by central distances are not harder to obtain from sequence than solvent exposure as measured by accessible surface areas. For two burial layers of C_β atoms, for example, we have obtained a prediction accuracy close to 70%, as shown in Figure 4(b) of Ref. 16 which is comparable to reported values for accessible surface areas, for example, Ref. 36. Prediction accuracy decreases as the number of layers increases, as expected, but prediction quality, as appropriately measured by prediction information, actually increases

significantly up to at least four or five layers, as shown in Figure 6 of the same Ref. 16. Comparing our informational analysis of central distances with a similar analysis of accessible surface areas,[37] we found that single residue identities are less informative of distances than of surfaces but correlations between adjacent distances are indeed stronger.[16] Our present results displayed in Figure 2(b) show that prediction is improved for size-dependent training sets, confirming the relevance of chain length in the informational transfer between sequence and burials in our prediction scheme. At the same time, they refute the hypothesis that our current increase in prediction quality could reflect some putative intrinsic easiness of prediction for very small structures. Burial correlations, which are explicitly accounted for in the HMM, could partially explain how central distances can be "felt" differently by eventually equally unexposed chain segments. An expected dependence of burial correlation lengths on globular size is consistent with the dependence of prediction quality on the range of chain lengths in the training set.

Some practical questions also arise naturally from our results, indicating possible directions for future research. Particularly relevant, the range of applicability of the present approach in terms of the quality of burial prediction and sequence-independent constraints must also be investigated. Accordingly, the possibility of improvement in parameters associated to these two rather independent components acquire special importance. Regarding burial prediction, it is unlikely that another statistical scheme could perform significantly better than our HMM, when using the same training set of native structures, because the estimated available burial information in local amino acid sequences has been shown to be efficiently extracted by our procedure.[16] Some improvement can be expected from considering some nonlocal information, as we actually do in the present study with chain length by using only small proteins in the training set. Similarly, further improvement could be obtained if training sets more representative of the sequences to be predicted could be constructed, possibly using additional structural information that might happen to be available, such as structural class. Alternatively, a more efficient utilization of predicted burials could be attempted by some preferential utilization of more reliable predictions, as quantified by $p_i^a(\beta_p)$.

Regarding sequence-independent constraints, we have occasionally obtained conformations forming hydrogen bonds but with low fraction of standard secondary structure. This observation is suggestive that our current constraints are not always able to prevent hydrogen bond formation with non-standard backbone dihedrals. Since our hydrogen bond is defined exclusively in terms of appropriate distance and orientation between donor and acceptor, they could favor by themselves nonstandard secondary structure, such as left-handed α-helices or the γ-helix and original pleated sheet described by Pauling *et al.* in the 50s.[29,38] Left-handed α-helices are avoided by the repulsion between C_β and backbone oxygen atoms. Pauling and Corey discarded the two other, posteriorly unobserved, secondary structures in terms of unfavorable backbone dihedrals,[30] an effect not included in our current potential. It appears, therefore, that the chain might occasionally escape to protein-unlike structures still satisfying our hydrogen bond constraints, as particularly perceptible in our simulations of CsBc. This problem is likely to be aggravated as burial type prediction becomes more inaccurate and, conversely, penalization of these incorrect structures should increase the range of burial prediction accuracy still sufficient for successful folding simulations. It is not presently clear if such penalization could arise simply from a more detailed consideration of different atomic repulsion distances or if a sequence-independent torsional potential around backbone dihedrals should also be included.

CONCLUSION

Our results demonstrate the possibility of reaching and distinguishing the native-like structures using sequence-dependent burial propensities combined to sequence-independent geometrical constraints. The clear division between sequence-dependent and sequence-independent information, or between literature and grammar, will be useful for both the production of more accurate burial reading from sequence and the development of more restrictive sequence-independent grammatical constraints. The key original hypothesis that atomic burial propensities might constitute the only information to be obtained directly from sequence is corroborated and might become the basis for a new conceptual framework for improving prediction and understanding the fundamentals of protein folding.

REFERENCES

1. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science 2012;338:1042–1046.
2. Dill KA, Ozcan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys 2008;37:289–316.
3. Shakhnovich E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 2006;106: 1559–1588.
4. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–75.
5. Bryngelson JD, Onuchic J, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
6. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)-round IX. Proteins 2011;79:1–5.
7. Kinch L, Shi SY, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins 2011;79:59–73.

8. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-level characterization of the structural dynamics of proteins. Science 2010;330:341–346.
9. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science 2011;334:517–520.
10. Bowman GR, Voelz VA, Pande VS. Taming the complexity of protein folding. Curr Opin Struct Biol 2011;21:4–11.
11. Koonin EV, Wolf YI, Karev GP. The structure of protein universe and genome evolution. Nature 2002;420:218–223.
12. Vendruscolo M, Dobson C. Protein dynamics: Moore's law in molecular biology. Curr Biol 2010;21:R68–R70.
13. Pereira de Araújo AF, Gomes A LC, Bursztyn AA, Shakhnovich EI. Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. Proteins 2008;70: 971–983.
14. Pereira de Araújo AF, Onuchic JN. A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. Proc Natl Acad Sci USA 2009;106: 19001–19004.
15. Gomes ALC, de Rezende JR, Pereira de Araújo AF, Shakhnovich EI. Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. Proteins 2007;66:304–320.
16. Rocha JR, van der Linden MG, Ferreira DC, Azevêdo PH, Pereira de Araújo AF. Information-theoretic analysis and prediction of protein atomic burials: on the search for an informational intermediate between sequence and structure. Bioinformatics 2012;28:2755–2762.
17. Crooks GE, Brenner SE. Protein secondary structure: entropy, correlations and prediction. Bioinformatics 2004;20:1603–1611.
18. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
19. Larkin M, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal w and clustal x version 2.0. Bioinformatics 2007;23:2947–2948.
20. Whitford PC, Miyashita O, Levy Y, Onuchic JN. Conformational transitions of adenylate kinase: switching by cracking. J Mol Biol 2007;366:1661–1671.
21. Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. Proteins 2009;75:430–441.
22. Schrödinger LLC. The PyMOL molecular graphics system, version AQ2 1.3r1. 2010.
23. Hardin C, Eastwood MP, Prentiss MC, Luthey-Schulten Z, Wolynes PG. Associative memory Hamiltonians for structure prediction without homology: α/β proteins. Proc Natl Acad Sci USA 2003;100: 1679–1684.
24. Yang JS, Chen WW, Skolnick J, Shakhnovich EI. All-atom ab initio folding of a diverse set of proteins. Structure 2007;15:53–63.
25. Simons K, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of casp iii targets using ROSETTA. Proteins 1999;3:171–176.
26. Hubner IA, Deeds EJ, Shakhnovich EI. High-resolution protein folding with a transferable potential. Proc Natl Acad Sci USA 2005; 102:18914–18919.
27. Kabsch WSC. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
28. Dill KA. Dominant forces in protein folding. Biochemistry 1990;29: 7133–7155.
29. Pauling L, Corey RB, Branson HR The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 1951;37:205–211.
30. Pauling L, Corey RB. Configurations of polypeptide chains with favored orientations around single bonds—2 new pleated sheets. Proc Natl Acad Sci USA 1951;37:729–740.
31. Finkelstein A, Ptitsyn O. Why do globular-proteins fit the limited set of folding patterns. Prog Biophys Mol Biol 1987;50:171–190.
32. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A. Geometry and symmetry presculpt the free-energy landscape of proteins. Proc Natl Acad Sci USA 2004;101:7960–7964.
33. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
34. Pereira de Araújo AF. Folding protein models with a simple hydrophobic energy function: the fundamental importance of monomer inside/outside segregation. Proc Natl Acad Sci USA 1999;96:12482–12487.
35. England J. Allostery in protein domains reflects a balance of steric and hydrophobic effects. Structure 2011;19:967–975.
36. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
37. Crooks GE, Wolfe J, Brenner SE. Measurements of protein sequence-structure correlations. Proteins 2004;57:804–810.
38. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 1951;37:251–256.