Universidade de Brasília
Programa de Pós-graduação em Biologia Molecular

Enterramentos atômicos como possíveis intermediários informacionais no código do enovelamento proteico.

Autor:
Diogo César Ferreira

Orientador:
Antônio Francisco Pereira de Araújo

Brasília / DF
28 de Fevereiro de 2019

A Deus

e a minha família.

Agradecimentos

A Deus e a minha familia, por tudo.

Ao orientador, o professor Antônio Francisco, pela orientação e pela oportunidade de desenvolver este projeto.

Ao LBTC-UnB e aos colegas e amigos que fizeram e fazem parte dele, Lindomar, Marx, Juliana, Diogo Martins e a equipe do professor Werner, pela companhia ao longo de todos estes anos.

A Universidade de Brasília e ao Programa de Pós-Graduação em Biologia Molecular, pela estrutura.

Ao NCC/Grid-Unesp, pela disponibilização dos recursos computacionais.

A CAPES, pelo financiamento deste projeto.

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.

Enterramentos atômicos como possíveis intermediários informacionais no código do enovelamento proteico.

Diogo César Ferreira

Fevereiro de 2019

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

O código do enovelamento, ou seja, as regras através das quais a estrutura tridimensional de uma proteína está codificada em sua sequência de aminoácidos, permanece uma das grandes questões ainda não resolvidas no problema do enovelamento proteico. Neste trabalho buscamos contribuir para a compreensão do código do enovelamento partindo da hipótese de que os enterramentos atômicos, definidos como a distância dos átomos até o centro geométrico da estrutura, são o fator preponderante na codificação da estrutura terciária na sequência primária de uma proteína.

Na primeira parte deste trabalho, através de simulações de dinâmica molecular particularmente desenvolvidas para o estudo dos enterramentos atômicos, caracterizamos um conjunto de proteínas em termos da resolução mínima necessária para a descrição de suas conformações nativas através de potenciais derivados de enterramentos atômicos discretizados em camadas equiprováveis. Obtivemos um valor de 3 a 4 camadas de enterramento necessária para a descrição de algumas cadeias pequenas (de até 66 resíduos) e 4 a 5 camadas para cadeias médias (113 e 104 resíduos). Estes resultados mostram também que esta divisão em camadas equiprováveis (mesmo número de átomos por camada) implica que o número de camadas requeridas na descrição da estrutura terciária de uma proteína cresce com o seu tamanho.

Estimamos também a redundância da informação estrutural presente nas sequências de enterramentos divididos desta forma, através da capacidade das estruturas de enovelarem e se manterem estáveis ao fornecermos potenciais de enterramento errados. Estimamos uma redundância de 75% a 80% para as proteínas analisadas, o que deve ser uma medida também da proporção da informação estrutural que precisa ser efetivamente codificada na sequência na forma de enterramentos atômicos.

Na segunda parte do trabalho propomos uma nova representação de enterramentos atômicos baseada em tecnologias de codificação de sinais amplamente utilizadas na indústria, a qual chamamos de representação diferencial dos enterramentos atômicos. Aqui utilizamos ao invés de camadas as diferenças de enterramento entre átomos ou resíduos adjacentes na cadeia, o que permite uma maior resolução nos sinais de enterramento, com o uso de apenas 2 ou 3 símbolos no alfabeto da representação. Mostramos que é possível realizar predições estruturais utilizando a representação diferencial e apresentamos também um método de otimização por Monte Carlo dos sinais obtidos que permite combinar as diferentes representações de enterramento apresentadas neste trabalho.

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.

Atomic burials as possible informational intermediates in the protein folding code.

Diogo César Ferreira

February, 2019

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

The protein folding code, that is, the rules through which the three-dimensional structure of a protein is encoded in its amino acid sequence, remains one of the largely unanswered questions in the protein folding problem. In the present work we seek to contribute to the understanding of the folding code, from the hypothesis that atomic burials, defined as the distance between atoms and the structural geometric center, are the ruling factor in encoding a protein's tertiary structure in its primary sequence.

In the first part of this work, through a molecular dynamics methodology particularly developed for studying atomic burials, we characterize a set of proteins in terms of the minimal required resolution to describe their native conformations with folding potentials obtained from atomic burials represented as equiprobable layers of burials. We have found the value of 3 to 4 layers of burials required for the description of the smaller protein chains (up to 66 residues) and 4 to 5 layers for the medium sized chains (113 and 104 residues). These results also show that such equiprobable layer (same number of atoms per layer) representation implies that the amount of layers required to describe the tertiary structure of a protein chain increases with its size.

We also estimate the structural information redundancy present in the burial sequences divided in this way, by evaluating the structure's ability to fold and to remain

folded as we provide increasing fractions of incorrect burial potentials. We estimate the redundancy to be 75% to 80% for the evaluated proteins, which should also be a measurement of the proportion of structural information which must be effectively encoded in the primary sequence in the form of atomic burials.

In the second part of this work we propose a new representation scheme for atomic burials, based on signal encoding technologies widely used in telecommunications industry, which we call the differential atomic burial representation. Here, instead of layers, we use burial differences between adjacent atoms or residues in a chain, which allows for a greater resolution in the recovered burial signals while only requiring 2 or 3 symbols in the representation's alphabet. We show that it is possible to make structural predictions from sequence using the differential atomic burial representation and we also present a Monte Carlo optimization method for the obtained signals, which allows us to combine the various burial representation schemes available in this work.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

- BLAST: *Basic Local Alignment Search Tool*
- C: Carbono da carboxila de um aminoácido
- C_α: Carbono-α de um aminoácido
- C_β: Carbono-β de um aminoácido
- D_{KL}: Divergência de Kullback-Leibler
- DNA: Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*)
- HMM: Modelo Oculto de Markov (*Hidden Markov Model*)
- LL: *log likelihood*
- L_{min}: Número mínimo de camadas de enterramento necessárias para uma proteína dobrar nas simulações de enovelamento
- N: Nitrogênio da ligação peptídica de um aminoácido
- PDB: *Protein Data Bank*
- RMSD: Desvio Quadrático Médio (*Root Mean Square Deviation*)
- RNA: Ácido Ribonucleico (*Ribonucleic Acid*)

Somente na Seção 3.3 - Otimização das predições:

- D: Termo de energia da predição da representação diferencial com 2 direções
- L: Termo de energia da predição com 2 camadas
- TG: Termo de energia referente à distribuição global dos enterramentos atômicos

- 2D: Representação diferencial com 2 direções
- 2L+C_β: Representação com 2 camadas e direção de C_β
- 4L: Representação com 2 camadas

Capítulo 1

Introdução

1.1 O enovelamento proteico

As proteínas são macromoléculas biológicas que estão envolvidas na grande maioria dos processos celulares, exercendo papéis que vão desde funções estruturais e de transporte de moléculas até a catálise enzimática e controle de vias metabólicas. Para que possam desempenhar adequadamente sua função as proteínas devem passar pelo processo de enovelamento, processo através do qual uma proteína adquire uma conformação estável, bem definida e biologicamente ativa, denominada de estrutura nativa. Devido a sua ubiquidade e versatilidade, e visto que a função das proteínas está atrelada ao enovelamento, o seu estudo apresenta grande interesse e potencial aplicação em diferentes áreas como na biotecnologia, na agropecuária, na medicina e na indústria (1, 2).

Se por um lado temos que a função de uma proteína é determinada pela estrutura, por outro temos que a estrutura de uma proteína é determinada pela sua sequência de aminoácidos e a informação referente às sequências proteicas é codificada, armazenada e transmitida geneticamente através de sequências de ácidos nucleicos. Watson; Crick (3) propuseram a estrutura helicoidal da dupla fita de DNA, estabilizada por ligações de hidrogênio entre os pares de nucleotídeos que a compõem: adenina com timina e guanina com citosina. Através desse trabalho elucidou-se o pareamento das bases complementares, que é um fator preponderante na formação da estrutura tridimensional do RNA e consiste no mecanismo central dos processos de transcrição e tradução do DNA durante a síntese proteica: a partir de uma cadeia de DNA é sintetizada diretamente, por pareamento de

bases, uma fita de RNA mensageiro (transcrição). Esta, por sua vez, é traduzida em uma sequência de aminoácidos segundo um conjunto de regras conhecidas como código genético (4) que mapeiam as trincas de nucleotídeos (códons) presentes no RNA mensageiro nos aminoácidos que serão incluídos na cadeia proteica. Portanto, a física envolvida no pareamento consiste na formação de ligações de hidrogênio entre purinas e pirimidinas estabelecendo uma complementaridade entre as bases nitrogenadas que, quando associada ao código genético permite, a determinação tanto do RNA transcrito quanto da sequência proteica resultantes dos processos de transcrição e tradução puramente a partir da sequência de DNA.

Neste contexto, passou-se a buscar também uma melhor compreensão dos mecanismos envolvidos na determinação de estruturas proteicas e sua relação com as sequências de aminoácidos. Os estudos acerca de estruturas proteicas remontam à década de 1950 com a resolução das primeiras estruturas cristalinas de proteínas globulares realizadas por Kendrew et al. (5). Estes autores determinaram a conformação tridimensional da mioglobina observando que, apesar da baixa resolução, tratava-se de um arranjo molecular assimétrico e irregular, com grande teor de complexidade; resultados subsequentes (6, 7) permitiram análises mais detalhadas que vieram a corroborar também os modelos de estrutura secundária propostos por Pauling; Corey; Branson (8). Esses estudos pioneiros levantaram as principais questões sobre os princípios físicos subjacentes ao enovelamento proteico, muitas das quais permanecem amplamente não resolvidas e têm sido citadas como um dos temas científicos de maior importância na atualidade (9–11).

Um experimento particularmente relevante foi realizado por Anfinsen et al. (12), no qual efetuou-se a completa desnaturação e redução das quatro pontes dissulfeto da ribonuclease através de ureia e β-mercaptoetanol. Mostrou-se então que assim que esses agentes desnaturante e redutor eram removidos do meio, a proteína tinha a capacidade de recuperar sua atividade biológica. Uma vez que a atividade biológica está associada à conformação nativa, o experimento, que foi posteriormente reproduzido para cadeias peptídicas não sujeitas a modificações pós-traducionais intensas (13), evidenciou a espontaneidade e reversibilidade do enovelamento para o estado nativo nessas proteínas. Estes estudos constituíram a base do que posteriormente seria formulado como a **hipótese termodinâmica** do enovelamento proteico (14), sugerindo que a conformação nativa de uma proteína dentro de seu contexto fisiológico é inteiramente determinada por sua

estrutura primária, tanto na forma de interações intramoleculares quanto de interações com o solvente, que levam ao estado de menor energia livre de Gibbs dentro do espaço de conformações disponível (13).

Em contrapartida, no ano de 1969 Levinthal (15) argumentou que uma cadeia polipeptídica desnaturada não poderia se enovelar de forma aleatória dado o elevado grau de liberdade existente na cadeia, ou seja, o tempo necessário para que ela encontrasse a conformação de menor energia livre dentro do imenso espaço de conformações disponível seria incompatível com a vida. No cenário mais otimista, onde cada resíduo pudesse ser classificado como "correto" ou "errado" do ponto de vista conformacional, uma cadeia de 100 resíduos de aminoácidos precisaria explorar em média 2^{100} estados. Isto se traduz em um tempo da ordem de 10^{10} anos para encontrar o estado nativo de forma aleatória, assumindo-se uma duração de 1 picossegundo em cada conformação (13, 16). Esta suposta contradição com a hipótese termodinâmica do mínimo de energia livre global ficou posteriormente conhecida como o "paradoxo" de Levinthal (16, 17). Uma proposta de solução foi a **hipótese cinética** do enovelamento proteico, segundo a qual a estrutura nativa de um polipeptídio não seria determinada pela estabilidade, mas sim pela cinética do enovelamento, onde o espaço de busca conformacional estaria confinado a uma via rápida de enovelamento regida por restrições físicas locais. Ao final dela o estado nativo apareceria como um mínimo de energia livre não necessariamente global (13, 18).

A dicotomia entre as hipóteses termodinâmica e cinética do enovelamento deu margem à elaboração de modelos matemáticos desenvolvidos com o intuito de se avaliar os possíveis perfis de energia associados ao processo. Dill (16) e Zwanzig; Szabo; Bagchi (17) propuseram que a busca em todo o espaço conformacional não é relevante para se atingir o mínimo global de energia quando consideramos um viés energético adverso à exploração de configurações localmente desfavoráveis ao longo do enovelamento, o que permite o acesso ao estado de menor energia em tempo hábil e implica que as barreiras cinéticas estão impostas apenas sobre um espaço pequeno de conformações similares. Com isto estabeleceu-se as bases da noção de paisagem de energia e de que o enovelamento é governando por um **funil energético** (19), onde cada cadeia de uma população desenovelada corresponde a uma conformação desnaturada diferente que se enovela através de mudanças conformacionais entre estados geometricamente semelhantes que diminuem gradualmente a sua energia livre até que se atinja o mínimo global,correspondente à estrutura nativa. Posteriormente,

alterações foram propostas ao modelo, alterando sua forma para um funil "enrugado" que melhor representa as armadilhas cinéticas presentes no espaço conformacional (13, 18, 20). A teoria do funil de energia uniu as hipóteses termodinâmica e cinética em um arcabouço comum e permanece amplamente aceita como uma das representações que melhor descrevem o processo do enovelamento proteico (10, 11).

1.2 Simulações computacionais de enovelamento

Historicamente, o estudo do enovelamento proteico andou em paralelo com o advento de metodologias computacionais propostas para a investigação deste problema. Os primeiros modelos de simulação computacional consistiram em representações simples, mas que visavam modelar realisticamente a cadeia de resíduos de aminoácidos. Levitt; Warshel (21), que receberam o prêmio Nobel em 2013, publicaram a primeira simulação computacional de enovelamento, na qual utilizou-se uma representação da cadeia polipeptídica onde cada resíduo de aminoácido correspondia a um par de esferas conectadas, uma para o C_α e outra para o centroide da cadeia lateral. As ligações entre C_α adjacentes ao longo da cadeia apresentavam certa liberdade de rotação e a simulação propriamente dita consistiu na minimização da energia dos ângulos de torção destas ligações relativamente aos ângulos obtidos a partir da conformação nativa, acoplada com vibrações térmicas e com estimativas de interação com o solvente dadas pela solubilidade calculada experimentalmente para cada resíduo de aminoácido (uma vez que a simulação foi realizada sem o solvente explícito). Nesse estudo os autores conseguiram estruturas consideravelmente próximas da estrutura nativa, com desvio quadrático médio (RMSD) entre 5 Å e 7 Å, embora em muitos casos com a topologia incorreta. De todo modo, o experimento abriu as portas para a avaliação do problema do enovelamento sob uma perspectiva computacional, estabelecendo os principais elementos envolvidos em uma simulação: um modelo de representação molecular, uma estratégia para a exploração do espaço conformacional (dinâmica molecular ou Monte Carlo, por exemplo) e uma função de custo associada, na forma de uma função de energia potencial ou de um campo de força.

Posteriormente, Taketomi; Ueda; Gō (22) e Gō; Taketomi (23) publicaram um modelo ainda mais simplificado, onde uma proteína hipotética foi representada por uma cadeia de nós conectados, com 49 unidades e sem distinção de cadeia lateral, dando origem

aos chamados modelos minimalistas. Nesse estudo o espaço de conformações foi delimitado por um reticulado (***lattice model***) bidimensional e a conformação nativa foi determinada pelas interações entre pares de nós específicos em uma conformação compacta de dimensões 7×7. A exploração do espaço conformacional foi realizada através de um método de Monte Carlo, onde a energia era favorável quando nós com interações nativas encontravam-se adjacentes no reticulado bidimensional, de tal forma que o mínimo global de energia era aquele em que todas as interações nativas estavam satisfeitas. Esse potencial energético, explicitamente dependente da estrutura nativa, ficou conhecido como potencial Go (24) e serviu como base para estudos posteriores onde foram utilizados reticulados tridimensionais (cúbicos) para uma descrição abrangente das propriedades do espaço de conformações e do espaço de sequências (24, 25). Os estudos com *lattice models* e potenciais tipo Go permitiram uma compreensão básica de princípios gerais envolvidos no enovelamento proteico.

Posteriormente foi proposto (26) o conceito de "energy gap", onde evidenciou-se que o estado nativo não apenas deveria situar-se no mínimo global de energia livre, mas que também é necessário haver uma diferença significativa entre a energia da conformação nativa e a energia da conformação não-nativa ("misfolded") de menor energia. Esse conceito foi então aprimorado por Bryngelson et al. (27) para acomodar o fato de que a conformação nativa pode apresentar flutuações que não necessariamente comprometem estruturalmente ou funcionalmente a proteína. Aqui argumentou-se que a diferença de energia livre deve ser entre energia média no conjunto de conformações representando o estado nativo e a energia média no conjunto de conformações de menor energia sem uma similaridade notável com o estado nativo. Este conceito foi denominado de "stability gap", e então proposto que ele deve ser o principal fator determinante da estabilidade da conformação nativa de uma proteína (24, 25, 28). Além disso, a existência de um "stability gap" seria suficiente para que este estado seja acessível em uma dada temperatura, o que vai de acordo com a hipótese termodinâmica do enovelamento, revelando que o enovelamento é um processo cooperativo (24, 29) onde a estrutura rapidamente se dobra para uma estado semi-compacto e passa a buscar no espaço conformacional um estado de transição que permita o colapso para o estado nativo, em uma série de interações que favorecem sucessivamente o acesso à conformação nativa à medida que os contatos entre os nós são estabelecidos.

1.3 Efeito hidrofóbico e código do enovelamento

Foi sugerido por Kauzmann (30) que um dos fatores mais importantes para a estabilização da conformação nativa é a distribuição entre resíduos de aminoácidos com cadeias laterais hidrofílicas, cuja tendência é de expor-se ao solvente aquoso, e resíduos com cadeias laterais hidrofóbicas, que tendem a esconder-se do solvente, formando estruturas semelhantes às micelas presentes em soluções com detergentes. Kauzmann propôs um arcabouço termodinâmico com base em estudos acerca do **efeito hidrofóbico** (30, 31), que dizem respeito ao comportamento de moléculas hidrofóbicas em água, e concluiu que a estabilização da conformação nativa de uma proteína globular é especialmente mediada pela entropia do sistema (proteína + água), onde espera-se um ganho de entropia de 20 ordens de grandeza para cada cadeia lateral hidrofóbica que deixa o meio aquoso. Além disso, o autor aponta também para a desnaturação de proteínas em solventes apolares e, consistentemente com o que se sabe sobre o aumento da solubilidade de solutos hidrofóbicos em água a baixas temperaturas, para a menor estabilidade da estrutura nativa em temperaturas reduzidas, como evidencias da importância do efeito hidrofóbico na estabilidade das proteínas.

Durante muito tempo, entretanto, acreditou-se que não haveria uma força dominante no enovelamento proteico e que o efeito hidrofóbico teria um papel secundário no processo, sendo responsável apenas por direcionar a cadeia para uma configuração globular genérica e inespecífico demais para determinar os detalhes mais finos da estrutura terciária. Portanto, esta informação deveria ser transmitida através de interações específicas entre resíduos próximos no espaço, tais como a formação de ligações de hidrogênio e restrições de ângulos diedrais (24). Posteriormente, o advento dos modelos computacionais minimalistas em reticulados tornou possível uma enumeração completa do espaço conformacional e consequentemente uma investigação mais detalhada da termodinâmica de processos relacionados ao enovelamento proteico. Dentre os modelos desenvolvidos, podemos citar aquele conhecido como modelo HP, desenvolvido por Lau; Dill (32), onde as cadeias possuem nós que podem ser apenas de dois tipos: “H” (representando aminoácidos hidrofóbicos) ou “P” (representando os hidrofílicos, ou polares) e cujo objetivo é de se maximizar o número de contatos entre os nós hidrofóbicos (contatos H-H) e com isso estudar a estudar a importância do efeito hidrofóbico no dobramento de cadeias.

Estudos subsequentes com modelos semelhantes (33) mostraram que cadeias guiadas por um potencial hidrofóbico inespecífico (onde a contribuição de cada monômero para a energia de um contato não depende da identidade do outo monômero) podem exibir nos reticulados um enovelamento cooperativo semelhante ao de proteínas. Entretanto, a estrutura nativa alvo escolhida deve apresentar um alto grau de segregação estrutural entre os resíduos enterrados e expostos, com monômeros hidrofóbicos tendendo a ocupar posições completamente enterradas e hidrofílicos tendendo a ocupar posições completamente expostas. Estes e outros resultados de trabalhos experimentais com enovelamento de proteínas mutantes, onde resíduos foram substituídos com base em sua hidrofobicidade (34), e estudos relacionados com o comportamento de polímeros em soluções (35) também ajudaram a fortalecer a noção de que o enovelamento proteico pode ser principalmente mediado por um colapso determinado pelo efeito hidrofóbico, sendo ele não apenas suficiente, mas também o principal fator determinante da formação e estabilização da estrutura secundária, ligações de hidrogênio e dos contatos presentes na estrutura nativa. Além disso, ao contrário das restrições físicas locais independentes da sequência (ângulos diedrais, formação de ligações de hidrogênio e restrições estereoquímicas), o efeito hidrofóbico é um fenômeno dependente da sequência primária, o que aponta para o mesmo como o fator dominante também em um possível **código do enovelamento**, que seriam as regras através das quais a informação estrutural de uma proteína é codificada em sua sequência primária. Alguns autores se referem ao código do enovelamento como a "segunda metade" do código genético (24), entretanto, ao contrário deste, os detalhes precisos do código do enovelamento permanecem uma das questões ainda a serem solucionadas no âmbito do enovelamento proteico.

1.4 Enterramentos atômicos

Pela ação do efeito hidrofóbico, observamos a tendencia de resíduos hidrofílicos encontrarem-se expostos ao solvente e resíduos hidrofóbicos encontrarem-se enterrados na estrutura de uma proteína globular. Assim, a partir deste princípio simples, com o objetivo de investigarmos um código do enovelamento, torna-se interessante desenvolvermos uma medida quantitativa do grau de enterramento dos átomos de um modelo, a qual nos possibilite estabelecer uma relação entre a sequência de aminoácidos e a estrutura

tridimensional das proteínas. Neste contexto, é desejável que ela esteja correlacionada ao efeito hidrofóbico e que, sob uma perspectiva informacional, satisfaça simultaneamente duas condições: primeiro, a medida deve conter informação suficiente para codificar a estrutura tridimensional daquela proteína; segundo, a medida precisa ser suficientemente compacta de modo que a estrutura primária de uma proteína possa codificá-la. Nosso grupo então propôs o uso das distâncias entre os átomos e o centro geométrico da estrutura como uma forma de se quantificar seus enterramentos atômicos (Figura 1.1) e utilizaremos a partir de agora esta definição de enterramento atômico.

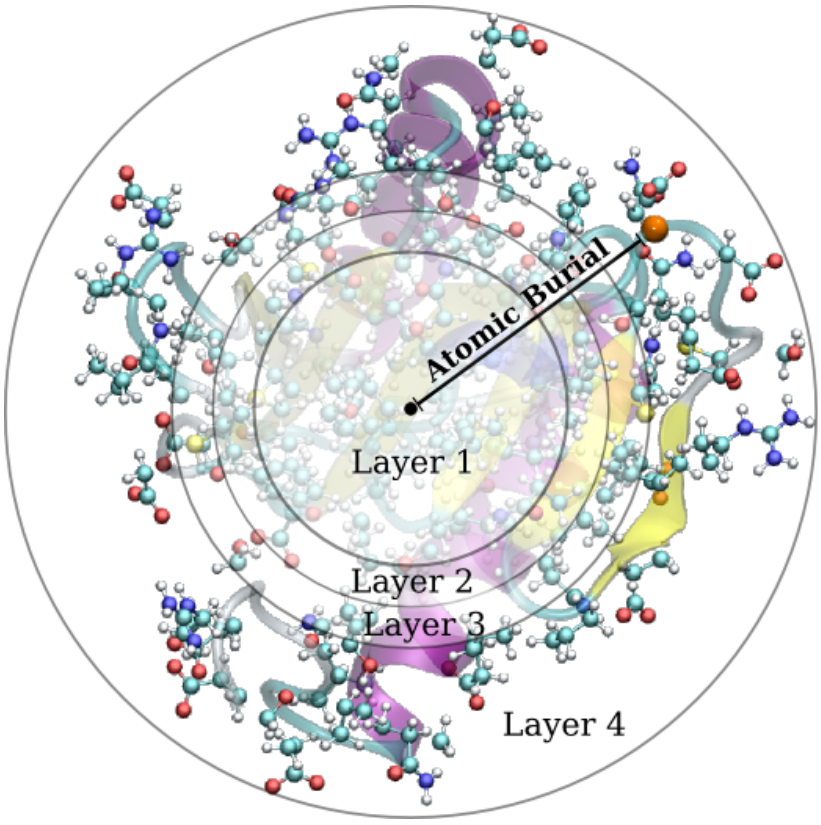

Figura 1.1. O enterramento de um átomo é definido pela sua distância até o centro geométrico da estrutura. Uma possível forma de se discretizar essas distâncias é a utilização de camadas de enterramento. Nesta representação, a largura das camadas é propositalmente desigual pois a divisão é feita de modo que cada camada contenha o mesmo número de átomos (imagem de Linden et al. (36)).

Os enterramentos atômicos assim definidos foram utilizados em um trabalho inicial (37) onde os átomos (exceto hidrogênio) de cadeias provenientes de um banco de dados de estruturas globulares foram ordenados, agrupados e contados conforme seus valores de enterramento atômicos (R_a) obtidos por

$$R_a = \sqrt{(x_a - x_0)^2 + (y_a - y_0)^2 + (z_a - z_0)^2} \tag{1.1}$$

dada uma estrutura de N átomos com coordenadas (x_a, y_a, z_a) e centro geométrico (x_0, y_0, z_0). Isto possibilitou o ajuste de uma função de densidade de probabilidade que descreve a distribuição dos enterramentos atômicos naquele banco de dados, definida por

$$P(r) = \frac{Ar^2}{1 + e^{\beta(r-\mu)}} \tag{1.2}$$

que consiste no produto de Ar^2, que representa a variação de volume das camadas esféricas de raio r, com a função de densidade atômica de Fermi $\frac{1}{1+e^{\beta(r-\mu)}}$ (38). Com os parâmetros $A = 1.73$, $\beta = 9.37$ e $\mu = 1.17$, ela é uma função de densidade de probabilidade de enterramentos atômicos no intervalo $0 \leq r \leq 2$, onde $r = \frac{R_a}{R_g}$ é o enterramento de um átomo (R_a) normalizado pelo raio de giro (R_g) da estrutura (Figura 1.2), obtido por

$$R_g = \sqrt{\frac{1}{N}\sum_{a=1}^{N}(x_a - x_0)^2 + (y_a - y_0)^2 + (z_a - z_0)^2} \tag{1.3}$$

Esta distribuição representa razoavelmente bem não só o banco de dados total mas também proteínas individuais, o que a torna particularmente útil para estudos envolvendo os enterramentos atômicos como será apresentado nas partes apropriadas desta tese. Gomes et al. (37) também analisaram as distribuições de enterramentos com base na identidade dos resíduos aos quais os átomos pertencem, que revelou uma escala de hidrofobicidade compatível, a despeito de algumas diferenças, com outras escalas de hidrofobicidade disponíveis na literatura. Esses resultados corroboram o uso desta medida de enterramento atômico como uma forma simples de aproximar a exposição ao solvente, que é diretamente associada ao efeito hidrofóbico.

Posteriormente foi mostrado que podemos derivar potenciais baseados nos enterramentos atômicos da estrutura nativa de uma proteína. Quando associados a restrições baseadas na formação de ligações de hidrogênio, os potenciais de enterramento são suficientes para se determinar a conformação nativa destas proteínas em simulações de computacionais de Monte Carlo (39) e de dinâmica molecular (40) partindo de conformações desnaturadas. Aqui os enterramentos nativos também foram discretizados em camadas equiprováveis, ou seja, de tal forma que cada camada continha o mesmo número de átomos (Figura 1.1) e ainda assim os potenciais derivados de camadas de enterramentos permaneceram suficientes para determinar a estrutura nativa, desde que a divisão fosse feita em 3 ou mais camadas. Nesses estudos, os potenciais de enovelamento contavam apenas com termos de energia baseados nos enterramentos atômicos e restrições físicas

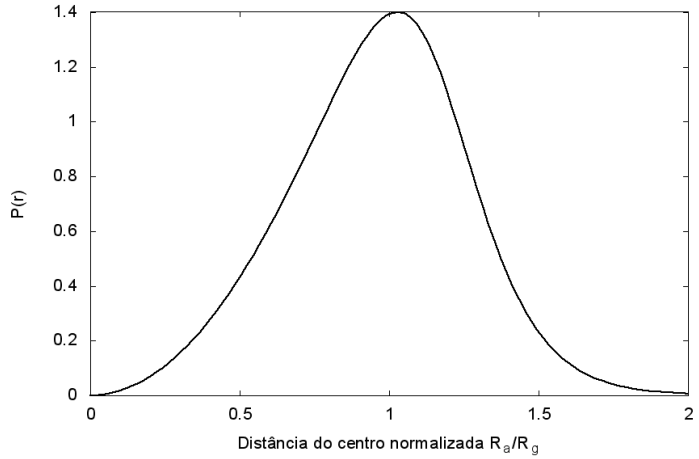

Figura 1.2. Curva de distribuição dos enterramentos atômicos com parâmetros $A = 1.73$, $\beta = 9.37$ e $\mu = 1.17$, para os quais a curva se comporta como uma função de densidade de probabilidades no intervalo $0 \leq r \leq 2$, onde $r = \frac{R_a}{R_g}$ é o enterramento de um átomo (R_a) relativo ao raio de giro (R_g) da estrutura. As distâncias em *angstroms* podem ser obtidas multiplicando-se r por R_g. Esta curva pode ser usada na definição dos limites de camadas de enterramento, já que a integral desta curva nos fornece a probabilidade de uma camada delimitada pelo intervalo de integração.

apropriadas, ou seja, ligações covalentes e ligações de hidrogênio. O fato de que nenhuma outra informação estrutural nativa foi fornecida para as simulações é um indicativo de que os enterramentos atômicos não somente contêm a informação necessária para a descrição da estrutura nativa, como também é possível, obtermos a estrutura nativa de uma proteína em simulações caso conheçamos seus enterramentos atômicos (Figura 1.3, item 3).

Com isto temos uma medida correlacionada ao efeito hidrofóbico e que satisfaz a condição de conter informação suficiente para que a estrutura nativa seja corretamente determinada. Resta avaliarmos se a estrutura primária pode de fato codificar a informação estrutural através dos enterramentos atômicos. De um ponto de vista quantitativo é razoável pensarmos que se isto for verdade, então a entropia da sequência primária, uma medida proposta por Shannon (41) relacionada com a quantidade de informação produzida por uma variável, deve ser um limite para a quantidade de informação estrutural disponível para ser obtida por meio de um esquema de predição de enterramentos a partir da sequência. Realizamos então uma análise (artigo em anexo (42)) baseada na Teoria da Informação acerca das correlações entre a sequência de aminoácidos e a sequência de enterramentos associados em um banco de proteínas não redundante (43). Nesse estudo mostramos primeiramente que a quantidade de informação por símbolo nas sequências de enterramento,

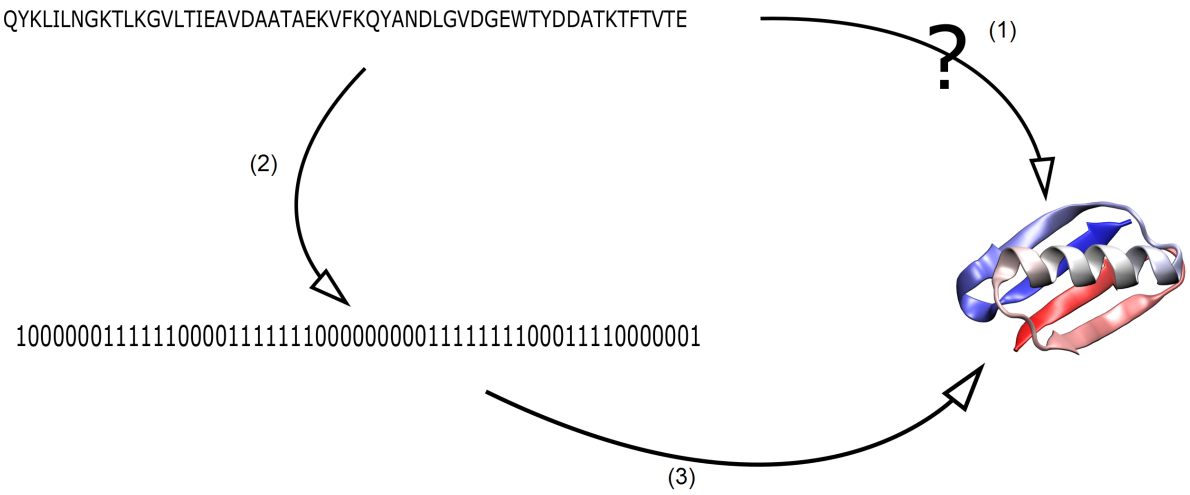

Figura 1.3. (1) Os mecanismos pelos quais a estrutura primária codifica a estrutura terciária, ou seja, o código do enovelamento não é plenamente conhecido. (2) Propomos então que a estrutura primária codifica a informação referente aos enterramentos atômicos, aqui representados como uma sequência de duas camadas de enterramento. (3) A partir dos enterramentos atômicos, podemos chegar na estrutura terciária de uma proteína através de simulações de dinâmica molecular.

ou seja, sua densidade de entropia (41, 44), é de aproximadamente 0.62 *bits/resíduo* ou 0.93 *bits/resíduo* quando consideradas, respectivamente, representações em 2 ou 3 camadas equiprováveis para os C_α. Estes valores são significativamente menores do que a densidade de entropia estimada para sequências de aminoácidos, que é de 4.2 *bits/resíduo* (45), o que inicialmente já confere à estrutura primária a capacidade de codificar por completo ao menos os enterramentos referentes aos C_α nessas representações de enterramento.

Nesse estudo também propusemos um método de predição de enterramentos a partir da sequência primária baseado em Modelos Ocultos de Markov (*Hidden Markov Models* - HMM), que são modelos estatísticos que encontram diversas aplicações em análises de sequências biológicas como predições estruturais de ácidos nucleicos, de estrutura secundária de cadeias proteicas e alinhamento de sequências(46). O programa é uma generalização do algoritmo de predição de estrutura secundária proposto por Crooks; Brenner (45) e pode ser usado para realizar predições de sequências de símbolos associados a cadeias de resíduos de aminoácidos, com especial suporte para representações de enterramentos atômicos. Essas predições não utilizam nenhuma informação adicional além da própria sequência de aminoácidos, o que as torna particularmente úteis na análise do papel dos enterramentos no contexto aqui apresentado. A partir de algumas das predições derivamos potenciais de enovelamento que foram posteriormente utilizados em simulações de dinâmica molecular (artigo em anexo (36)). Notamos aqui que as melhores predições foram capazes

de gerar potenciais que descreveram com sucesso a estrutura nativa de suas cadeias. Esses resultados mostram que não somente é possível que os enterramentos estejam codificados na sequência de aminoácidos como também, no caso das melhores predições, é possível extrair informação estrutural a partir da sequência de resíduos de aminoácidos a fim de caracterizar a estrutura terciária nativa destas sequências através dos enterramentos atômicos.

Neste contexto, enfrentamos atualmente o desafio de desenvolver esquemas capazes de prever de maneira mais eficiente a sequência nativa de enterramentos a partir da sequência primária de aminoácidos. Sabemos que há um grande volume de proteínas para as quais a predição dos enterramentos atômicos não é acurada, de tal forma que os potenciais de enterramento gerados a partir delas não é suficiente para obtermos a estrutura nativa em simulações de enovelamento. É razoável pensarmos que, ao fornecermos potenciais errados em uma simulação, estamos fornecendo uma quantidade menor da informação estrutural daquela cadeia. Portanto, neste trabalho buscaremos quantificar uma medida que nos possibilite estimar a quantidade de informação de enterramentos atômicos necessária para caracterizarmos com sucesso a estrutura terciária de proteínas. Proporemos também um novo esquema de representação de enterramentos atômicos que possibilite a realização de predições mais precisas. Esperamos com este trabalho não só avançar os conhecimentos acerca do papel dos enterramentos atômicos no enovelamento proteico, mas também contribuir para a compreensão do código do enovelamento como um todo.

1.5 Objetivos

Objetivo geral

O objetivo geral deste trabalho é contribuir para a compreensão do **código do enovelamento** e seu papel no contexto do problema do enovelamento de proteínas, uma vez que esta permanece uma das frentes amplamente não resolvidas deste problema. Utilizaremos para este fim a hipótese de que os enterramentos atômicos são o fator central na codificação da informação estrutural na sequência primária, dentro da perspectiva de que eles contêm predominantemente a informação dependente da sequência que é necessária e suficiente para a descrição da estrutura tridimensional de cadeias proteicas.

Objetivos específicos

1. Corroborar a generalidade da suficiência da informação sobre enterramentos atômicos, na forma de distâncias ao centro geométrico discretizadas em um pequeno número de camadas, para a determinação da estrutura terciária de proteínas de diferentes tamanhos e classes estruturais.

2. Verificar a resolução necessária, expressa em número de camadas, e uma possível dependência com o tamanho e/ou classe estrutural para a descrição da estrutura terciária através dos enterramentos.

3. Verificar a acurácia necessária para a descrição da estrutura através dos enterramentos atômicos, através da redundância detectável imposta pelas restrições independentes da sequência presentes no modelo proteico utilizado.

4. Estimar a fração de proteínas para as quais as predições de enterramento a partir da sequência apresenta a acurácia necessária para uma resolução apropriada. Para este fim, utilizaremos predições obtidas a partir de um algoritmo baseado em um *Hidden Markov Model* (HMM) previamente proposto.

5. Explorar a possibilidade de se utilizar uma representação diferencial dos sinais de enterramentos para se obter uma maior resolução na descrição dos enterramentos, correspondendo a um número maior de camadas, o que deve ser particularmente relevante para proteínas de tamanho médio e grande, sem aumentar o número de símbolos da descrição.

Capítulo 2

Resolução e redundância

Neste capítulo descreveremos brevemente os experimentos realizados em conformidade com os objetivos propostos nos itens de 1 a 4 deste trabalho. A metodologia utilizada e os resultados obtidos estão descritos detalhadamente no artigo intitulado *Information and redundancy in the burial folding code of globular proteins within a wide range of shapes and sizes* (47) incluso neste capítulo, publicado em 2016 como parte da pesquisa desempenhada no desenvolvimento deste estudo.

Nesta parte do trabalho, queremos quantificar a resolução necessária para que uma dada representação de enterramentos atômicos discretizada em camadas equiprováveis descreva com sucesso a estrutura terciária de uma cadeia proteica. Para tanto caracterizaremos, em termos de estabilidade e de acessibilidade, um conjunto de proteínas selecionadas de modo a abranger diferentes classes estruturais e faixas de tamanho. Definimos **estabilidade** como a capacidade de uma estrutura se manter na conformação nativa em simulações cujo estado inicial é a própria conformação nativa e definimos **acessibilidade** como a capacidade de uma cadeia atingir a conformação nativa em simulações cujo estado inicial é uma conformação desnaturada. Em outras palavras, buscaremos quantificar o menor número necessário L_{min} de camadas de enterramento na representação de uma estrutura para que o estado nativo de uma proteína seja acessível e estável. Como a largura das camadas diminui com o número de camadas utilizadas na representação, teremos com isto uma medida da resolução requerida para que uma sequência de camadas de enterramento resulte no dobramento correto de uma dada estrutura.

Em seguida quantificaremos a redundância da informação estrutural fornecida pelos

enterramentos atômicos na forma do erro tolerado pelas estruturas quando fornecemos informação incompleta ou errada. Caracterizaremos também em termos de estabilidade e de acessibilidade, o comportamento das estruturas ao serem simuladas com frações crescentes de erro nos potenciais de enovelamento fornecidos, expressas na forma de potenciais removidos ou potenciais inconsistentes com o enterramento nativo para um número crescente de átomos na cadeia. A quantidade de erro tolerado pela estrutura, ou seja, a fração máxima de potenciais errados fornecidos para as simulações em que a estrutura ainda é capaz de se atingir e se manter estável no estado nativo nos dá uma estimativa direta do quanto da informação estrutural dada pelos enterramentos é redundante e já decorre de fatores não dependentes da sequência como restrições físico-químicas nos sistemas proteicos.

2.1 Metodologia das simulações de enovelamento

Esta parte do trabalho depende em grande parte da realização de simulações de enovelamento através de um algoritmo de dinâmica molecular o qual descreveremos em detalhes nesta seção pois ele apresenta termos de energia particularmente desenvolvidos para abordar a questão dos enterramentos atômicos no processo de enovelamento. O algoritmo de dinâmica descrito a seguir foi implementado por Whitford et al. (48) na linguagem Fortran. Os termos específicos para o tratamento dos enterramentos atômicos foram adicionados posteriormente por Pereira de Araujo; Onuchic (40), conforme descrito no trabalho onde esta metodologia foi primeiramente descrita. O algoritmo modificado foi denominado MDBury e também foi um dos objetos da publicação (36) e da tese de doutorado de Linden (49), que foi o principal mantenedor do código durante o período do desenvolvimento de seu projeto.

Em uma simulação de dinâmica molecular, os átomos são representados como um conjunto de corpos pontuais em um espaço cartesiano. A partir de uma função de energia potencial, deriva-se as forças atuantes sobre os átomos em cada iteração, fazendo com que o sistema explore o espaço conformacional com o objetivo de se atingir uma configuração de energia potencial reduzida. A função de energia potencial implementada pelo MDBury não visa ser uma função fisicamente realista, mas sim permitir a aplicação do conceito dos enterramentos atômicos, de tal forma que o mínimo de energia seja suficientemente

próximo da conformação nativa da cadeia proteica simulada. Com isto, pretende-se tanto reduzir o tempo de processamento computacional quanto possibilitar uma separação bem definida entre termos de energia independentes da sequência primária e termos dependentes da sequência.

Os termos independentes da sequência modelam o conjunto mínimo de forças necessárias para que se mantenha a geometria de estruturas proteicas de forma genérica, sendo incapazes de permitir uma distinção entre sequências diferentes e entre uma configuração nativas ou não nativa. Por outro lado, o termo de energia dependente da sequência é derivado dos enterramentos atômicos os quais, pela hipótese principal deste trabalho, estão diretamente associados à estrutura primária de uma cadeia proteica. Sendo assim, esta separação nos permitirá, por sua vez, uma validação desta hipótese no que tange ao papel dos enterramentos no código do enovelamento. A função de energia potencial utilizada pelo MDBury é composta pela soma dos seguintes termos independentes:

$$V = V_{\text{ligações}} + V_{\text{ângulos}} + V_{\text{diedrais}} + V_{\text{repulsão}} + \underline{V_{\text{hidrogênio}} + V_{\text{enterramentos}}} \tag{2.1}$$

Os dois últimos termos destacados foram aqueles desenvolvidos especificamente para o algoritmo MDBury (dinâmica usando enterramentos atômicos), enquanto os outros apresentam características similares a termos de energia comumente utilizados em simulações convencionais (exceto diedrais, como veremos adiante).

Termos ligantes Os três primeiros termos da energia potencial são $V_{\text{ligações}}$, $V_{\text{ângulos}}$ e V_{diedrais} visam manter uma geometria realista com base nas ligações covalentes entre os átomos da proteína onde as entidades são como massas pontuais unidas por molas (potenciais de Hooke). Formalmente, os termos são:

$$V_{\text{ligações}} = \sum_{\text{ligações}} k_d(d - d_0)^2 \tag{2.2}$$

$$V_{\text{ângulos}} = \sum_{\text{ângulos}} k_\theta(\theta - \theta_0)^2 \tag{2.3}$$

$$V_{\text{diedrais}} = \sum_{\text{diedrais}} k_\chi(\chi - \chi_0)^2 \tag{2.4}$$

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 2.1). Os mínimos de energia destes termos se localizam nas conformações estruturais em que estas distâncias e ângulos

se encontram em seus valores ótimos, d_0, θ_0 e χ_0, obtidos pelo programa a partir do observado em estruturas estendidas padronizadas. As constantes de mola têm os valores $k_d = 100\epsilon\text{Å}^{-2}$, $k_\theta = 20\epsilon\text{rad}^{-2}$, $k_\chi = 10\epsilon\text{rad}^{-2}$ onde ϵ é a nossa unidade de energia. Estes valores são definidos de modo que quando as distâncias ou ângulos desviarem dos valores ótimos, o custo energético seja consistentemente alto em relação aos outros termos do potencial.

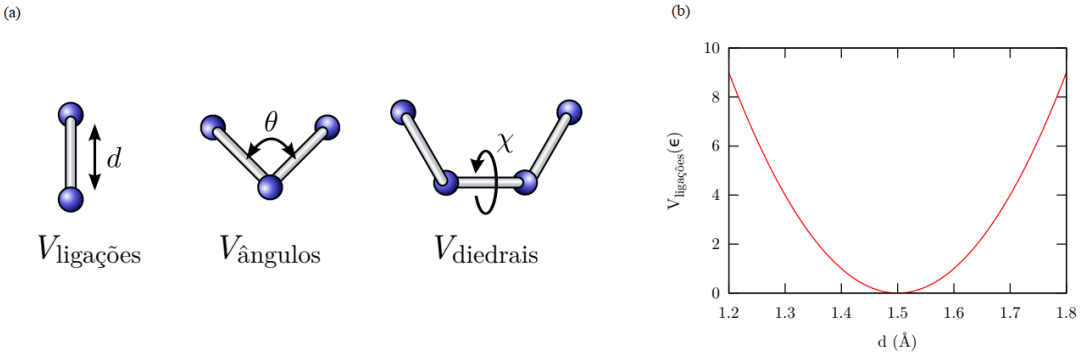

Figura 2.1. (a) Termos ligantes presentes no algoritmo. (b) Forma do potencial $V_{ligações}$, que mantém uma distância ótima nas ligações covalentes, neste caso 1.5 Å. Os outros dois termos ligantes têm uma forma análoga (49).

Os termos ligantes são aplicados apenas em conjuntos de átomos cuja configuração relativa se deseja manter aproximadamente constante ao longo da simulação: a equação 2.2 é aplicada a todos os pares de átomos unidos covalentemente; a equação 2.3 é aplicada a todos os vértices formados por três átomos unidos covalentemente; por fim, a equação 2.4 é aplicada aos ângulos diedrais que precisam permanecer fixos, como a ligação peptídica, os grupos planares das cadeias laterais dos resíduos aromáticos e nos diedrais formados pelos átomos N, C_α, C e C_β, de modo a manter a quiralidade correta da cadeia lateral. Em particular, entretanto, nenhuma força é aplicada diretamente aos ângulos ϕ e ψ a fim de se forçar qualquer configuração referente à conformação nativa.

Termo de repulsão atômica A repulsão entre orbitais de átomos próximos no espaço (repulsão de Pauli) é modelada pela seguinte equação:

$$V_{\text{repulsão}} = \sum_{\text{pares de átomos}} \epsilon_{\text{rep}} \left(\frac{\sigma_{\text{rep}}}{d} \right)^{12} \tag{2.5}$$

onde d é a distância entre dois átomos e $\epsilon_{\text{rep}} = 1.0\epsilon$. Esta equação corresponde à parte repulsiva do potencial conhecido como Lennard-Jones, tipicamente utilizado em dinâmica molecular para modelar a repulsão de Pauli (Figura 2.2). A implementação não inclui a

parte atrativa desse potencial, que seria utilizada para modelar a atração de van der Waals entre pares de átomos não ligados covalentemente.

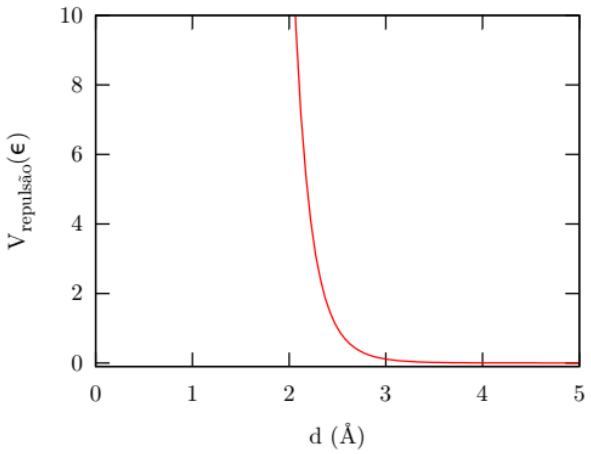

Figura 2.2. Forma do potencial $V_{repulsão}$, que modela a repulsão de Pauli entre átomos próximos (49).

O termo de repulsão atômica é aplicado a todos os pares de átomos que estão separados por mais de duas ligações covalentes e não pertencem ao mesmo diedral. 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, onde o valor é de $\sigma_{\text{rep}} = 3$Å. Esse último ajuste foi necessário para impedir o surgimento de algumas estruturas secundárias não realistas, formadas em simulações nas quais todos os átomos tinham o mesmo raio atômico.

Termo das ligações de hidrogênio Os modelos utilizados nas simulações de dinâmica molecular realizadas neste trabalho não incluem as moléculas do solvente no qual as proteínas estariam presentes e tampouco os átomos de hidrogênio nas moléculas das próprias cadeias proteicas. Contudo, as interações entre a proteína e o solvente e a formação de ligações de hidrogênio entre átomos de sua cadeia principal são essenciais tanto na compactação da cadeia quanto no surgimento e na estabilização de estrutura secundária durante o enovelamento, de modo que sem elas os enterramentos atômicos são insuficientes para que as cadeias simuladas dobrem corretamente, implicando que as ligações de hidrogênio não possam ser ignoradas nas simulações. Portanto, aqui as ligações de hidrogênio são modeladas como forças que atuam diretamente sobre os átomos de oxigênio e nitrogênio da cadeia principal.

A ruptura de ligações de hidrogênio com o solvente é modelada como um potencial que aplica uma penalidade de ϵ_{hb} (medida em ϵ, nossa unidade de energia) para todos os pares de potenciais doadores (N) e aceptores (O) da cadeia principal se eles estiverem enterrados a uma distância menor que μ_r do centro da estrutura sem formar ligações de hidrogênio. Formalmente, este termo é definido por:

$$V_{hidrogênio} = \sum_{\text{doadores/aceptores}} \epsilon_{hb} f(r, \Lambda) \tag{2.6}$$

$$f(r, \Lambda) = \begin{cases} F(r)(1 - \Lambda) & \text{para } \Lambda \leq 0.95 \\ 0 & \text{para } \Lambda > 1.05 \end{cases} \tag{2.7}$$

com uma região quadrática intermediária entre $0.95 < \Lambda \leq 1.05$ a fim de se manter a diferenciabilidade em todo o intervalo, Λ é o número total de ligações de hidrogênio formadas por um determinado átomo doador ou aceptor e $F(r)$ é uma função de Fermi dependente do enterramento atômico r definida como:

$$F(r) = \frac{1}{1 + e^{\beta_r(r - \mu_r)}} \tag{2.8}$$

que varia abruptamente de 1 para 0 em um intervalo controlado por β_r em torno de $r = \mu_r$. Aqui, utilizamos $\beta_r = 10\text{Å}^{-1}$ e $\mu_r = 15\text{Å}$ (Figura 2.3-a).

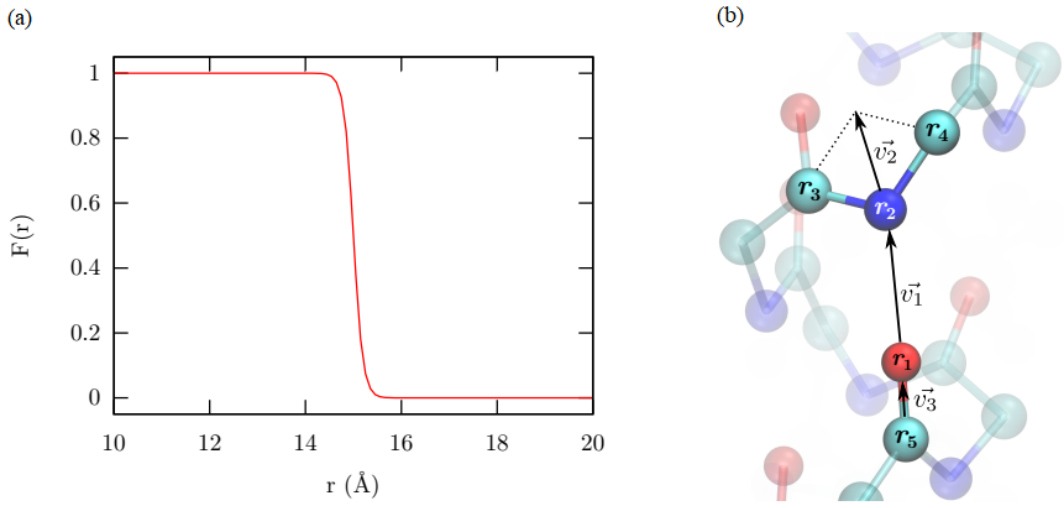

Figura 2.3. Componentes do potencial $V_{hidrogênio}$, aplicado sobre um átomo enterrado a uma distância r menor do que μ_r do centro da estrutura sem formar ligações de hidrogênio (49). (a) Forma da função $F(r)$ para $\beta_r = 10\text{Å}^{-1}$ e $\mu_r = 15\text{Å}$. (b) O critério para a determinação da formação de uma ligação de hidrogênio depende das coordenadas atômicas e dos vetores formados pelas posições relativas entre os átomos envolvidos.

O número de ligações de hidrogênio (Λ_i) formadas por um doador i é calculada por:

$$\Lambda_i = \sum_j F(h_j)F(\eta_j)F(\theta_j) \tag{2.9}$$

para todos os possíveis aceptores j. Aqui, os valores h, η e θ são definidos em termos das coordenadas atômicas de cinco átomos participantes da ligação: o oxigênio aceptor da carbonila ($\boldsymbol{r}_1$), o nitrogênio doador ($\boldsymbol{r}_2$), os dois átomos de carbono adjacentes a este nitrogênio ($\boldsymbol{r}_3$, $\boldsymbol{r}_4$) e o átomo de carbono adjacente ao oxigênio ($\boldsymbol{r}_5$), conforme mostra a Figura 2.3-b. Essas coordenadas definem três vetores, $\boldsymbol{v}_1 = \boldsymbol{r}_2 - \boldsymbol{r}_1$, $\boldsymbol{v}_2 = \boldsymbol{r}_3 + \boldsymbol{r}_4 - 2\boldsymbol{r}_2$ e $\boldsymbol{v}_3 = \boldsymbol{r}_1 - \boldsymbol{r}_5$. Definimos então $h = |\boldsymbol{v}_1|$, a norma de $\boldsymbol{v}_1$, η é o ângulo entre $\boldsymbol{v}_1$ e $\boldsymbol{v}_2$ e θ é o ângulo entre $\boldsymbol{v}_1$ e $\boldsymbol{v}_3$.

Nas simulações aqui realizadas, quando a estrutura inicial é a estrutura desnaturada, é aplicado um procedimento de *annealing*, onde o valor ϵ_{hb} varia de 0ϵ até 5ϵ ao longo de toda a trajetória. Isto é feito com o objetivo de se evitar a formação prematura de estrutura secundária, prendendo a cadeia em um mínimo local. Nas simulações em que a estrutura inicial é a conformação nativa, o *annealing* não é feito e $\epsilon_{hb} = 5\epsilon$ em toda a trajetória. Não foram incluídos nas simulações os doadores e aceptores presentes nas cadeias laterais dos aminoácidos.

Termo dos enterramentos atômicos O termo dos enterramentos atômicos é dado pela seguinte equação:

$$V_{\text{enterramentos}} = \sum_i B(r_i) \tag{2.10}$$

onde r_i é o enterramento do átomo i, ou seja, sua distância até o centro geométrico da estrutura. A função $B(r_i)$ tem valor zero quando o enterramento do átomo se aproxima de um intervalo de tolerância de tamanho $2\delta_i$ em torno de seu valor esperado de enterramento r_i^*, e cresce linearmente quando o átomo sai do intervalo de tolerância, que é dado por $(r_i^* - \delta_i, r_i^* + \delta_i)$. Além disto, uma pequena região de quadrática δ_q de tamanho $\delta_q = 0.5$Å é aplicada em torno desse intervalo a fim de se manter a diferenciabilidade da função em todos os pontos. Desta forma, átomos que estão dentro de seu intervalo especificado não sofrem nenhuma força oriunda deste termo, enquanto aqueles que estão fora sofrem uma

força constante de $1\epsilon\text{Å}^{-1}$ em direção a r_i^*. A função $B(r_i)$ é definida como:

$$B(r_i) = \begin{cases} -a_1 r^2 + b_1 & \text{para } r \leq r_1 \\ -a_2 r + b_2 & \text{para } r_1 < r \leq r_2 \\ a_3(r - r_3)^2 & \text{para } r_2 < r \leq r_3 \\ 0 & \text{para } r_3 < r \leq r_4 \\ a_4(r - r_4)^2 & \text{para } r_4 < r \leq r_5 \\ a_5 r - b_5 & \text{para } r > r_5 \end{cases} \tag{2.11}$$

com $r_1, ..., r_5, a_1, ..., a_5, b_1, ..., b_5 \geq 0$ definidos em termos de r_i^*, δ_i e δ_q para diferentes átomos i.

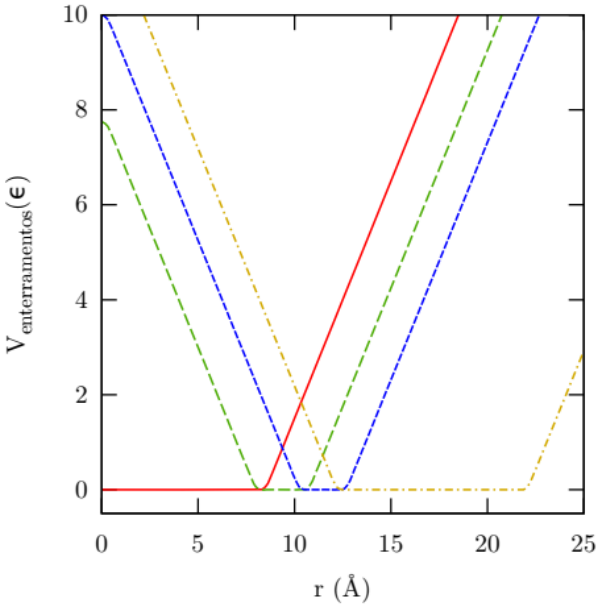

Figura 2.4. (a) Forma do potencial $V_{enterramentos}$. Aqui está exemplificado as formas do potencial para átomos classificados em 4 camadas de enterramento equiprováveis.

A Figura 2.4 mostra a forma deste potencial no caso em que os átomos são classificados em 4 camadas de enterramento equiprováveis. Neste exemplo, os tipos possíveis de classificação são os seguintes pares de valores normalizados $\left(\frac{r_i^*}{R_g}, \frac{\delta_i}{R_g}\right)$: $(0.378, 0.378)$ para a camada mais interna, $(0.859, 0.103)$, $(1.051, 0.089)$ e $(1.570, 0.430)$ para a camada mais externa. Estes valores são obtidos a partir da Equação 1.2 (Figura 1.2) e correspondem aos intervalos de integração da curva $\left(\frac{r_i^* - \delta_i}{R_g}, \frac{r_i^* + \delta_i}{R_g}\right)$ sem sobreposição que resultam no valor de 0.25 (para 4 camadas). O denominador R_g é o raio de giro da conformação nativa daquela estrutura.

As simulações presentes no artigo a seguir envolvem a manipulação deste termo da energia potencial de diferentes maneiras. Utilizar um número diferente de camadas significa modificar estes valores para o número respectivo, de forma que ainda seja mantida a propriedade de equiprobabilidade das camadas. Nos experimentos em que o potencial de enterramento de alguns átomos é apagado significa que estes átomos recebem o par de valores $(0, 2)$ e portanto não sofrerão força alguma proveniente deste termo de energia dentro dos limites determinados pelo dobro do raio de giro da estrutura nativa. Nos experimentos em que o potencial de enterramento de alguns átomos é embaralhado, significa que estes átomos são reclassificados ao acaso para algum dos outros tipos disponíveis conforme aquele número de camadas. Em todos os casos, a temperatura média do sistema é mantida constante e igual a 1ϵ através de um termostato de Berendsen (50).

Alguns aspectos da metodologia de simulação foram descritos nesta Seção uma vez que se trata de uma metodologia de dinâmica com componentes não convencionais e que não está totalmente especificada no artigo a seguir. A metodologia específica desta parte do trabalho e os resultados decorrentes destes experimentos estão descritos por completo e serão agora apresentados no artigo científico incluído a seguir, e publicado como parte da pesquisa desempenhada ao longo deste projeto.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Information and redundancy in the burial folding code of globular proteins within a wide range of shapes and sizes

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

[1] Laboratório de Biofísica Teórica e Computacional, Departamento de Biologia Celular, 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 and Departments of Physics and Astronomy, Chemistry and Biosciences Rice University, 6100 Main Street, Houston, Texas 77005

ABSTRACT

Recent *ab initio* folding simulations for a limited number of small proteins have corroborated a previous suggestion that atomic burial information obtainable from sequence could be sufficient for tertiary structure determination when combined to sequence-independent geometrical constraints. Here, we use simulations parameterized by native burials to investigate the required amount of information in a diverse set of globular proteins comprising different structural classes and a wide size range. Burial information is provided by a potential term pushing each atom towards one among a small number L of equiprobable concentric layers. An upper bound for the required information is provided by the minimal number of layers $L^{\min}$ still compatible with correct folding behavior. We obtain $L^{\min}$ between 3 and 5 for seven small to medium proteins with $50 \leq N_r \leq 110$ residues while for a larger protein with $N_r = 141$ we find that $L \geq 6$ is required to maintain native stability. We additionally estimate the usable redundancy for a given $L \geq L^{\min}$ from the burial entropy associated to the largest folding-compatible fraction of "superfluous" atoms, for which the burial term can be turned off or target layers can be chosen randomly. The estimated redundancy for small proteins with $L = 4$ is close to 0.8. Our results are consistent with the above-average quality of burial predictions used in previous simulations and indicate that the fraction of approachable proteins could increase significantly with even a mild, plausible, improvement on sequence-dependent burial prediction or on sequence-independent constraints that augment the detectable redundancy during simulations.

Proteins 2016; 84:515–531.

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

INTRODUCTION

Statistical mechanics of coarse-grained polymer models have contributed significantly to our general understanding of protein folding, as extensively reviewed.[1–3] An important insight is that random heteropolymers are not expected to display protein-like folding behavior because global energy minima resulting from random interactions would not be sufficiently deep to provide thermodynamic co-operativity nor stability at a temperature sufficiently high for kinetic accessibility. Evolution is normally assumed, therefore, to have acted on possible amino acid sequences to select a nonrandom subset with large global stability when compared with unfolded local minima, that is, sequences displaying minimal energetic frustration.[4,5] This expectation has been corroborated by simulations of minimalist lattice models with pairwise contact potentials with native-dependent sequence design[6,7] and geometrically realistic models with structure-based, native-dependent, Gō-type potentials.[8,9] Application of this simple idea in the development of sequence-dependent potentials with transferable parameters among different sequences, as required for simulations of sequences with unknown structures, turned out to be far from trivial, however. A possible source of

Diogo C. Ferreira and Marx G. van der Linden contributed equally to this work and should be considered first authors.

*Correspondence to: Antônio F. Pereira de Araújo, Laboratório De Biofísica Teórica E Computacional, Departamento De Biologia Celular, Universidade De Brasília, Brasília, DF, 70910-900, Brazil. E-mail:aaraujo@unb.br

Received 22 October 2015; Revised 28 December 2015; Accepted 19 January 2016
Published online 27 January 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24998

complication is that real protein sequences tend to look random[10–12] and potentials parameterized in terms of a finite set of atomic types reflecting sequence directly will also look random and incompatible with protein-like folding behavior. Attempts to cope with this problem in the context of knowledge-based parameterization tend to increase the number of atomic types and/or interactions, as exemplified by the "as Gō as possible" contact μ-potential[13] or the exposure-dependent effective energies of water-mediated contacts in associative memory folding hamiltonians.[14] Although easier to be accommodated favorably in any finite collection of structural examples, a larger number of types might have as an undesirable effect a substantial increase in the number of possibilities from which parameter values must be effectively chosen by any particular amino acid sequence.

From an informational perspective, conversely, the apparent randomness of protein sequences might seem less surprising. Shannon's theorems for discrete communication channels imply that long messages of a quite general class, possibly constrained by internal correlations, can be made to look arbitrarily close to random as seen from the channel by efficient encoding.[15–17] The number of probable signals, or channel states, can then increase up to the largest value compatible with the alphabet of a noiseless channel, or to a possibly smaller value that optimizes the number of non-confusable signals in the presence of noise, adjusting in any case the rate of information transmission to the maximum provided by the channel capacity. It appears natural, in this context, to consider specific protein sequences just as signals being transmitted through a communication channel. The fact that they are hard to distinguish from random sequences suggests an efficient encoding of structural messages with the appropriate average information content, or entropy, resulting in a rate of input information close to the maximal for 20 uncorrelated symbols distributed according to observed amino acid frequencies, that is $\approx$4.1 bits/residue[12,18] or $\approx$0.5 bits/atom. It is a reasonable possibility that an appropriate folding potential could be parameterized in terms of these hypothetical nonrandom structural messages that, even though encoded in sequences, would be more directly correlated with tertiary structures. The resulting number of atomic types could even appear to be large, in the sense that equivalent atoms in different occurrences of the same amino acid could be perceived differently by the resulting potential depending on sequence context,[19] but the number of parameters to choose from would be small, reflecting the sequence-compatible entropy of structural messages.

We have been exploring the possibility that atomic burials, as quantified by central distances, could constitute such messages.[18,20–22] In terms of a simple analogy with human communication, which is actually similar to the one discussed in Ref. 11, we investigate if the set of viable burial sequences could constitute a "language" capable of expressing structural "ideas" while satisfying the "grammar" of sequence-independent constraints and being encodable in the "script" generated by the alphabet of amino acids.[21] In this direction, native-like conformations were obtained in folding simulations guided by burial potentials pushing each atom toward its native central distance, either regarded as a continuous parameter[20] or discretized into a small number L of equiprobable burial layers.[21] This result suggested that atomic burials could indeed be regarded as a practical uniquely decodable code for tertiary structures, at least for the limited number of small proteins that had been investigated. The average information content of burial sequences, estimated from burial local correlations in small globular proteins for small L, was additionally found to be comparable to the entropy of protein sequences and therefore sufficiently small to be at least in principle further encoded in them. Burial information was later shown to be actually extractable from sequence by simple statistical prediction schemes, providing an estimated average reduction in burial uncertainty around 15% for 2 or 3 burial layers of C_α or C_β atoms,[18] suggesting that the resulting sequence-dependent burial predictions could be appropriate for folding simulations if the burial folding code turned out to be sufficiently redundant. Recent *ab initio* simulations using discrete burial information obtained directly from sequence with a burial prediction algorithm based on a Hidden Markov Model (HMM),[18] which was adapted from an algorithm for secondary structure prediction,[23] have been encouraging in this respect. Native-like conformations were obtained and identified for three small globular proteins, comprising all three structural classes, with an above average burial prediction accuracy around 56% for $L = 4$ burial layers.[22] If this approach for *ab initio* folding simulations using sequence-dependent burial propensities happens to be more general, and therefore applicable to a significant fraction of monomeric globular proteins, it could become a powerful scheme for tertiary structure prediction. More importantly, however, it would provide strong support for the underlying assumption about the general encoding of tertiary structures in amino acid sequences exclusively through burials.

A comparison with available prediction schemes might be instructive, since our simulations resemble previous schemes that minimize a simplified heuristic energy function containing both sequence-dependent and sequence-independent terms. As emphasized in previous reports,[20–22] however, they stand out in assuming that burial propensities constitute the only information to be obtained from sequence. In particular, there are no sequence-dependent terms to favor specific tertiary interactions as in most previous schemes, for example, Refs. 13,14,24, nor any sequence-dependent bias on backbone dihedral angles as in popular schemes involving libraries of backbone fragments, for example, Ref. 24. Our scheme involves therefore a significantly smaller amount of

sequence-dependent information, implying that sequence-dependent parameters are, on one hand, more likely to be obtainable from sequence but, on the other hand, less restrictive with respect to tertiary structure determination. It could be noted that a modest improvement of *ab initio*, "template-free," prediction algorithms along two decades of CASP experiments, contrasting to more successful modeling based on homologous templates, as recently reviewed,[25] is indeed consistent with the hypothesis that previous schemes have attempted to obtain more information from the primary sequence than is actually available. Furthermore, mathematical decompositions of "knowledge-based" parameters intended to extract specific, "high information," pairwise contact preferences from the statistics of known structures, for example, Refs. 26,27, have suggested a dominance of unspecific, "low information," hydrophobic/polar burial propensities.[27,28] Recent interest in improving contact prediction using presently available large, sufficiently "deep," multiple sequence alignments[29,30] further supports the notion that such a restrictive piece of information is not obtainable from single sequences.

The dominance of unspecific burial propensities in statistical pairwise parameters is consistent with the established notion that hydrophobicity constitutes a major physical factor in protein folding.[31] Burial propensities obtained from HMM predictions, formally considered as "knowledge-based," are also likely to reflect mainly the distribution of amino acid hydrophobicities in sequence fragments around each residue. It is noteworthy that sequence-dependent approximate traces of C_α central distances can actually be obtained from a purely analytical polymer model that takes into consideration nothing more than amino acid hydrophobicities, chain connectivity and excluded volume.[32,33] To compensate for the intrinsically small amount of sequence-dependent burial information, however, sufficiently restrictive sequence-independent constraints acquire crucial importance in the burial folding scheme, as particularly emphasized by the requirement of hydrogen bond formation by buried backbone polar atoms, independently of partner. This is again physically reasonable, since the enforcement of backbone hydrogen bonds has long been known to provide a dramatic reduction in the conformational space of a polypeptide, as observed already by Pauling more than six decades ago.[34] The whole approach turns out to be consistent, therefore, with the simple physical picture of protein folding resulting from minimization of exposed hydrophobic surface with concomitant satisfaction of hydrogen bond restraints.[35]

Accordingly, our restrictive sequence-independent hydrogen bond term, as previously detailed,[22] is actually reminiscent of Pauling's original work. Our burial term, however, is different in some possibly relevant informational aspects from similarly inspired burial terms, and burial layers, that have been used in previous prediction schemes, for example, Refs. 14,27. Regarding the specific piece of structural information assumed to be related to burials, central distances are possibly more restrictive about tertiary structures when compared with structural parameters more directly associated to hydrophobicity such as solvent accessibility, or even local contact densities or contact numbers, since central distances contain global instead of local structural information. This can be easily realized from the observation that the central distance of each atom depends on the whole structure and not just on its local environment. It should be noted, nevertheless, that contact numbers have been shown to be sufficiently restrictive at least in the context of lattice models.[36,37] Probably more important, however, is the manner in which individual atomic burial propensities, and therefore the specific burial layer towards which each atom will be pushed, are assumed to be obtainable from sequence. HMM predictions depend on local sequence and might well assign different layers, or "burial types," to atoms that could be considered of the same "atomic type" in previous schemes, such as "C_α of Leucine" or "C_γ of Valine." Additionally, as discussed above, the number of possible layers to be chosen from is small, implying that a large number of atoms of different "atomic types" must turn out, conversely, to be perceived as the same "burial type" by the burial folding potential.

In any case, it is apparent that the fraction of proteins to which this general approach might be applicable should increase both with the quality of the burial prediction scheme, or the amount of burial information obtainable from sequence, and of sequence-independent constraints that filter out protein-unlike conformations and increase the detectable redundancy in our simulations. Here, we investigate this interplay between required information and available redundancy in a diverse set of globular proteins, comprising all structural classes and a wide size range from $\approx$50 to $\approx$140 amino acid residues. We use molecular dynamics simulations with a varying amount of native burial information as provided by a potential term pushing each atom toward one among a small number L of equiprobable concentric layers, as in previous investigations. Sequence-independent information is again provided by geometrical constraints on covalent structure and hydrogen bond formation. An upper bound for required burial information is provided by the minimal number of layers, $L^{\min}$, still compatible with correct folding when all atoms are pushed to their native layers with our current sequence-independent constraints. Redundancy $\rho(L)$ for a given $L \geq L^{\min}$ is similarly estimated from the largest folding-compatible fraction of "superfluous" atoms, for which the burial term can be turned off or target layers can be chosen randomly. While $\log_2 L$ quantifies the required prediction precision, or the restriction to be imposed on each atom, the redundancy $\rho(L)$ reflects the tolerable inaccuracy, imposing a lower limit for the quality of burial prediction required for appropriate folding behavior and structural determination. Presently estimated redundancy can therefore be compared directly to the

relative uncertainty provided by any burial prediction scheme, as will be done here for the HMM used in our previous simulations.

THEORETICAL BACKGROUND

Redundancy in burial information for a given number of layers, $\rho(L)$, can be estimated directly from the maximal fraction $f_1^{\max}(L)$ of atoms for which the burial potential term can be "turned off" while folding behavior is preserved in our simulations, or

$$\rho_1(L)=f_1^{\max}(L). \tag{1}$$

This is analogous to estimate the redundancy of written text from the maximal fraction of randomly erased letters that can be successfully restored by a reader, as described by Shannon,[15] which is actually an ingenious procedure that provides a lower bound for the redundancy of the written language, or an upper bound for its entropy, relying solely on grammatical rules known unconsciously by the reader, bypassing therefore explicit consideration of such rules, or the resulting probability distribution of grammatical sentences generated by such rules that ultimately underlies both informational measures. Successful restoration for a given fraction f_1 of erased symbols indicates that the correct message is the only N-long sequence of L symbols recognized as "meaningful" by the receiver within the group of $\Omega_1(N, L, f_1)=L^{Nf_1}$ possibilities that are compatible with the corrupted signal. Considering the total number $\Omega_0(N, L)=L^N$ of possible sequences we have that

$$f_1=\frac{\log_2\Omega_1(N, L, f_1)}{\log_2\Omega_0(N, L)}=\frac{H_1(L, f_1)}{H_0(L)} \tag{2}$$

equals the ratio between the entropy associated to the number of "meaningless" sequences from the receiver perspective, $H_1(L, f_1)$, where $H_1(L, f)=f\log_2 L$ is the entropy per symbol of uniformly distributed sequences compatible with a corrupted signal from which a fraction f of symbols has been erased, and the entropy per symbol of uniformly distributed all possible sequences, $H_0(L)=\log_2 L$.

In other words, each "meaningful" N-long sequence, whose total number is $\Omega(N,L)$, lies within a region in sequence space containing additional $\Omega_1(N, L, f_1)-1$ "meaningless," but recognizably close, companions. Since the total number of "meaningful" and "meaningless" sequences in the combination of all such mutually exclusive regions cannot be larger than the total number of sequences, $\Omega_0(N, L)=L^N$, it is clear that

$$\Omega(N, L)\Omega_1(N, L, f_1) \le \Omega_0(N, L), \tag{3}$$

or

$$H(L)+H_1(L, f_1) \le H_0(L) \tag{4}$$

for all $f_1 \le f_1^{\max}$, and

$$\frac{H_1(L, f_1)}{H_0(L)} \le \frac{H_1(L, f_1^{\max}(L))}{H_0(L)} \le \frac{H_0(L)-H(L)}{H_0(L)}, \tag{5}$$

or

$$f_1 \le \left[f_1^{\max}(L)=\rho_1(L)\right] \le \rho(L). \tag{6}$$

Accordingly, $\rho_1(L)$ provides indeed a lower bound for $\rho(L) \equiv 1-H(L)/H_0(L)$, the actual redundancy as defined by Shannon,[15] where $H_0(L)=\log_2 L$ is the the maximal entropy per symbol, corresponding to uniformly distributed sequences, and $H(L)$ is the actual entropy per symbol, corresponding to "meaningful" or, more formally, typical long sequences satisfying whatever constraints happen to determine the underlying probability distribution. The only assumption is that this distribution should satisfy the general asymptotic equipartition property,[15,16] implying that for large N the sequence ensemble is dominated by a number close to $\Omega(N, L)=2^{NH(L)}$ of approximately equiprobable typical sequences. While $\rho(L)$ corresponds to the maximal fraction of erasures that could be restored by a hypothetical perfectly reliable receiver, $\rho_1(L)$ provides the expected restorable fraction by the available receiver, that is, folding simulations with sequence-independent constraints. Here we initially use $f_1=1/2$, 3/4, 7/8, and 15/16 and randomly choose the Nf_1 atoms for which the burial term is turned off.

Note that $f_1^{\max}(L)$ is different from the maximal tolerable inaccuracy in a given burial prediction scheme. Since the receiver of a long sequence of predicted burials containing a fraction f_2 of errors does not know the positions of these errors, the number of possible sequences that must be recognized as meaningless is now larger than before. The associated entropy of sequences compatible with a corrupted signal with a fraction f of errors, with each error uniformly distributed among $L-1$ wrong possibilities, becomes

$$H_2(L, f)=-f\log_2 f-(1-f)\log_2(1-f)+f\log_2(L-1), \tag{7}$$

with contributions from the uncertainty about error positions and from possible errors given a set of positions. The usable redundancy to preserve folding behavior in this context can then be estimated by

$$\rho_2(L)=\frac{H_2(L, f_2^{\max}(L))}{\log_2 L} \tag{8}$$

where $f_2^{\max}(L)$ is now the largest fraction of atoms that can be pushed toward a wrong target layer, randomly chosen from the $L-1$ possibilities, while preserving

folding behavior. To facilitate the comparison with $\rho_1(L)$, we randomly choose Nf_2 atoms with different f_2 values in such a way that $H_2(L,f_2)/\log_2 L = 1/2$, 3/4, 7/8, and 15/16, as well as the negative control $H_2(L,f_2)=\log_2 L$. In a more realistic prediction scenario wrong symbols would not be equally likely when $L>2$ because reasonable prediction schemes, when mistaken, are expected to choose wrong layers that are close to the correct native layer and the entropy of sequences compatible with a corrupted signal would be smaller than $H_2(L,f)$, satisfying the so called Fano's inequality.[16] The minimal required redundancy to resolve the remaining burial uncertainty after a prediction with a given inaccuracy, that is, a given fraction f of incorrectly predicted burials, should therefore be clearly larger than $H_1(f,L)/H_0(L)=f$ but somewhat smaller than $H_2(f,L)/H_0(L)$.

If an available burial prediction scheme happens to provide probabilities for each atom to occupy different burial levels then the negative average log-likelihood of observed burials,[18,23] $-\langle LL\rangle = -\langle \log_2 P_a(b_L^*)\rangle_a$, where $P_a(b_L^*)$ is the predicted probability for atom a to occupy its native burial layer b_L^*, can be used as an estimate for the remaining burial uncertainty after prediction and also, as a consequence, for the minimal required redundancy to resolve this uncertainty. We should have therefore in this case

$$f < -\frac{\langle LL\rangle}{\log_2 L} \lesssim \frac{H_2(L,f)}{\log_2 L}, \tag{9}$$

and folding behavior could be expected if

$$-\frac{\langle LL\rangle}{\log_2 L} < \rho_3(L) \approx \rho_2(L), \tag{10}$$

or

$$\frac{I_p(L)}{\log_2 L} > 1-\rho_3(L) \approx 1-\rho_2(L), \tag{11}$$

where $\rho_3(L)$ is the usable redundancy, detectable by our folding simulations, to correct for mistakes in sequences of burial symbols with the noise structure generated by our burial prediction scheme, which is assumed to be reasonably close to $\rho_2(L)$. Note that $\rho_2(L)$ is presently estimated by our simulations with noisy potentials and is independent of any particular burial prediction scheme while the relative prediction uncertainty, $-\langle LL\rangle/\log_2 L$, and the related prediction information, $I_p(L)=\log_2 L+\langle LL\rangle$, reflect the quality of the given burial prediction independently of folding simulations. These measures of prediction quality are more general than the prediction accuracy, $1-f$, since they depend on the whole distribution of predicted burials. High predicted probabilities for observed burials increase prediction quality, as expected, since they contribute to decrease prediction uncertainty $-\langle LL\rangle$ increasing prediction information $I_p(L)$. Predicted probabilities for observed burials close to $1/L$, conversely, correspond to poor predictions indistinguishable from a random burial choice independent of sequence, in which case $-\langle LL\rangle \approx \log_2 L$ and $I_p(L)\approx 0$. Additionally, if most observed atoms would happen to have predicted probabilities smaller than $1/L$ then the prediction would be even worse than random, resulting in $-\langle LL\rangle/\log_2 L > 1$ and $I_p(L)<0$.

We finally note that we are currently interested in comparing the available redundancy with the remaining uncertainty on burial prediction obtained from single sequences, which is most relevant for the general understanding about the genetic encoding of tertiary structures. A similar analysis, however, could also be performed to any kind of additional burial knowledge that happens to be available, such as evolutionary burial information possibly detectable in multiple sequence alignments,[38] which are becoming an increasingly reliable source of various kinds of evolutionary structural information,[5,29,30,39] or even experimental information.

RESULTS

We investigate eight globular proteins, whose native structures are shown in Figure 1, ranging in size from $N_r = 54$ to $N_r = 141$, comprising all structural classes and including particularly complex topologies. Molecular dynamics simulations, using geometrically realistic models with all nonhydrogen atoms represented as single beads of unit mass were performed as in previous investigations,[21,22] with a program adapted from a Fortran code previously used in simulations with structure-based C_α models[40] and modified afterwards to handle all atoms. As detailed in Ref. 22, the potential contains simple terms on bond lengths, angles, rigid dihedrals and pairwise repulsion, with non-standard terms for burials and hydrogen bond formation, and was derived from the all-atom structure-based model described in Ref. 41, with removal of the native-dependent contact and dihedral energy terms and the addition of terms for hydrogen bonds and atomic burials. The hydrogen bond term is annealed linearly during each folding trajectory, a convenient procedure to avoid kinetic trapping at the expense of realistic modeling of path-dependent features of the process, as previously discussed.[22] No annealing is performed in unfolding trajectories, which begin from the native structure with full hydrogen bonds. The intervals in central distances defining each burial layer are presently adjusted to accommodate the same number of atoms in each layer, while the radius within which hydrogen bond formation is enforced is still derived from the radius of gyration for each protein estimated from its number of residues, as in our previous simulations. All size-independent parameters are also the same as in this previous investigation.[22] Total simulation time is however 10 times smaller, or 2 ×

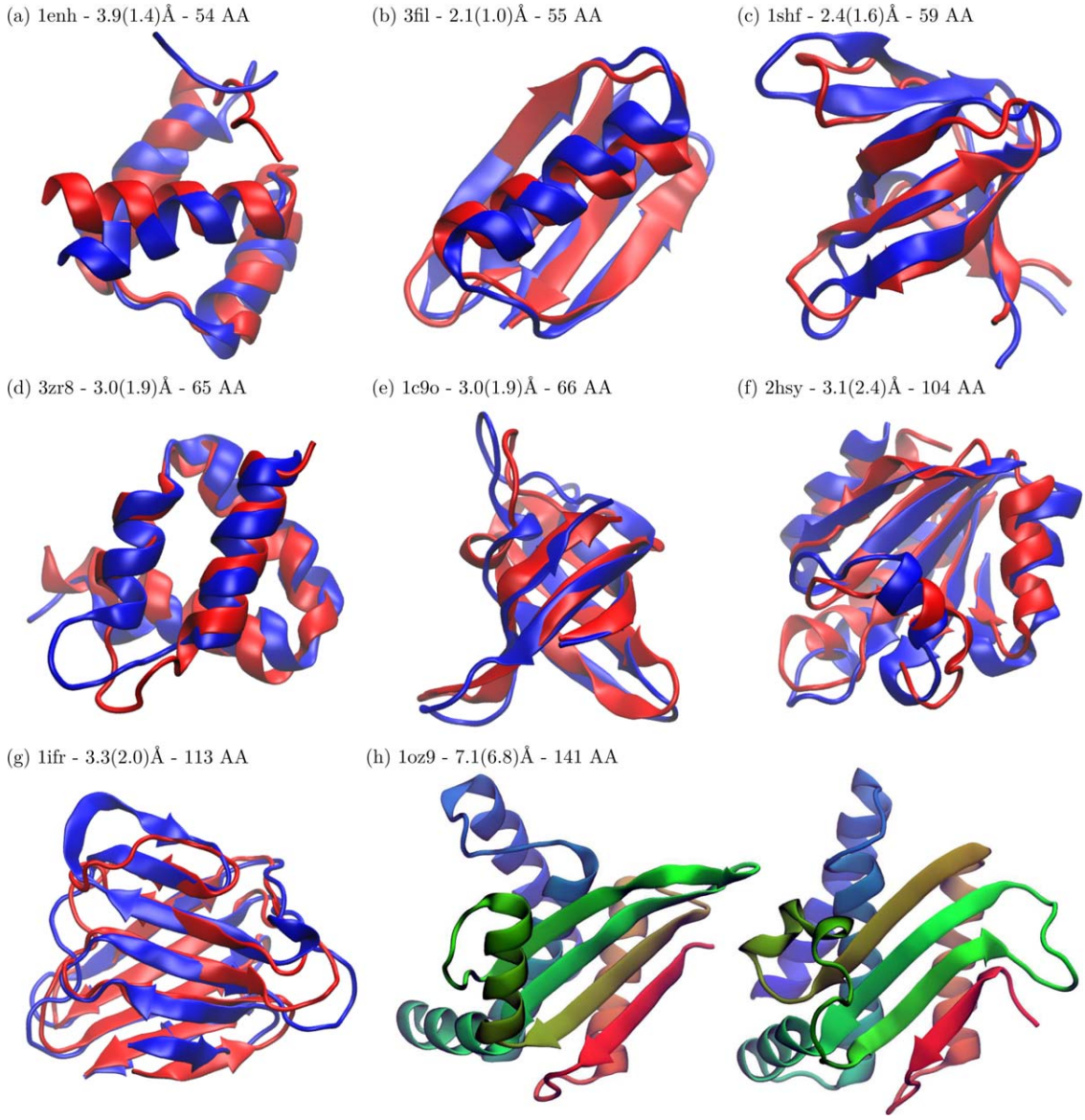

Figure 1

Structures used in this investigation. The lowest energy conformation among the last 10% of folding trajectories is shown in blue, superimposed to the native structure shown in red, for 1enh (a), 3fil (b), 1shf (c), 3zr8 (d), 1c9o (e), 2hsy (f), and 1ifr (g). For 1oz9, we selected the structure with largest fraction of secondary structure from the folding trajectory with lowest ⟨RMSD⟩. It is shown side by side with the native structure, native to the left, with a colour code indicating position along the sequence from red to blue (h). The number of burial layers in the folding trajectories from which the simulated structure was selected was $L = 4$, except for 2hsy and 1oz9 in which case $L = 5$ and $L = 8$, respectively, were used instead. We also show for each protein its number of residues, the RMSD of the selected conformation and, in parenthesis, the lowest RMSD observed in folding trajectories.

10^8 instead of 2×10^9 time steps, to viabilize both the required large number of simulations for the estimates of available redundancy in small proteins as well as the small number of significantly more demanding simulations for the larger proteins.

We estimate $L^{\min}$ for these eight proteins as the lowest value of L for which the native state is both accessible and stable during our simulations with all atoms being pushed to its correct layer. Stability is indicated by final conformational ensembles in all unfolding trajectories dominated by low C_α-RMSD structures while a similarly native-like final conformational ensemble for at least one folding trajectory indicates accessibility. Figure 2 shows trajectories beginning from the native structure (left) and from an extended initial conformation (right) with $L = 2$, 3, and 4 and a perfectly accurate burial potential

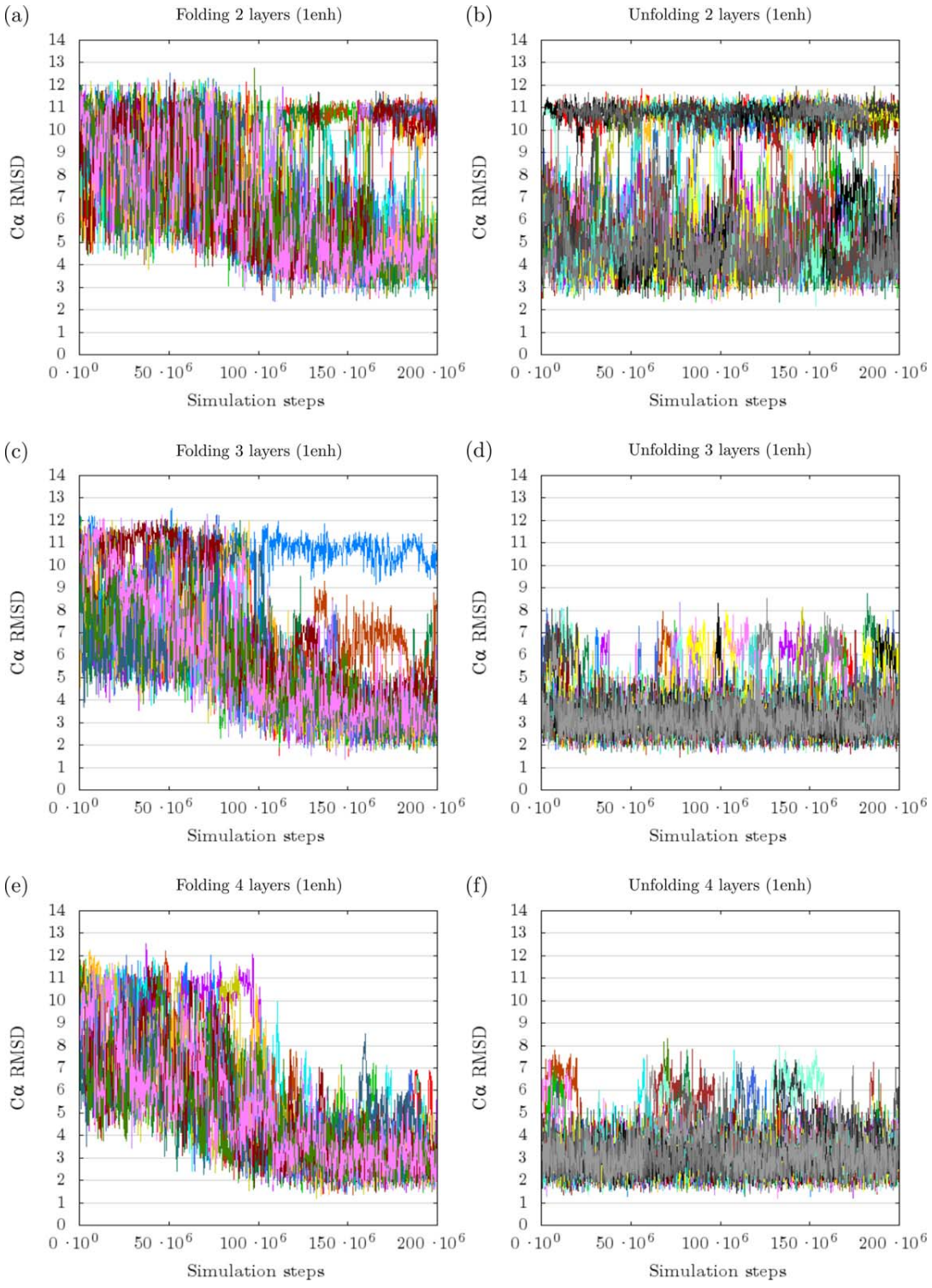

Figure 2

Folding and unfolding trajectories for the all-α engrailed homeodomain, PDB code 1enh. RMSD is shown as function of simulation time step for folding trajectories, beginning from an extended initial conformation (a, c, and e), and unfolding trajectories, beginning from the native structure (b, d, and f), with a native burial potential with $L = 2$ (a and b), $L = 3$ (c and d), and $L = 4$ (e and f). The hydrogen bond term increases linearly during folding trajectories and remain constant in unfolding trajectories. Twenty five independent trajectories are shown in each plot. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

for the α-helical engrailed homeodomain, pdbcode 1enh, where C_α root mean square deviation from the native structure (C_α-RMSD, or simply RMSD) is shown as a function of simulation time for 25 trajectories in each plot. RMSD remains low most of the time for all unfolding trajectories with $L = 3$, around 3 Å, while for $L = 2$ abrupt interconversions are observed between two narrow RMSD ranges, corresponding respectively to native-like conformations around 4 Å and previously described mirror-like structures[20,21] around 11 Å. The minimal number of layers required to preserve the stability of the native structure for this protein, with our current sequence-independent constraints, is therefore $L = 3$. Additionally, low RMSD conformations dominate the final conformational ensemble at least in some of the folding trajectories with $L = 3$, indicating that an ensemble of native-like conformations is also kinetically accessible during our simulations. We use this double criterion of native-like ensemble stability and accessibility to consider that $L^{\min} = 3$ for 1enh, as previously reported.[21] Note however that even for $L = 2$ the chain is already constrained within just two narrow and symmetrically related conformational ensembles. The native-like ensemble is accessible but not stable. The dominant conformational ensemble for $L = 3$ still increases slightly in abundance and similarity for $L = 4$, remaining both stable and accessible. Trajectories for 1enh are conveniently summarized in Figure 3(a) by a folding/unfolding double plot of the average RMSD along the last 10% steps in folding or unfolding trajectories, $\langle \text{RMSD} \rangle$, further averaged over the 25 trajectories in each plot, $\langle \langle \text{RMSD} \rangle \rangle$, as a function of L with standard deviation of the second average represented by error bars with minimal and maximal individual $\langle \text{RMSD} \rangle$ values represented by circles. Stability is indicated therefore by a low maximal unfolding $\langle \text{RMSD} \rangle$ while accessibility is conversely indicated by a low minimal folding $\langle \text{RMSD} \rangle$.

Accordingly, instability for $L = 2$ is indicated by a maximal unfolding $\langle \text{RMSD} \rangle \approx 11$ Å while the minimal folding $\langle \text{RMSD} \rangle < 5$ Å indicates accessibility if we consider this value as a sufficient indication for ensemble native-likeness. For $L = 3$, the maximal unfolding $\langle \text{RMSD} \rangle < 5$ Å indicates that the native ensemble is now stable during our simulation time while the minimal folding $\langle \text{RMSD} \rangle \approx 3$ Å shows that it remains accessible and actually more uniformly similar to the native conformation. A large maximal folding $\langle \text{RMSD} \rangle \approx 11$ Å corresponds to the single trajectory that converged to the mirror ensemble. For $L = 4$ stability and accessibility are maintained and now the maximal folding $\langle \text{RMSD} \rangle < 5$ Å reflects the observation that no trajectory was trapped in the mirror ensemble. The concentration of incorrectly folded trajectories around a mirror image of the native structure becomes more evident in Figure 3(b) where RMSD values in each trajectory were computed both with respect to the native structure and a single mirror conformation and averages were taken over the combination that retained the smaller of these two values for each conformation. The combined ensemble is seen to be stable for $L = 2$ to within ≈5Å RMSD from one of the two references, as indicated by low maximal unfolding $\langle \text{RMSD} \rangle$, and is highly accessible, as indicated not only by low minimal but also maximal folding $\langle \text{RMSD} \rangle$. Analogous folding/unfolding double plots for other small proteins, the α/β 3fil (c), all-β 1shf (d), all-α 3zr8 (e) and all-β 1c9o (f), are also shown in the same figure and can be interpreted similarly. For the α/β 3fil we observe instability and inaccessibility for $L = 2$ while for $L^{\min} = 3$ and $L = 4$ the native ensemble is both stable and accessible, as determined by both the maximal unfolding and minimal folding $\langle \text{RMSD} \rangle < 5$ Å. The all-β protein 1shf, conversely, is both unstable and inaccessible for $L = 2$, stable but inaccessible for $L = 3$, and both stable and accessible only for $L^{\min} = 4$. The native ensemble of all-α 3zr8 is again both unstable and unaccessible for $L = 2$, becomes accessible but remains unstable for $L = 3$ and, finally, is both stable and accessible for $L^{\min} = 4$. The all-β 1c9o is the only protein in this collection for which the native ensemble could almost be considered stable and accessible with our present constraints for $L = 2$, since both maximal unfolding and minimal folding $\langle \text{RMSD} \rangle$ are just above 5Å. Both values decrease perceptibly, to ≈3Å, as L increases to 3 or 4 and we decide to maintain the sharp threshold and consider $L^{\min} = 3$ for 1c9o.

Concomitant stability and accessibility during our simulations, as presently estimated, might be affected by kinetics and could depend on total simulation time in addition to the number of attempted trajectories. Longer trajectories, or slower annealing protocols, are expected to increase the probability of reaching the lowest energy conformational ensemble while avoiding kinetic traps. Accordingly, if native-like ensembles happen to be kinetically inaccessible during short trajectories but lower in energy than unfolded alternatives as computed with the same potential, that is, if they are energetically distinguishable, they should eventually become accessible in sufficiently long trajectories. Increasing the number of trajectories for a given simulation time, conversely, increases the number of folding attempts and, therefore, the probability of a folding event being observed by chance in at least one trajectory. If obtained native-like ensembles are energetically indistinguishable, however, their observation would accordingly have even lower probability in longer simulations. Furthermore, energetically indistinguishable native-like ensembles obtained by chance as kinetic traps could not be identified without previous knowledge of the native structure. It is relevant for the present analysis, therefore, that present $L^{\min}$ estimates would not be greatly affected if energetic distinguishability were applied as a more stringent criterion for protein-like folding behavior.

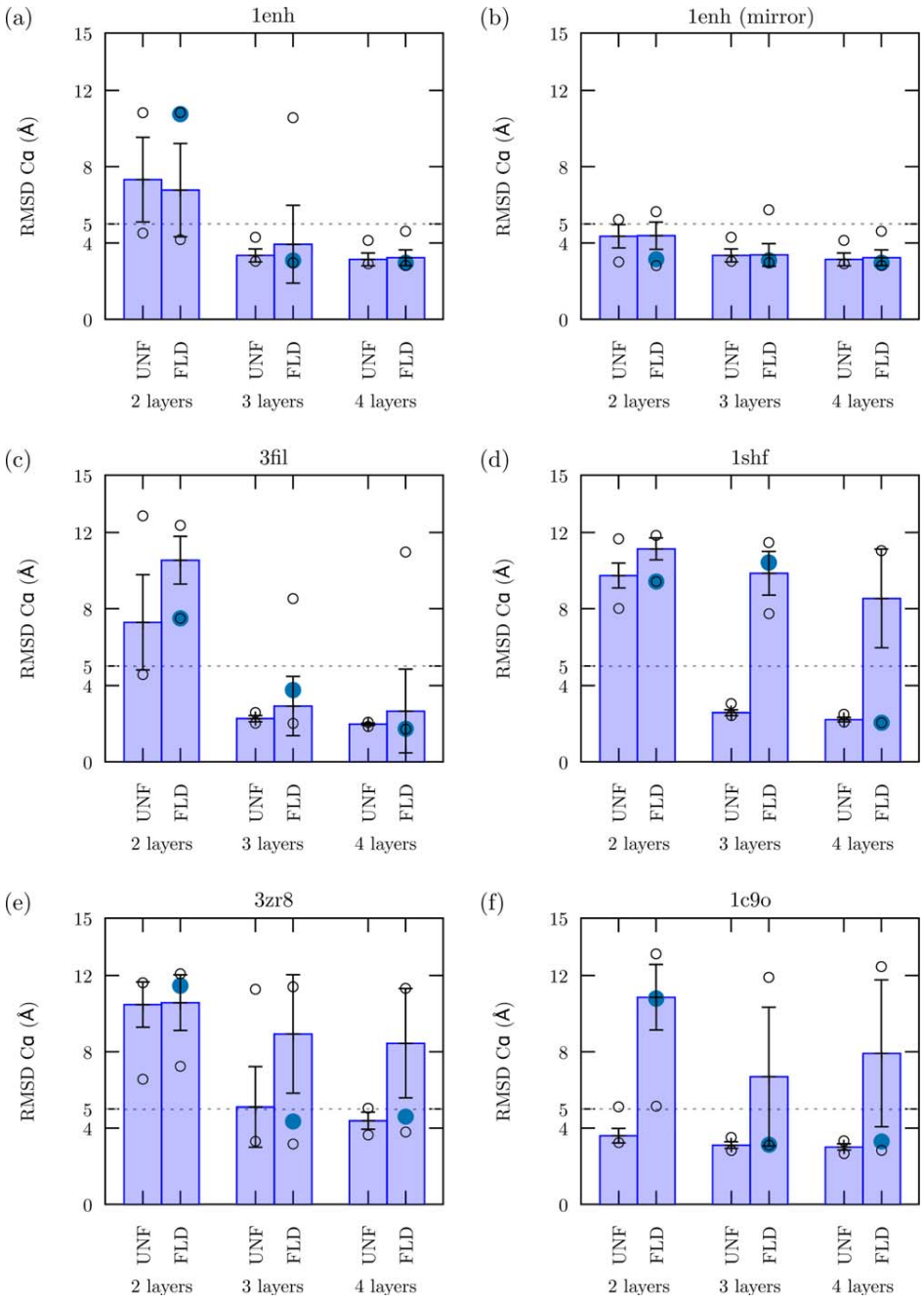

Figure 3

Folding and unfolding trajectories with a noiseless burial potential for small proteins. Each bar indicates the ⟨⟨RMSD⟩⟩ for the final 10% of unfolding (UNF) and folding (FLD) trajectories for small proteins 1enh (a and b), 3fil (c), 1shf (d), 3zr8 (e), and 1c9o (f), with a native burial potential with $L = 2$, $L = 3$, and $L = 4$ burial layers, as indicated. The double average corresponds to an initial RMSD average in each trajectory, ⟨RMSD⟩, further averaged over 25 independent trajectories. Error bars represent the standard deviation of this second average. Minimal and maximal individual ⟨RMSD⟩ values are indicated by open circles. A large filled circle indicates the ⟨RMSD⟩ for the folding trajectory with minimal final average energy. Averages shown in (b) are based on RMSD values both from the native structure and from a single mirror conformation, whichever is smaller for any given sampled conformation. Each bar in each plot condensates relevant information from a whole group of trajectories that would occupy a whole plot if presented as in Figure 2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Energetic distinguishability for the noiseless potentials might be suggested by a low ⟨RMSD⟩ for the folding trajectory with lowest average energy, which is shown in Figure 3 by a filled circle superimposed to the bar representing the distribution of folding ⟨RMSD⟩ values. The accessible and almost native-like ensemble for 1c9o with

$L = 2$, for example, corresponding to the lowest folding $\langle$RMSD$\rangle$ close to 5Å, as discussed above, turns out to be, anyway, energetically indistinguishable since at least one other ensemble with lower average energy, that is, the lowest energy final ensemble, was found to have high $\langle$RMSD$\rangle \approx 11$ Å. For $L^{\min} = 3$, however, the lowest energy final ensemble corresponds to a low, native-like, $\langle$RMSD$\rangle$. A particularly ambiguous result is provided by concomitant stability and inaccessibility, as observed for 1shf in 25 folding and unfolding trajectories with $L = 3$, since it is apparent that longer and/or more simulations should eventually detect either instability or accessibility. Accordingly, accessibility was indeed observed in an additional round of 50 folding trajectories (not shown) which could suggest a smaller $L^{\min}$ for this all-β protein but the obtained native-like ensembles turned out to be energetically indistinguishable. In any case, we find that native-like ensembles are stable, accessible and distinguishable for all these small proteins when $L = 4$, which is a sufficiently small number of layers to be directly attacked by our current discrete burial prediction scheme. Distinguishability is particularly emphasized by native-like conformations shown in Figure 1 superimposed to the native structure for each of these five proteins, with RMSD between 2 and 4 Å as indicated, which were obtained as the lowest energy conformation within the last 10% of all their folding trajectories with $L = 4$, while the lowest obtained RMSD is inside parenthesis. Consistent estimates for $L^{\min}$ resulting from the simpler criterion for protein-likeness justifies its use in our estimates for usable redundancy based on simulations with noisy potentials in which case the stricter criterion is not applicable since, as described further below, each trajectory is governed by a different potential.

Figure 4 shows folding/unfolding plots for two larger proteins, the all-β globular tail of nuclear lamin 1ifr (a), and the α/β bacterial thyorredoxin 2hsy (b), indicating that the sufficiency of a modest amount of burial information to make the native sate both stable and accessible extends beyond the previously investigated small size range. Interestingly, $L^{\min} = 4$ was found to be sufficient to provide stability and accessibility for the particularly complex all-β Ig-like fold,[42] with $N_r = 117$, while for the somewhat smaller and apparently simpler α/β domain,[43] with $N_r = 103$, the native ensemble was stable but not accessible with $L = 4$ in 25 folding attempts. Furthermore, for $L^{\min} = 5$ a single folding trajectory converged to a low RMSD ensemble, with $\langle$RMSD$\rangle < 5$ Å, but the resulting β-sheet topology turned out to be incorrect. In an additional round of 50 folding trajectories, however, some trajectories did converge to native-like ensembles with the correct topology with $L^{\min} = 5$ but not $L = 4$. The conformation with lowest energy in the final folding trajectories for 1ifr with $L = 4$ and 2hsy with $L = 5$ are shown in Figure 1 superimposed to their native structures. For the largest investigated protein, the α/β 1oz9 with $N_r = 141$ residues, $L = 6$ layers were required to maintain unfolding trajectories below the $\langle$RMSD$\rangle < 5$ Å threshold we have been using as a pre-condition for native-likeness, as shown in Figure 5(a). Regarding accessibility, however, we did not obtain any folding trajectory converging to this low $\langle$RMSD$\rangle$ range even when increasing the number of layers to 7 and 8 or in an additional round of 50 trajectories with $L = 8$ (not shown). An energetic consideration of final folding and unfolding ensembles becomes therefore particularly relevant for this larger protein. Figure 5(b) shows final $\langle$RMSD$\rangle$ plotted against average final energy for different folding and unfolding trajectories for 1oz9 with $L = 8$, showing that average final energies in the observed folding attempts are higher than average native energies obtained from unfolding trajectories, with a reasonable correlation between $\langle$RMSD$\rangle$ and average energy, strongly suggesting that the stable and distinguishable native-like ensemble should become accessible in sufficiently longer folding trajectories. It is also encouraging to note that the final folding ensemble with lowest $\langle$RMSD$\rangle \approx 7$ Å is in the lower energy range of folding trajectories and adopts an α/β conformation differing from the native topology by a single exchange between two strands in the β-sheet, namely, the second and third strands along the sequence, as shown in Figure 1 by the conformation from this trajectory with the largest fraction of standard secondary structure displayed beside the native structure.

Folding and unfolding simulations for three of the small proteins with a varying amount of noise added to the burial potential term with $L = 4$ are summarized in Figure 6. Each bar in each plot corresponds again to $\langle\langle$RMSD$\rangle\rangle$ in 25 trajectories with minimal and maximal $\langle$RMSD$\rangle$ indicated by circles. Each plot corresponds to a given protein, that is, 1enh (a), 3fil (b), and 1c9o (c). The first group of bars in each panel represents unfolding simulations with different fractions f_1 of atoms for which the sequence-dependent burial potential term is "turned off" and the atom is simply pushed to a common burial well comprising all layers. The second group in each panel also represents unfolding simulations but with a fraction f_2 of atoms for which the target layer is chosen randomly. As described in the theoretical background section, f_2 values in the second group are chosen in such a way that $H_2(f_2) = f_1$, according to Eq. (7), for f_1 values used in the first group and noise levels added in both groups are therefore expected to probe similar levels of available redundancy, as given by f_1. Note that $f_1 = 0$ and $f_2 = 0$ correspond to the same noiseless condition and represented $\langle$RMSD$\rangle$ values are the same already shown for unfolding trajectories in Figure 3. Finally, the third group in each panel represents folding simulations with the same fractions f_2 of atoms for which target layers are chosen randomly and values for $f_2 = 0$ are the same already shown in Figure 3 for folding trajectories. Differently from the averages shown in Figure 3,

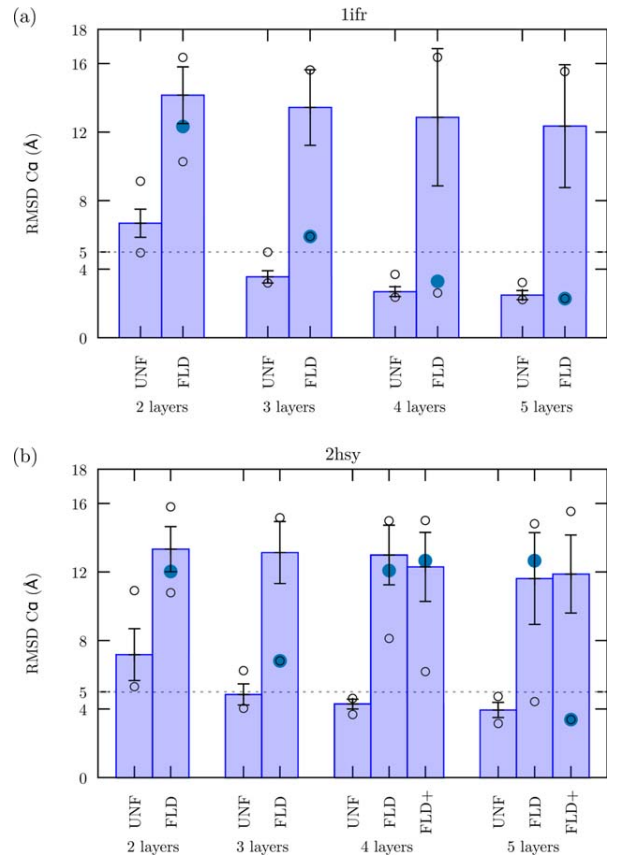

Figure 4

Folding and unfolding trajectories with a noiseless burial potential for two larger proteins. Each bar indicates the $\langle\langle \text{RMSD} \rangle\rangle$ for the final 10% of unfolding (UNF) and folding (FLD) trajectories for two somewhat larger proteins, 1ifr (a) and 2hsy (b), with a native burial potential with $2 \leq L \leq 5$ burial layers, as indicated. The double average corresponds to an initial RMSD average in each trajectory, $\langle \text{RMSD} \rangle$, further averaged over 25 independent trajectories. Error bars represent the standard deviation of this second average. Minimal and maximal individual $\langle \text{RMSD} \rangle$ values are indicated by open circles. A large filled circle indicates the $\langle \text{RMSD} \rangle$ for the folding trajectory with minimal final average energy. The extra bars for 2hsy in $L = 4$ and $L = 5$, labeled FLD+, correspond to an enlarged collection of folding simulations with 50 additional trajectories. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

however, which are taken over different trajectories with the same potential, averages for non-zero values of noise are taken over individual trajectories governed by different burial potentials, each generated from a different realization of noise. We now use $\langle\langle \text{RMSD} \rangle\rangle$ itself instead of maximal $\langle \text{RMSD} \rangle$ to probe for stability, considering $\langle\langle \text{RMSD} \rangle\rangle < 5$ Å as a rough indication that the native structure might remain stable for about half of these realizations of burial potential, while accessibility in folding trajectories is still indicated by at least one trajectory converging to a native-like ensemble. Stability might be therefore overestimated by this procedure, since longer unfolding trajectories could increase the fraction of unfolding events and the value of $\langle\langle \text{RMSD} \rangle\rangle$, but accessibility is conversely underestimated since a single folding event would be more likely to be observed if a larger number of simulations were attempted.

It is apparent from the first group of bars in each panel of Figure 6 that the available redundancy to preserve stability against erasures according to these criteria is larger than 3/4 but smaller than 7/8 for 3fil (b) and 1c9o (c), considering that $\langle\langle \text{RMSD} \rangle\rangle$ is smaller and larger than 5 Å, respectively, for these two fractions of erased burials. For 1enh (a) $\langle\langle \text{RMSD} \rangle\rangle$ is just above the threshold already for the smaller fraction, suggesting a somewhat smaller redundancy but still close to the lower end of the previous range, around 3/4. From the second group of bars in each plot we observe that the available redundancy to preserve stability against scrambling also appears to lie between 3/4 and 7/8 for 1enh and 3fil, while for 1c9o our present procedure provides a higher value, between 7/8 and 15/16. Lower limits for the estimated redundancy to preserve accessibility for the same scrambling noise scheme, obtained from the last group of bars in each plot, also lies between 3/4 and 7/8 for 3fil and 1enh. For 1c9o, however, the estimated value is very low since we did not observe folding events for $f_2 > 0$ in these collections of 25 trajectories. The resulting combination of high upper limit for stability with low lower limit for accessibility in the estimates for scrambling redundancy for 1c9o might reflect a significant kinetic barrier between folded and unfolded ensembles for this all-β protein. We obtain therefore, based on the results for 1enh and 3fil, $3/4 \leq \rho_2(L{=}4) < 7/8$ as an initial estimate for the available redundancy to preserve folding behavior. Note that this range is not inconsistent with our present results for 1c9o, although a similarly precise estimate in this particular case would require longer simulations and not be practical. We then attempt to narrow down the estimated range for 1enh and 3fil performing additional folding and unfolding simulations with noisy potentials probing additional intermediate levels of redundancy against scrambling, that is, 4/5, 5/6, and 6/7. We note from the double plots shown in Figure 7 that both proteins remain stable and accessible for $H_2(f_2){=}4/5$ while for $H_2(f_2){=}5/6$ we already have unfolding $\langle\langle \text{RMSD} \rangle\rangle$ at 5 Å for 1enh and larger than this threshold for 3fil. We suggest therefore that a usable redundancy close to 0.8, or $\rho_2(L{=}4) \approx 4/5$, is provided by our current sequence-independent constraints for small proteins with $L = 4$.

We now observe that the quality of atomic burial predictions obtained from a Hidden Markov Model that were used in our previous *ab initio* folding simulations are also consistently described in terms of the relations between inaccuracy and uncertainty provided by Eqs. (7) and (9). Prediction accuracy, $1 - f$, for $L = 4$ in a group of 278 small globular proteins with $N_r \leq 80$ was found

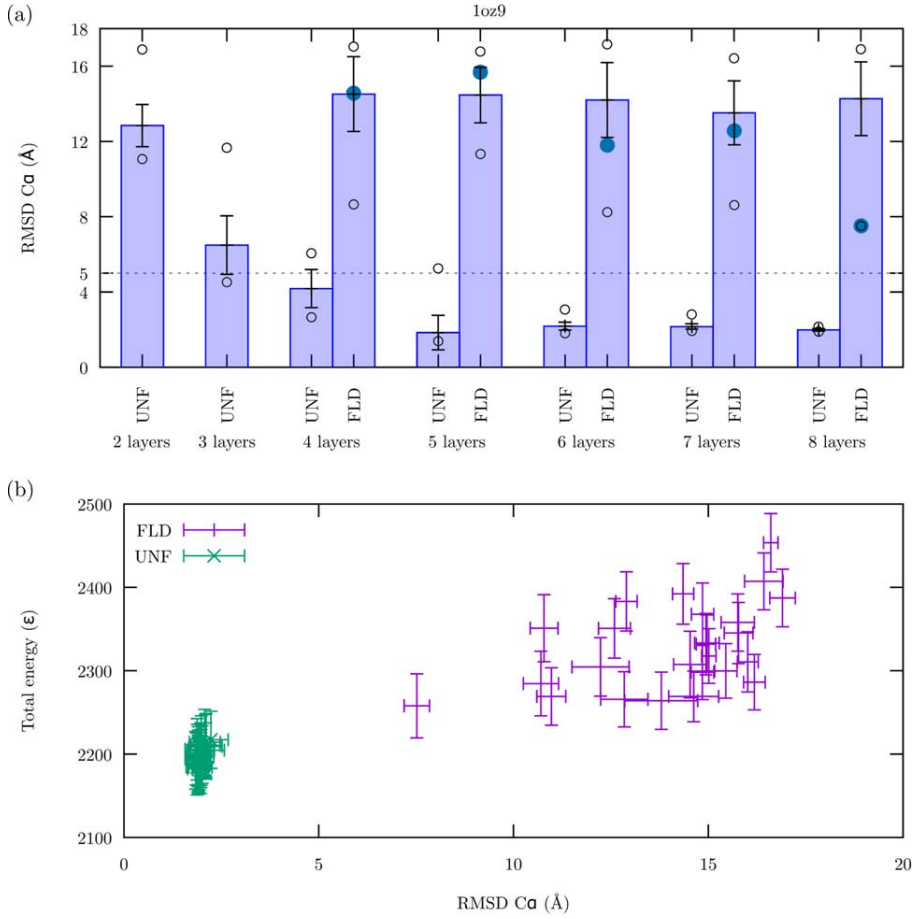

Figure 5

Folding and unfolding trajectories with a noiseless burial potential for the large α/β 1oz9. (**a**) Each bar indicates the $\langle\langle \text{RMSD} \rangle\rangle$ for the final 10% of unfolding (UNF) and folding (FLD) trajectories with a native burial potential with $2 \leq L \leq 8$ burial layers, as indicated. The double average corresponds to an initial RMSD average in each trajectory, $\langle \text{RMSD} \rangle$, further averaged over 25 independent trajectories. Error bars represent the standard deviation of this second average. Minimal and maximal individual $\langle \text{RMSD} \rangle$ values are indicated by open circles. A large filled circle indicates the $\langle \text{RMSD} \rangle$ for the folding trajectory with minimal final average energy. (**b**) Average final energy and average final RMSD, with standard deviations represented by error bars, for the 25 folding (FLD) and unfolding (UNF) trajectories for this large protein, with $L = 8$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to range from as low ≈0.25 to ≈0.6 with an average ≈ 0.44±0.07 and successful folding simulations were observed for three proteins with above average prediction accuracy around 0.56.[22] As shown in Figure 8(a) for 223 proteins from that group, with $50 \leq N_r \leq 80$, prediction uncertainty $-\langle LL \rangle$ for $L = 4$ is indeed somewhat smaller but follows the trend provided by the expression of $H_2(f)$ for the large majority of these predictions when the inaccuracy f is distinctively smaller than $(L-1)/L = 0.75$, or accuracy $(1-f) > 1/L = 0.25$, with relative uncertainty $-\langle LL \rangle / \log_2 L$ increasing from less than 0.75(3/4) for the best predictions with inaccuracy $f \lesssim 0.45$ to around 0.875(7/8) for close to average inaccuracy $f \approx 0.55$. For poor, random-like, predictions with inaccuracy $f \approx (L-1)/L = 0.75$, conversely, no correlation is apparent between f and $-\langle LL \rangle / \log_2 L$, with many examples of the latter assuming values larger than 1, an indication of "worse than random" predictions. The fraction of small proteins whose relative uncertainties for $L = 4$ are expected to be sufficiently small to be corrected during folding simulations by a detectable redundancy around 4/5, as presently estimated, is therefore small, corresponding the fraction of proteins for which $-\langle LL \rangle / \log_2 L < 0.8$, or just above 20% (46/223), as more clearly observed in the histogram of frequencies shown in Figure 8(b). Note that the fraction of approachable proteins could increase significantly if the detectable redundancy becomes closer to to 7/8 with eventually improved sequence-independent constraints in our simulations, since more than 50% (115/223) of the observed predictions would require a redundancy between 0.8 and 0.9.

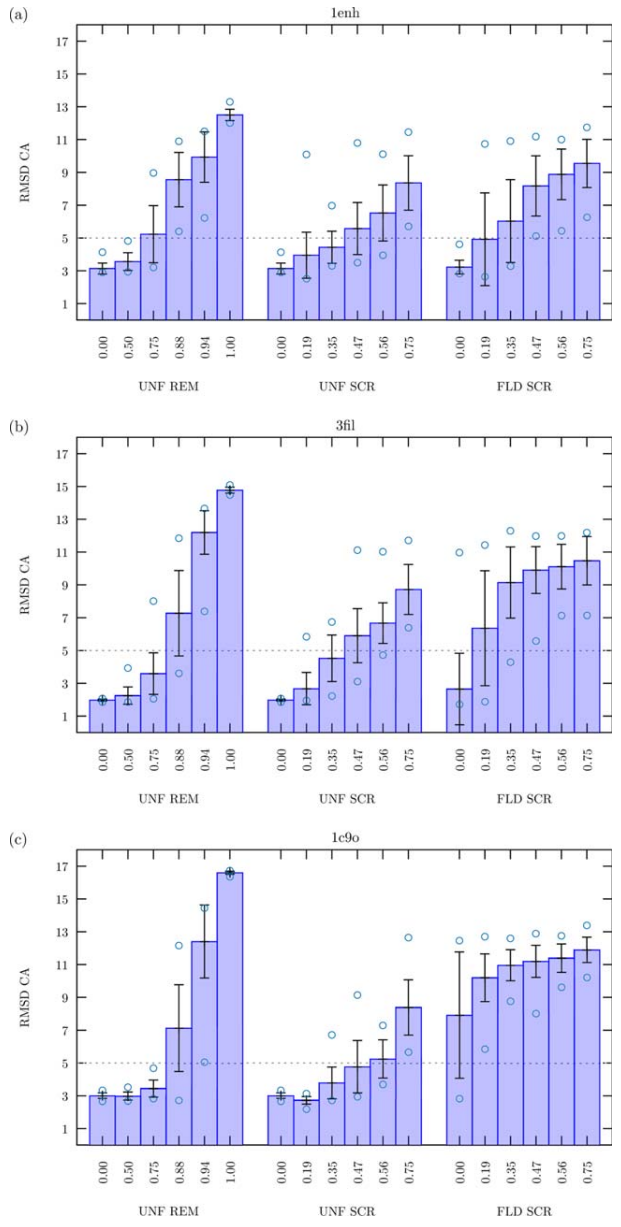

Figure 6

Folding (FLD) and unfolding (UNF) trajectories with increasingly noisy burial potentials with $L = 4$ layers. Each bar represents $\langle\langle \mathrm{RMSD} \rangle\rangle$ in 25 trajectories with minimal and maximal $\langle \mathrm{RMSD} \rangle$ indicated by circles. Each panel corresponds to a given protein, that is, 1enh (a), 3fil (b), and 1c9o (c). The first group of bars in each panel represents unfolding simulations with different fractions f_1 of atoms for which the native-dependent burial potential term is "turned off" or "removed" (REM). The second group of bars represents unfolding simulations with a fraction f_2 of atoms for which the target layer is chosen randomly, or "scrambled" (SCR), with f_2 values chosen in such a way that $H_2(f_2)=f_1$, according to Eq. (7), for f_1 values used in the first group. The third group represents folding simulations with the same fractions f_2 of scrambled atoms. Noise levels added for each group in all panels are expected to probe similar levels of available redundancy, as given by f_1, that is, 0, 1/2, 3/4, 7/8, 15/16, and 1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

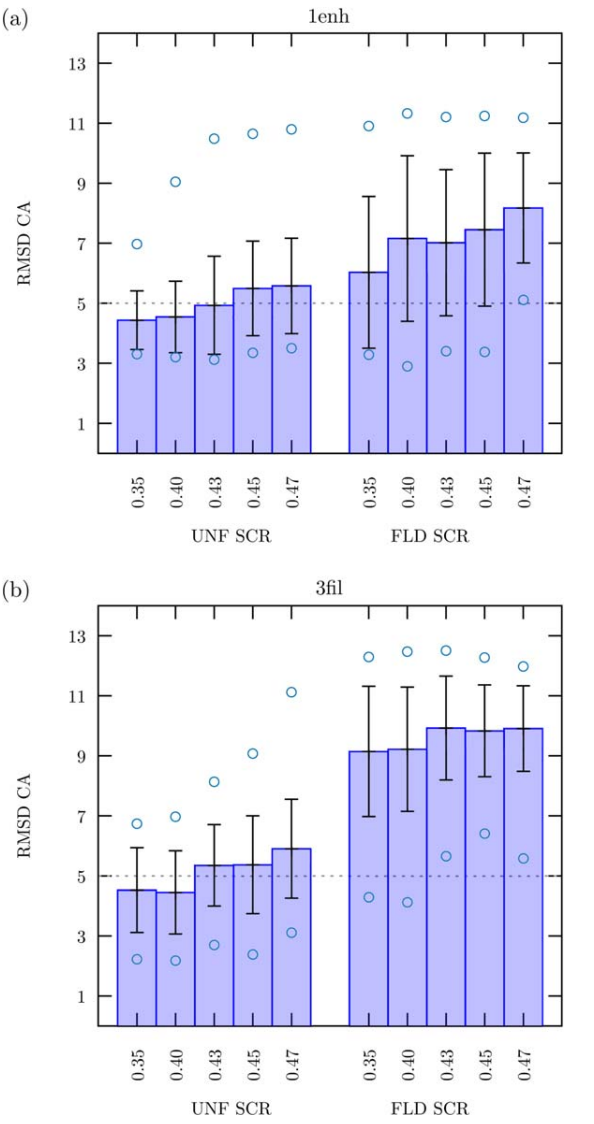

Figure 7

Folding (FLD) and unfolding (UNF) trajectories with noisy burial potentials with $L = 4$ layers probing for redundancy between 3/4 and 7/8 for 1enh(a) and 3fil (b). Each bar represents again $\langle\langle \mathrm{RMSD} \rangle\rangle$ with minimal and maximal $\langle \mathrm{RMSD} \rangle$ indicated by circles. The fraction f_2 of scrambled atoms, for which the target layer is chosen randomly, with $H_2(f_2)=3/4$, 4/5, 5/6, 6/7, and 7/8, according to Eq. (7), is indicated below each bar. Averages shown for $f_2 = 0.35$ and 0.47, or $H_2(f_2)=3/4$ and 7/8, respectively, are the same already shown in Figure 6 and were computed from 25 trajectories while for the additional intermediate values of $f_2 = 0.40$, 0.43, and 0.45, or $H_2(f_2)=4/5$, 5/6, and 6/7, in this order, averages were computed from 50 trajectories. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DISCUSSION

Our present simulations with noiseless burial potentials demonstrate that the sufficiency of a small amount of burial information to generate energetically distinguishable native-like conformational ensembles, when

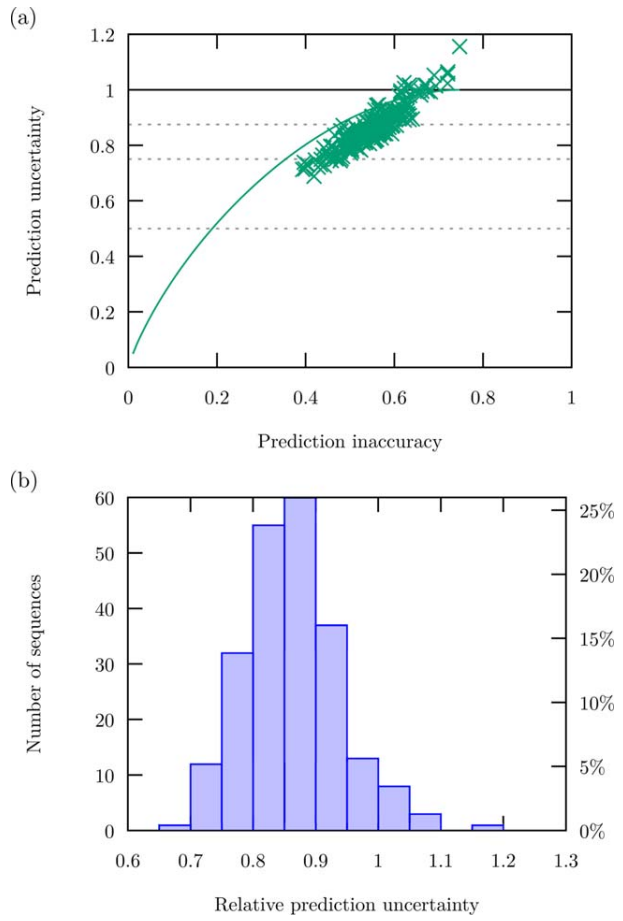

Figure 8

Relative uncertainty for available burial predictions. Relative prediction uncertainty, $-\langle LL\rangle/\log_2 L$, is plotted as a function of prediction inaccuracy, f, for burial predictions with $L = 4$ of 223 small globular proteins with $50 \leq N_r \leq 80$ obtained by a Hidden Markov Model as described elsewhere.[22] Each point corresponds to a protein. The curve represents $H_2(f)$ as provided by Eq. (7). Dashed lines mark relative uncertainty values of 0.5(1/2), 0.75(3/4) and 0.875 (7/8). A redundancy not smaller than the relative uncertainty would be required to correct for prediction mistakes and, since $L{=}4 > L^{\min}$, to result in correct folding behavior. Prediction mistakes for a small fraction of proteins could be appropriately corrected by a redundancy of 0.75 while 0.875 would be sufficient for most proteins, as more clearly observed in the histogram of frequencies obtained from the same data shown in (**b**). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

combined to physically motivated sequence-independent constraints, extends beyond a minute group of small or particularly simple examples. Furthermore, there is no apparent correlation between the amount of required information and structural class for proteins of similar sizes, as indicated by the small $L^{\min} \leq 4$ for all five small proteins with $N_r \leq 80$. A notable distinction is observed, however, in that the native ensemble of α-helical proteins for $L < L^{\min}$ can be accessible but unstable during our simulations while for β-sheets there is a tendency for the native ensemble to become stable before becoming accessible as $L < L^{\min}$ increases. This discrepancy is consistent with a larger kinetic barrier between folded and unfolded ensembles in the latter case, which is most likely to arise from the requirement of energetically expensive hydrogen bond disruptions during a spatial rearrangement of β-strands, but not of α-helices, inside the hydrophobic core. It is also interesting that the associated interconversion between symmetrically related ensembles for the α-helical proteins for $L < L^{\min}$, that is, 1enh with $L = 2$ or 3zr8 with $L = 3$, generates a combined ensemble that could be described as having native-like secondary structure in the absence of rigid tertiary interactions. This is a well-known defining feature of molten globules, which might be experimentally observed under weak denaturant conditions.[44] While sufficient burial information could explain how ensembles dominated by a single conformation arise under native conditions, this additional observation indicates that a sub-optimal amount of burial information could similarly explain, at least qualitatively, preeminent features of enlarged conformational ensembles induced by weak denaturants. Actually the whole structural continuum englobing the so-called "trinity" formed by ordered tertiary conformations, molten globules and intrinsically disordered chains, whose last member tends to arise from low complexity sequences with an insufficient number of apolar residues,[45] could be similarly described in terms of different amounts of burial information.

Our present results for the three larger proteins strongly suggest that globular size does not impose any intrinsic limitation on the obtainability of tertiary structures from a small amount of atomic burial information, with apparently just a mild increase of $L^{\min}$ with chain length as indicated by probable $6 \leq L^{\min} \leq 8$ no more than twice as large for 1oz9 when compared with the small proteins about half its length. Burial prediction for these larger values of L will require some adaptation of our current prediction scheme, since a direct increase of L in our HMM could increase the number of hidden states beyond the limit imposed by statistical saturation in a finite training set.[18] A possible approach to generate burial predictions for larger values of L would be to use our current predictions for small L to estimate a continuous burial probability density determined by a small number of adjustable parameters[46] from which the prediction for any number of layers could be derived. The resulting quality of this approach must still be investigated. A major additional limitation for our present approach to tackle large proteins in general appears to be, however, the total simulation time. For larger proteins direct simulations with our present code could become impractical and alternative approaches to sample conformational space more efficiently might become necessary. In any case, sequences of atomic L-burials with a rather small number of layers $L \geq L^{\min}$ appear to be

indeed sufficient to determine the tertiary structure of monomeric globular proteins in general and it becomes most relevant to investigate to what extent *L*-burial sequences could actually be encoded in, and obtainable from, amino acid sequences.

Sequences of atomic *L*-burials could be encoded in amino acid sequences, and reliably decoded in our simulations, if a sufficiently large amount of sequence-independent, redundant, information resulting from correlations present in actual burial sequences could be detected by our sequence-independent constraints. The amount of required non-redundant information could then become smaller than sequence entropy and eventually compatible with the more restrictive threshold provided by actual burial predictions from sequence. In other words, the possibility of a reliable encoding of atomic *L*-burials in amino acid sequences is intrinsically dependent on burial redundancy. In his classic exposition of information theory Shannon described how the redundancy of ordinary English had been estimated to be around 0.5 from several independent observations, including the entropy computed from frequencies of short blocks of letters in written text and the maximal fraction of erased letters that could be successfully restored by an anglophone reader. These two approaches are analogous, respectively, to our previous estimate for the entropy of burials in blocks of adjacent atoms in globular proteins and to our current estimate for the usable burial redundancy in folding simulations. In our case, however, the two estimates are not coincident. The entropy of *L*-burial sequences obtained from *L*-burial fragments of adjacent backbone atoms in a collection of small proteins was found to be $\approx$0.4 and $\approx$0.6 bits/atom for $L=2$ and $L=3$, respectively,[21] corresponding to a redundancy of $\approx$0.6 independently of L, that is, $(1-0.4/\log_2 2)$ or $(1-0.6/\log_2 3)$. This is smaller than our present estimate for the available redundancy to correct for inaccurate burials in folding simulations, that is $4/5=0.8>0.6$. It is apparent, therefore, that burial correlations extending beyond short fragments of adjacent atoms are detectable by sequence-independent constraints in our simulations and that the amount of non-redundant information required to specify a particular sequence of atomic burials is actually smaller than previously suggested. While for a redundancy of 0.6 the corresponding entropy of *L*-burial sequences is smaller than the primary sequence limit of $\approx$0.5 bits/atom for $L=2$ but not $L=3$, as discussed above, $L=4$ burial layers become entropically compatible with amino acid sequences if the redundancy is 0.75, since $(1-0.75)\times\log_2 4=0.5$. The number of compatible layers further increases to $L=5$, or possibly $L=6$, for the presently estimated detectable redundancy around 0.8, since $(1-0.8)\times\log_2 5=0.46$ and $(1-0.8)\times\log_2 6=0.52$. We had previously noted that a limited number of protein shapes imposes a strong constraint on possible burial sequences providing a large amount of sequence-independent burial information, both local and non-local.[22] This is equivalent to a large burial redundancy that could eventually be used, if detectable, to correct for predicted burial inaccuracy. Our present results, therefore, demonstrate that at least a fraction of this burial redundancy originated from non-local information is actually detectable by our current simulations.

The entropy of primary sequences is expected to be somewhat larger than the theoretical limit imposed on burial predictions by the mutual information between sequences and central distances, since a significant number of different sequences might be compatible with a given tertiary structure,[47,48] and therefore with given set of distances. Furthermore, it is useful to consider directly a more pragmatic limit imposed by the quality of actual predictions as quantified by prediction information, that is, the reduction in burial uncertainty due to prediction, $I_p=\log_2 L+\langle LL\rangle$ or, equivalently, by the (relative) remaining burial uncertainty on prediction $-\langle LL\rangle/\log_2 L$, which are both derived from the average log-likelihood of burials observed in available structures according to predicted probabilities. The quality of burial predictions obtained from a Hidden Markov Model (HMM) on a large collection of globular proteins with different sizes was previously found to provide an average reduction in burial uncertainty close to 15% for $L=2$ and $L=3$ layers of C_α or C_β atoms,[18] or a relative burial uncertainty close to 85%. The suggested required redundancy of at least 0.85 is slightly above our present estimate for the redundancy detectable in our simulations around $4/5=0.8$. If we restrict ourselves to small proteins and consider predictions for different proteins individually, we also find that 0.8 is smaller than required to compensate, for most proteins, the inaccuracy provided by our burial prediction scheme extended to all atoms with $L=4$. The distribution of prediction uncertainties shown in Figure 8 is actually consistent with our previous successful *ab initio* simulations with sequence-dependent burial potentials for proteins with above-average burial predictions[22] and suggests that the fraction of approachable proteins would indeed be significantly enlarged if detectable redundancy could become closer to 0.9. As detectable redundancy depends on sequence-independent constraints in our simulations it is possible that a more careful treatment of these constraints could increase the fraction of proteins approachable by our general scheme. No redundancy could be sufficient for correcting all predictions in this collection, however, since our worst predictions do not provide any burial information at all. Alternatively, a larger fraction of approachable proteins could also arise from an increase in average burial prediction quality, that is, a reduction in prediction uncertainty that could bring more points below the detectable redundancy threshold.

CONCLUSION

The obtainability of the tertiary structure of globular proteins from a small amount of atomic burial information is not limited to small chains or particularly simple topologies. Furthermore, there is no obvious correlation between the number of required burial layers and structural class, or complexity, for proteins with similar size. The major practical limitation for larger proteins appears to arise from overall slower kinetics and a high computational cost for the required longer simulations for these larger systems and not from any intrinsic difficulty in the general scheme connecting sequence to structure through atomic burials. Burial redundancy detectable during our folding simulations for small proteins is suggested to be sufficient for correcting for burial inaccuracies from actual predictions for a small fraction of proteins, which is consistent with our previous *ab initio* successful simulations. The fraction of approachable proteins could increase significantly with a plausible improvement on sequence-dependent burial prediction or on sequence-independent constraints that augment the detectable redundancy during simulations.

ACKNOWLEDGMENT

This research was supported in part by FINEP and FAPESP through the computational resources provided by the Center for Scientific Computing (NCC/GridUNESP) of the So Paulo State University (UNESP).

REFERENCES

1. Pande VS, Grosberg A, Tanaka T. Statistical mechanics of simple models of protein folding and design. Biophys J 1997;73:3192–3210.
2. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–75.
3. Shakhnovich E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 2006;106: 1559–1588.
4. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a Random Energy Model (with applications to protein folding). J Phys Chem 1989;93:6902–6915.
5. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc Natl Acad Sci USA 2014;111: 12408–12413.
6. Gutin AM, Abkevich VI, Shakhnovich EI. Evolution-like selection of fast-folding model proteins. Proc Natl Acad Sci USA 1995;92:3066–3076.
7. Shakhnovich EI. Proteins with selected sequences fold into unique native conformation. Phys Rev Lett 1994;72:3907–3910.
8. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298: 937–953.
9. Shimada J, Kussell E, Shakhnovich EI. The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. J Mol Biol 2001;308:79–95.
10. Ptitsyn O, Volkenstein M. Protein structures and neutral theory of evolution. J Biomol Struct Dyn 1986;4:137–156.
11. Pande VS, Grosberg AY, Tanaka T. Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? Proc Natl Acad Sci USA 1994;91:12972–12975.
12. Weiss O, Jimenez-Montano M, Herzel H. Information content of protein sequences. J Theor Biol 2000;206:379–386.
13. Kussell E, Shimada J, Shakhnovich EI. A structure-based method for derivation of all-atom potentials for protein folding. Proc Natl Acad Sci USA 2002;99:5343–5348.
14. Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG. Water in protein structure prediction. Proc Natl Acad Sci USA 2004;101:3352–3357.
15. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–623.
16. Cover TM, Thomas JA. Elements of information theory. Wiley-Interscience; Cambridge, UK, 2006.
17. MacKay DJC. Information theory, inference, and learning algorithms. Cambridge University Press; Hoboken, NJ, USA, 2003.
18. 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. doi: 10.1093/bioinformatics/bts512.
19. Dokholyan NV. What is the protein design alphabet? Proteins 2004; 54:622–628.
20. Pereira de Araújo AF, Gomes ALC, 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.
21. 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.
22. van der Linden MG, Ferreira DC, Oliveira LC, Onuchic JN, Pereira de Araújo AF. *Ab initio* protein folding simulations using atomic burials as informational intermediates between sequence and structure. Proteins 2014;82:1186–99.
23. Crooks GE, Brenner SE. Protein secondary structure: entropy, correlations and prediction. Bioinformatics 2004;20:1603–1611.
24. Simons K, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;3:171–176.
25. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science 2012;338:1042–1046.
26. Miyazawa S, Jerningan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
27. Papoian GA, Ulander J, Wolynes PG. Role of water-mediated interactions in protein-protein recognition landscapes. J Am Chem Soc 2004;125:9170–9178.
28. Li H, Tang C, Wingreen NS. Nature of the driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79:765–768.
29. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 2011;108:E1293–E1301.
30. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci USA 2013;110:15674–15679.
31. Dill KA. Dominant forces in protein folding. Biochemistry 1990;29: 7133–7155.
32. England J. Allostery in protein domains reflects a balance of steric and hydrophobic effects. Structure 2011;19:967–975.
33. Perunov N, England JL. Quantitative theory of hydrophobic effect as a driving force of protein structure. Prot Sci 2014;23:387

34. 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.
35. Dyson HJ, Wright PE, Scheraga HA. The role of hydrophobic interactions in initiation and propagation of protein folding. Proc Natl Acad Sci USA 2006;103:13057–13061.
36. 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.
37. Garcia LG, Treptow WL, Pereira de Araújo AF. Folding simulations of a three-dimensional protein model with a non-specific hydrophobic energy function. Phys Rev E 2001;64:011912
38. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
39. Bleicher L, Lemke N, Garratt R. Using amino acid correlation and community detection algorithms to identify functional determinants in protein families. PLoS One 2011;6:e27786
40. Whitford PC, Miyashita O, Levy Y, Onuchic JN. Conformational transitions of adenylate kinase: switching by cracking. J Mol Biol 2007;366:1661–1671.
41. 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.
42. Dhe-Paganon S, Werner ED, Chi YI, Shoelson SE. Structure of the globular tail of nuclear lamin. J Biol Chem 2002;277:17381–17384.
43. Amorim GC, Pinheiro AS, Netto LE, Valente AP, Almeida FC. NMR solution structure of the reduced form of thioredoxin 2 from Saccharomyces cerevisiae. J Biomol NMR 2007;38:99–104.
44. Ptitsyn O. Molten globule and protein folding. Adv Prot Chem 1995;47:83–229.
45. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
46. Gomes ALC, de Rezende JR, Pereira de Araujo AF, Shakhnovich EI. Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. Proteins 2007;66:304–320.
47. Larson SM, England JL, Desjarlais JR, Pande VS. Throughly sampling sequence space: large-scale protein design of structural ensembles. Prot Sci 2002;11:2804–2813.
48. Koehl P, Levitt M. Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 2002;99:1280–1285.

Capítulo 3

Representação diferencial

3.1 Representação diferencial dos enterramentos atômicos

Em uma analogia, podemos visualizar a sequência de enterramentos atômicos ao longo de uma cadeia proteica como um sinal analógico sendo transmitido ao longo do tempo (Figura 3.1). Os enterramentos atômicos, bem como a informação estrutural de uma proteína, estão definidos dentro de um espaço contínuo (analógico) de possibilidades, ou seja, todas as possíveis distâncias entre o centro geométrico de uma estrutura até um determinado átomo, ou todas as suas possibilidades de coordenadas atômicas. Contudo a informação estrutural de uma proteína é transmitida através de um canal discreto e com capacidade limitada: a sequência de aminoácidos. Podemos pensar que ao longo do tempo as sequências proteicas tenham sido selecionadas de modo que esta informação seja transmitida de forma eficiente, codificando o sinal estrutural de uma maneira que ele seja compatível com a sequência de aminoácidos e que possa ser posteriormente recuperado.

A representação de sinais analógicos através de sinais digitais é extensivamente empregada em diversos contextos de engenharia, tal como transmissão de voz, codificação de áudio, armazenamento de dados e no uso de dispositivos que envolvem circuitos digitais de forma geral, como computadores. Uma técnica frequentemente utilizada na digitalização de sinais analógicos é a modulação por código de pulsos e suas variações, dentre as quais a modulação por código de pulsos diferencial (*Differential Pulse-Code Modulation*), proposta por Cassius C. Cutler em 1950, que utiliza as diferenças entre valores amostrados, ao invés

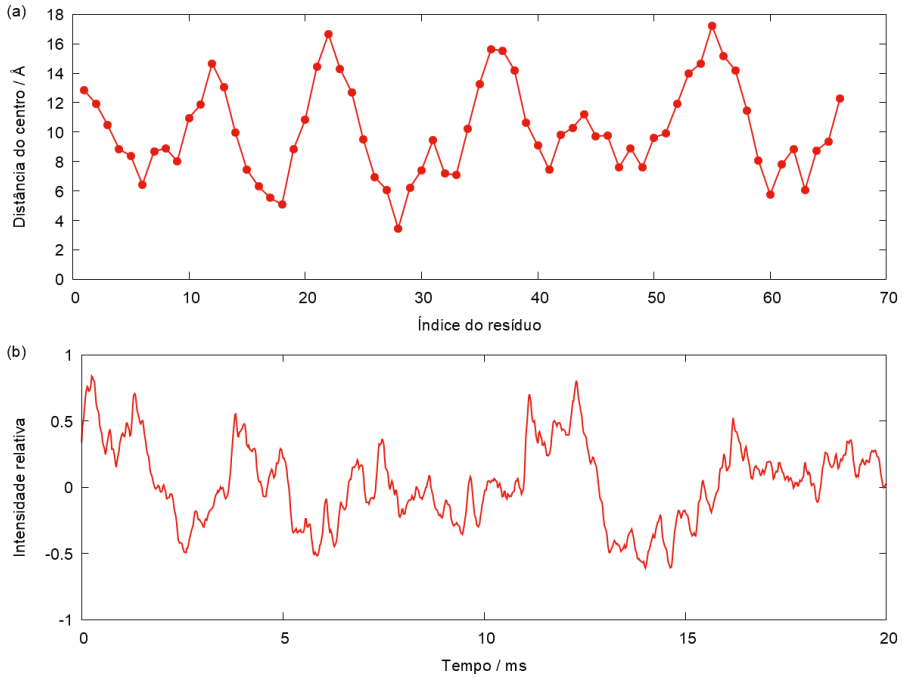

Figura 3.1. (a) Enterramentos atômicos de C_α ao longo da cadeia de resíduos de uma estrutura e (b) sinal de áudio digital ao longo do tempo, digitalizado através de técnicas de modulação por código de pulsos.

dos valores absolutos, como método de codificação de um sinal(51). A digitalização de sinais analógicos é interessante no contexto das telecomunicações, visto que ela permite compressão sem perdas dos dados, e o uso de diferenças de sinais possibilita que sua transmissão seja feita sob altas taxas de compressão, dada a menor entropia devido à quantidade reduzida de símbolos na codificação (41, 52).

Até então, temos utilizado o modelo que representa os enterramentos atômicos na forma de camadas equiprováveis de enterramento (Figura 3.2). Este é um meio de se digitalizar o sinal estrutural proveniente dos enterramentos atômicos onde, para cada átomo, é associada uma camada de enterramento dentro de um espaço finito de possibilidades (o número de camadas considerado para a estrutura em questão). Esta representação, que denota a posição **absoluta** de cada átomo na sequência, nos permite, por exemplo, realizar as predições baseadas em HMM apresentadas anteriormente e tem mostrado resultados interessantes no contexto do problema do enovelamento. Entretanto, a representação por camadas possui algumas limitações: a escolha do número de camadas é arbitrário e pode não ser adequado ao tamanho da estrutura com a qual se está trabalhando. A largura de cada camada pode variar, por definição, dependendo do tamanho da proteína e o número de camadas necessárias para a descrição da estrutura aumenta com o tamanho da

sequência. Consequentemente, ao avaliarmos estruturas cada vez maiores precisaremos também de cada vez mais símbolos para sua descrição, o que resultará inevitavelmente no comprometimento da estatística e está limitado pela capacidade computacional disponível.

Apresentaremos nesta seção uma nova forma de representação dos enterramentos atômicos, inspirada em tecnologias de processamento de sinais amplamente adotadas, a qual denominaremos de **representação diferencial dos enterramentos atômicos**. Esta metodologia consiste em associarmos à sequência de átomos da cadeia avaliada uma sequencia de símbolos, também escolhidos dentre um alfabeto finito, que denote a posição **relativa** entre dois átomos desta cadeia, ou seja, uma sequência de símbolos que expressam se um átomo está mais ou menos enterrado que o anterior (Figura 3.2). A partir da sequência de diferenças de enterramento (posições relativas) também é possível recuperarmos uma segunda sequência contendo os níveis de enterramento dos átomos da cadeia (posições absolutas), que se trata de uma descrição dos enterramentos análoga à sequência de camadas utilizada anteriormente.

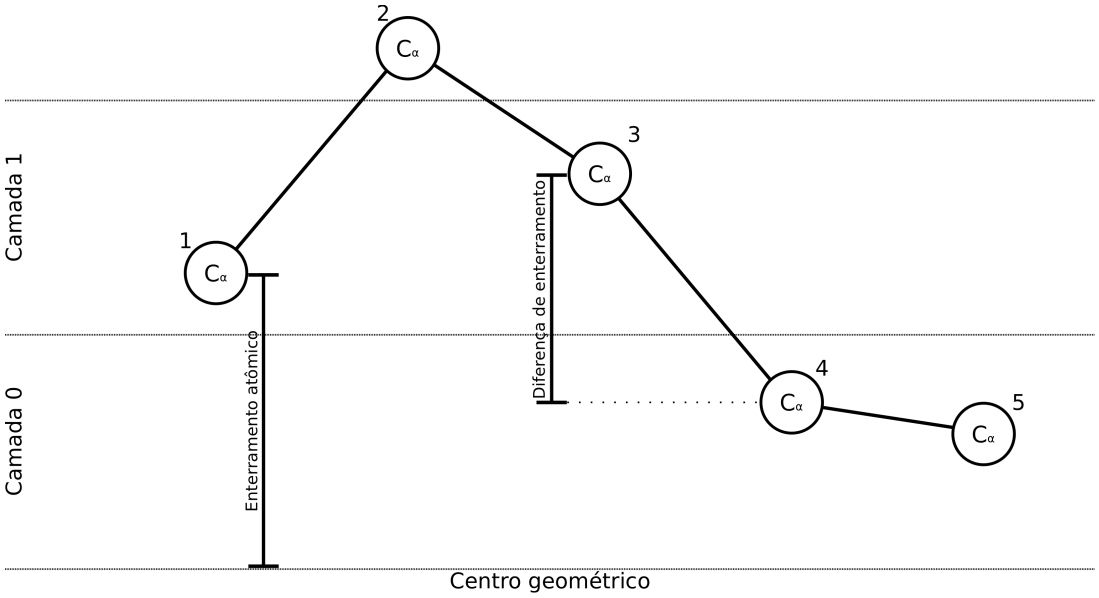

Figura 3.2. Esquema de representações de enterramentos atômicos. A representação por camadas equiprováveis quantifica as distâncias absolutas e são divididas de forma que cada uma tenha o mesmo número de átomos. A representação diferencial quantifica as distâncias relativas entre dois átomos, indicando se um átomo está mais enterrado ou mais exposto do que um átomo vizinho.

Exporemos agora as definições e o algoritmo em termos formais. Seja $\boldsymbol{A} = (a_1, \ldots, a_n), a_i \in \mathcal{Q}$ uma cadeia finita de átomos formando uma sequência de n elementos definida no alfabeto $\mathcal{Q}$ de tipos de átomo. Seja $\boldsymbol{B} = (b_1, \ldots, b_n), b_i \in \mathbb{R}$ a sequência dos

respectivos enterramentos atômicos de $\boldsymbol{A}$ expressos na forma de distâncias ao centro geométrico da estrutura, definidos no conjunto dos números reais.

Queremos obter uma sequência de diferenças $\boldsymbol{D} = (d_1, d_2, \ldots, d_{n-1}), d_i \in \mathcal{D}$ cujos símbolos estão definidos no alfabeto $\mathcal{D}$, de tal forma que d_i seja um descritor da diferença entre dois átomos adjacentes, b_{i-1} e b_i. A partir de $\boldsymbol{D}$, deve ser possível recuperarmos uma sequência $\boldsymbol{S} = (s_1, \ldots, s_n), s_i \in \mathbb{R}$ de níveis de enterramento atômico definidos no conjunto dos números reais de modo que s_i seja um descritor do valor de b_i. Chamaremos o processo de construção da sequência $\boldsymbol{D}$ de modulação e será realizado por um algoritmo que será detalhado posteriormente.

Por fim, o processo de geração da sequência $\boldsymbol{S}$ a partir de $\boldsymbol{D}$ será chamado de demodulação, e consiste simplesmente na acumulação dos valores de $\boldsymbol{D}$ ao longo da sequência. Se escolhermos $\mathcal{D} \subset \mathbb{Z}$, então, os elementos de $\boldsymbol{S}$ serão dados por

$$s_i = \begin{cases} 0 & \text{para } i = 1 \\ s_{i-1} + d_{i-1} & \text{para } 2 \leq i \leq n \end{cases} \tag{3.1}$$

É importante ressaltar que o valor inicial s_1 pode ser atribuído arbitrariamente uma vez que a escolha de valores diferentes acarreta apenas em uma transformação linear da sequência $\boldsymbol{S}$ deslocando-a para cima ou para baixo. Inclusive, é necessário aplicar tal transformação para que a sequência $\boldsymbol{S}$ seja expressa na mesma escala que os enterramentos atômicos a partir dos quais ela foi gerada.

Apresentaremos agora a descrição detalhada do algoritmo de escolha dos símbolos da sequência de diferenças. Sejam $\mathcal{D}_2 = \{-1, 1\}$ e $\mathcal{D}_3 = \{-1, 0, 1\}$ os alfabetos nos quais discretizaremos as diferenças de enterramento, onde o valor -1 representa que um átomo está mais enterrado que o anterior, 0 representa que está tão enterrado quanto e 1 representa que está mais exposto. O algoritmo é dado por

$$d_i = \begin{cases} -1 & \text{se } b_{i+1} \leq b_1 + ks_i \\ 1 & \text{se } b_{i+1} > b_1 + ks_i \end{cases} \tag{3.2}$$

no caso do alfabeto de 2 direções, e por

$$d_i = \begin{cases} -1 & \text{se } b_{i+1} < b_1 + ks_i - t \\ 0 & \text{se } (b_1 + ks_i - t) \leq b_{i+1} \leq (b_1 + ks_i + t) \\ 1 & \text{se } b_{i+1} > b_1 + ks_i + t \end{cases} \tag{3.3}$$

no caso do alfabeto de 3 direções. As constantes k e t controlam, respectivamente, o tamanho de cada incremento na representação e a proporção dos pares de átomos que serão considerados igualmente enterrados. Neste trabalho estudaremos a sequência de átomos dada pelos C_α, onde o maior valor possível de diferenças de enterramento é de 3.8Å. Portanto, utilizaremos os valores $k = \frac{3.8}{2} = 1.9$ e $t = \frac{3.8}{3} = 1.267$, que são os valores que resultam em uma distribuição aproximadamente equiprovável dos símbolos dos alfabetos de direções nos bancos de dados avaliados. Por fim, para colocarmos $\boldsymbol{S}$ novamente na escala da proteína em questão, realizaremos a seguinte operação sobre os elementos s_i de $\boldsymbol{S}$:

$$s'_i = k \cdot s_i + (R_e \cdot R_m) \tag{3.4}$$

onde $R_e = 2.7\sqrt[3]{n}$ é uma estimativa do raio de giro de uma proteína de tamanho n, descrita por Gomes et al. (37) e $R_m = 0.96$ é uma estimativa da mediana dos valores de enterramento atômico normalizados entre 0 e 2 descrita no mesmo artigo. Para compararmos os resultados obtidos com a representação diferencial e os resultados anteriores provenientes da representação por camadas, escolhemos algumas estruturas especialmente selecionadas a fim de abrangermos diferentes faixas de tamanho e classes estruturais. Estas estruturas são apresentadas na Figura 3.3.

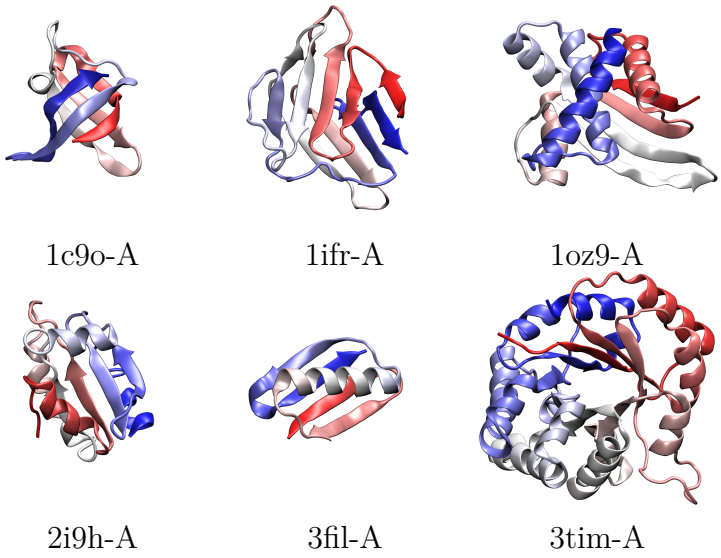

Figura 3.3. Estruturas para as quais apresentamos o sinal de enterramentos obtido a partir da representação diferencial e seus respectivos códigos PDB e identificadores da cadeia.

Apresentamos na Figura 3.4 uma comparação entre os sinais obtidos por meio da representação diferencial de enterramentos, o sinal dos enterramentos nativos e a representação em 4 camadas equiprováveis, para a proteína 1ifr-A. Como forma de compararmos a qualidade dos sinais de enterramento nativo obtidos pelas diferentes representações, calculamos o desvio quadrático médio (RMSD) entre o sinal de enterramento real e aquele proveniente de cada uma das representações.

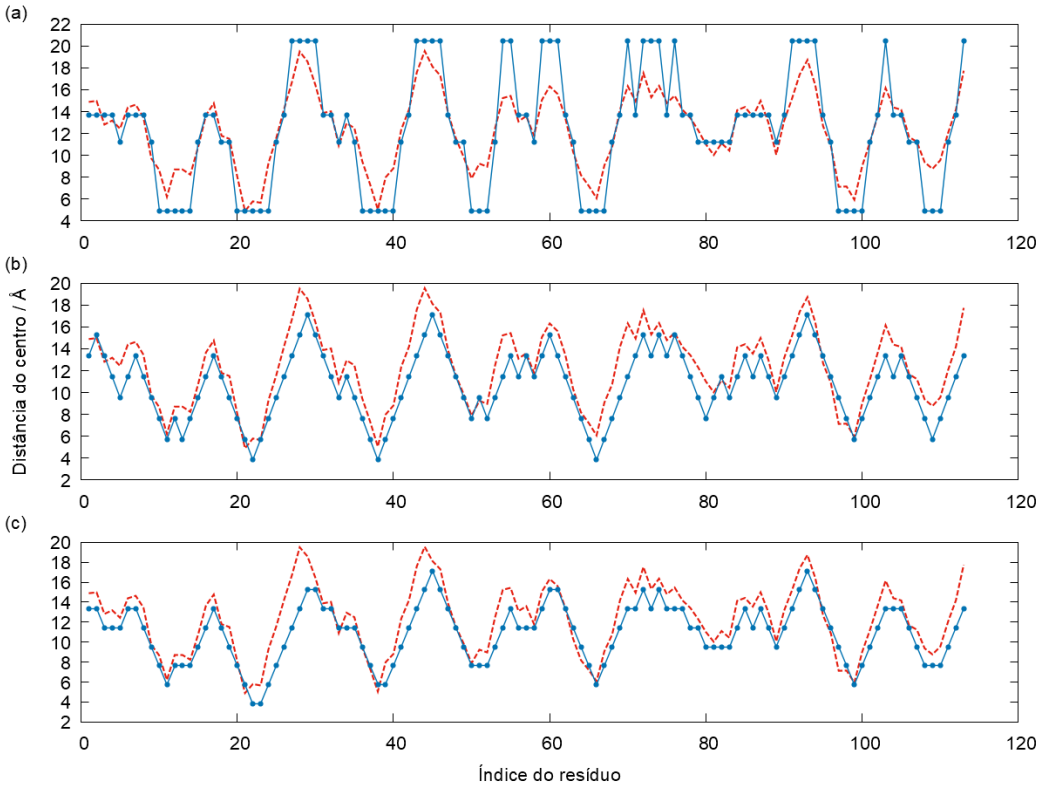

Figura 3.4. Proteína 1ifr-A. Sinais de enterramento atômico obtidos através da (a) representação por 4 camadas equiprováveis, RMSD=2.5Å, (b) representação diferencial com 2 direções, RMSD=1.4Å, e (c) representação diferencial com 3 direções, RMSD=1.4Å.

O sinal proveniente da representação por camadas equiprováveis apresenta um RMSD de 2.5 Å com o sinal real e necessita de 4 símbolos no alfabeto. Embora uma resolução maior poderia ser obtida utilizando-se mais camadas, o tamanho do alfabeto aumenta linearmente com o número de camadas na representação. Por outro lado, os sinais provenientes das representações diferenciais obtiveram ambas um RMSD de 1.4 Å com os enterramentos reais utilizando apenas 2 e 3 símbolos no alfabeto da sequência. Este ganho de resolução também é observado nas outras proteínas analisadas, conforme mostra a Tabela 3.1.

Tabela 3.1. Desvio quadrático médio (RMSD) entre o sinal de enterramento nativo e os sinais provenientes das diferentes representações de enterramento atômico para as proteínas apresentadas.

Código PDB	RMSD / Å		
	4 camadas	Diferencial (2 símbolos)	Diferencial (3 símbolos)
1c9o-A	2.122	1.902	2.036
1ifr-A	2.482	1.362	1.387
1oz9-A	2.437	2.342	2.027
2i9h-A	2.516	1.845	1.615
3fil-A	1.742	1.489	1.352
3tim-A	3.151	2.142	1.877

3.2 Predições de enterramento

Uma vez definida esta nova forma de representarmos os enterramentos atômicos, podemos utilizá-la para realizar predições de estrutura a partir da sequência. As predições apresentadas nesta seção foram feitas através do *software* HMMPred, publicado por van der Linden em seu trabalho de 2013(49). O algoritmo consiste em duas etapas: uma etapa de treinamento, na qual as matrizes de transição são preenchidas com as probabilidades inferidas a partir de um banco de dados, e uma etapa de predição, na qual as probabilidades de cada símbolo do alfabeto dos estados escondidos são calculadas para uma dada a sequência de resíduos avaliada.

Para a realização da etapa de treinamento, foi utilizada a lista de sequências obtida a partir do PDBSelect(43) de 2009, que consiste em um banco de estruturas selecionadas do *Protein Data Bank* - PDB(53) que tem como objetivo minimizar repetições nas sequências de aminoácidos, selecionando-se apenas sequências representativas com menos de 25% de identidade na estrutura primária. Além disso também foram executados os seguintes tratamentos adicionais sobre a lista obtida: aplicado o filtro de identidade (30%) fornecido pelo próprio PDB através da ferramenta BLAST/BLASTClust (54), removidas as proteínas de membrana, removidas as proteínas não globulares (37) e removidas as cadeias cuja sequência primária estivesse incompleta, ou seja, faltando as coordenadas do C_α para algum resíduo (possuem "gaps"). Por fim , as cadeias foram posteriormente separadas em dois grupos conforme o tamanho: entre 50 e 120 resíduos, totalizando 386 cadeias; entre 120 e 250 resíduos, totalizando 442 cadeias. Estes foram os conjuntos de fato utilizados na etapa de treinamento do algoritmo. O filtro que removeu as sequências com "gaps" foi o que causou a maior redução no tamanho dos conjuntos de treinamento, porém este filtro é necessário para que possamos calcular efetivamente as diferenças de enterramento entre os resíduos.

As predições foram feitas a partir da estrutura primária das cadeias avaliadas, considerando-se apenas o enterramento do C_α para cada resíduo. A representação das sequências de enterramento atômico associadas a cada sequência de resíduos foi feita através de símbolos da representação diferencial dos enterramentos atômicos, onde os seguintes modelos foram testados: representação com 2 direções de enterramento (mais ou menos enterrado) e representação com 3 direções de enterramento (mais, menos, ou

igualmente enterrado). Cada direção na representação diferencial compõe uma relação entre dois resíduos adjacentes, mas o programa HMMPred requer que a sequência de símbolos de enterramento tenha o mesmo tamanho que a sequência primária, o que potencialmente resultaria em uma escolha assimétrica na atribuição dos símbolos aos resíduos, ou seja, atribuir a cada resíduo ou a direção em relação ao anterior, ou em relação ao próximo; essa assimetria também incute na necessidade de se descartar os dados referentes aos resíduos de uma das extremidades da cadeia, dependendo da escolha, uma vez que não haveria uma direção de enterramento para com a qual ele poderia ser pareado. Para se resolver esta questão, foi utilizado um alfabeto em que cada símbolo denota tanto a direção de entrada quanto a de saída para cada resíduo (Figura 3.5). Para fins de comparação, foram realizadas também predições de camadas utilizando as seguintes configurações: 2 camadas de C_α, 3 camadas, 4 camadas, 2 camadas e direção do C_β, 3 camadas e direção do C_β, 4 camadas e direção do C_β.

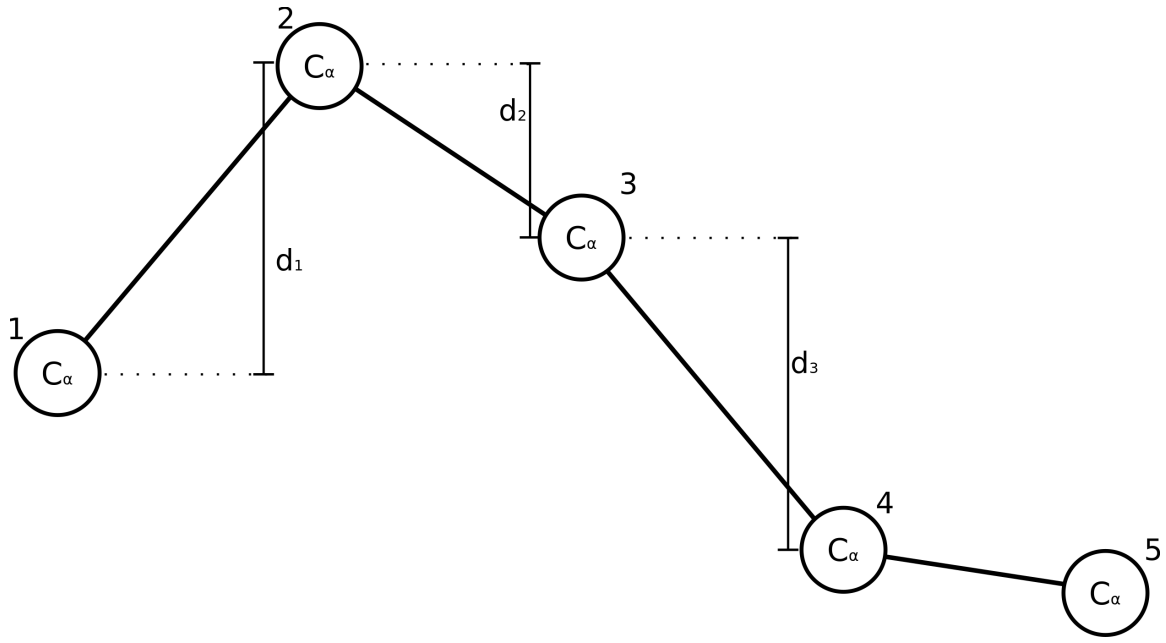

Figura 3.5. Para as predições, a fim de se manter a simetria, os resíduos foram associados a símbolos que compreendem tanto a direção de entrada quanto a direção de saída. Na imagem, o átomo 2 é associado a d_1 e d_2, e o átomo 3 é associado a d_2 e d_3.

Para a análise dos resultados, as probabilidades obtidas a partir das predições de direção foram marginalizadas com relação à direção relativa ao resíduo anterior,

$$P_a(\delta) = P_a(\delta_a) + P_a(\delta_{a+1}) \qquad \forall \delta \in \mathcal{D} \tag{3.5}$$

onde $P_a(\delta)$ é a probabilidade de um átomo a de C_α ser classificado com a direção δ em

um alfabeto $\mathcal{D}$ de direções, $P_a(\delta_a)$ é a probabilidade predita da direção ao seu antecessor e $P_a(\delta_{a+1})$ é a probabilidade predita da direção ao seu sucessor na cadeia.

Da mesma forma, as probabilidades obtidas a partir das predições de camadas com orientação do C_β foram marginalizadas, para cada átomo, com relação à camada do C_α,

$$P_a(\ell) = \sum_{\text{direções de } C_\beta} P_a(\ell) \qquad \forall \ell \in L \tag{3.6}$$

onde $P_a(\ell)$ é a probabilidade de um átomo a de C_α ser classificado na camada ℓ em um alfabeto de L camadas. Ou seja, somamos, para cada átomo, as probabilidades probabilidades das diferentes predições de orientação do C_β. As predições de camadas sem orientação do C_β não foram marginalizadas.

A qualidade das predições foi avaliada em termos de acurácia, informação da predição absoluta e relativa, coeficiente de correlação linear e desvio quadrático médio (RMSD) com os enterramentos observados. No caso das predições de camadas, consideramos como corretamente atribuída quando a camada de maior probabilidade predita coincide com a camada observada na estrutura nativa. A acurácia (f_c) então é dada pela razão entre o número de camadas corretamente atribuídas a cada resíduo (n_c) e o número total de resíduos (n),

$$f_c = \frac{n_c}{n} \tag{3.7}$$

e no caso das predições de direções, consideramos como corretamente atribuída quando a direção de maior probabilidade predita entre um par de resíduos adjacentes coincide com a direção observada na estrutura nativa, e a acurácia é dada pela razão entre o número de direções corretamente atribuídas e o número total de direções,

$$f_c = \frac{n_c}{n-1} \tag{3.8}$$

As medidas de informação da predição são calculadas através de uma estimativa da incerteza do conjunto de probabilidades obtida pelo *log likelihood* médio ($\langle LL \rangle$), conforme descrito por Moddemeijer (55), segundo a fórmula:

$$\langle LL \rangle = \frac{-\sum_{i=1}^{n} \log_2 p_{prd}(b_{nat}(i))}{n} \tag{3.9}$$

onde $p_{prd}(b_{nat}(i))$ é a probabilidade predita do descritor observado na estrutura nativa na posição i da sequência. A informação relativa da predição é definida pelo complemento da

razão entre $\langle LL \rangle$ e o valor máximo da entropia no alfabeto do descritor utilizado, dada por $1 - \frac{\langle LL \rangle}{\log_2 |\mathcal{B}|}$, onde $|\mathcal{B}|$ é o tamanho do alfabeto e a informação absoluta é dada por $\log_2 |\mathcal{B}| - \langle LL \rangle$. Na Tabela 3.2 estão sumarizados os para as predições no banco de cadeias entre 50 e 120 resíduos e na Tabela 3.3 para as predições no banco de cadeias entre 120 e 250 resíduos. As estimativas de erro apresentadas nesta Seção foram obtidas através de *bootstrapping* (56) com 50 reamostragens.

Tabela 3.2. Resultados das predições no banco de cadeias entre 50 e 120 resíduos para os diferentes alfabetos de enterramento considerados. Os dados estão agrupados conforme o número de símbolos em cada representação.

Representação dos enterramentos	Acurácia / %	Informação da predição / *bits*	Informação relativa da predição / %
Diferencial (2 símbolos)	70.6 ± 0.3	0.186 ± 0.005	18.6 ± 0.5
2 camadas	70.8 ± 0.3	0.188 ± 0.005	18.8 ± 0.5
2 camadas + direção C_β	71.3 ± 0.3	0.199 ± 0.005	19.9 ± 0.5
Diferencial (3 símbolos)	55.6 ± 0.3	0.261 ± 0.007	16.5 ± 0.4
3 camadas	54.2 ± 0.4	0.252 ± 0.008	15.9 ± 0.5
3 camadas + direção C_β	54.6 ± 0.4	0.260 ± 0.008	16.4 ± 0.5
4 camadas	43.7 ± 0.4	0.284 ± 0.007	14.2 ± 0.4
4 camadas + direção C_β	44.1 ± 0.4	0.283 ± 0.008	14.1 ± 0.4

Tabela 3.3. Resultados das predições no banco de cadeias entre 120 e 250 resíduos para os diferentes alfabetos de enterramento considerados. Os dados estão agrupados conforme o número de símbolos em cada representação.

Representação dos enterramentos	Acurácia / %	Informação da predição / *bits*	Informação relativa da predição / %
Diferencial (2 símbolos)	68.5 ± 0.2	0.149 ± 0.003	14.9 ± 0.3
2 camadas	68.2 ± 0.3	0.143 ± 0.003	14.3 ± 0.3
2 camadas + direção C_β	68.8 ± 0.2	0.149 ± 0.004	14.9 ± 0.4
Diferencial (3 símbolos)	53.9 ± 0.3	0.219 ± 0.004	13.8 ± 0.3
3 camadas	51.6 ± 0.3	0.197 ± 0.004	12.4 ± 0.3
3 camadas + direção C_β	51.4 ± 0.3	0.195 ± 0.004	12.3 ± 0.3
4 camadas	41.3 ± 0.3	0.220 ± 0.004	11.0 ± 0.2
4 camadas + direção C_β	40.9 ± 0.3	0.213 ± 0.005	10.7 ± 0.2

Observamos que, tanto em termos de acurácia quanto em termos de informação e incerteza, a qualidade das predições é comparável entre as diferentes representações, quando utilizados o mesmo número de símbolos. A qualidade das predições com orientação de C_β para 2 camadas evidencia também a importância de incluirmos informação sobre as

cadeias laterais, que não foram consideradas nas predições de direções. Há, contudo, certas limitações ao se utilizar alfabetos com um número grande de símbolos decorrentes tanto da complexidade do algoritmo, o que se reflete na capacidade computacional, quanto do espaço amostral dos bancos de dados usados para o treinamento do programa. Este é o caso da saturação na estatística das predições de 4 camadas e orientação de C_β, cujo alfabeto usado na predição possui 8 símbolos. Embora a representação de 3 direções utilize 9 símbolos na predição, este efeito é atenuado pois os símbolos são altamente correlacionados, mas o impacto da complexidade do algoritmo permanece, impossibilitando a adição de mais descritores nestas predições. De todo modo estes resultados mostram que não há limitações no uso deste algoritmo de predição para a representação diferencial.

A predição de diferenças de enterramento, entretanto, não deve ser a predição final do sinal. Estamos interessados em recuperar a partir dela a sequência de níveis que expressa o valor de enterramento absoluto dos átomos na cadeia. No caso do sinal observado isso é feito de maneira trivial, uma vez que só há uma direção relativa associada a cada resíduo. Contudo, no caso das predições temos para cada resíduo uma distribuição de probabilidades dentre os possíveis símbolos do alfabeto utilizado. Um primeiro teste seria simplesmente considerar a direção de maior probabilidade para construirmos um traço de níveis de enterramento para cada cadeia, da mesma forma que é feito quando realizamos predições com camadas de enterramento, e que foi o critério utilizado no cálculo das acurácias das predições. A construção do traço $\boldsymbol{S}$ é feita da mesma forma que o sinal quando tratamos dos enterramentos nativos, dada por

$$s_i = \begin{cases} 0 & \text{para } i = 1 \\ s_{i-1} + d_{i-1} & \text{para } 2 \leq i \leq n \end{cases} \tag{3.10}$$

Calculamos em seguida os coeficientes de correlação linear e o RMSD entre a sequência de enterramentos nativos e o sinal predito para cada cadeia. Nas predições de camadas, o sinal é dado pelo centro das camadas de maior probabilidade de cada resíduo. A distância do centro das camadas até o centro geométrico da respectiva estrutura é estimada através da curva de distribuição dos enterramentos atômicos, conforme descrito por Gomes et al. (37) e por Linden et al. (36) e também discutido na introdução desta teste (Equação 1.2). Os resultados estão sumarizados na Tabela 3.4.

Observamos que as predições com menor RMSD, e portanto aquelas cujos sinais de

Tabela 3.4. Correlação e RMSD médios entre os enterramentos reais e o traço obtido da predição para os diferentes alfabetos de enterramento considerados.

	Representação dos enterramentos	Tamanho / resíduos 50 a 120	120 a 250
Correlação	Diferencial (2 símbolos)	0.31 ± 0.01	0.225 ± 0.009
	2 camadas	0.483 ± 0.006	0.428 ± 0.005
	2 camadas + direção C_β	0.492 ± 0.007	0.436 ± 0.005
	Diferencial (3 símbolos)	0.33 ± 0.01	0.234 ± 0.009
	3 camadas	0.532 ± 0.007	0.470 ± 0.005
	3 camadas + direção C_β	0.534 ± 0.007	0.465 ± 0.005
	4 camadas	0.544 ± 0.007	0.480 ± 0.007
	4 camadas + direção C_β	0.539 ± 0.008	0.466 ± 0.007
RMSD / Å	Diferencial (2 símbolos)	5.06 ± 0.09	7.2 ± 0.1
	2 camadas	5.36 ± 0.03	6.98 ± 0.04
	2 camadas + direção C_β	5.33 ± 0.04	6.94 ± 0.04
	Diferencial (3 símbolos)	4.53 ± 0.06	6.7 ± 0.1
	3 camadas	4.87 ± 0.04	6.49 ± 0.05
	3 camadas + direção C_β	4.87 ± 0.04	6.53 ± 0.04
	4 camadas	4.65 ± 0.05	6.26 ± 0.04
	4 camadas + direção C_β	4.68 ± 0.04	6.34 ± 0.04

enterramento mais se aproximam dos sinais reais são aquelas realizadas com a representação diferencial de 3 símbolos sobre o conjunto de proteínas pequenas, conjunto este que apresenta maior taxa de acerto nas predições em geral. Notamos, entretanto, que esta não é a melhor maneira de recuperarmos o sinal de enterramento a partir das predições, como evidenciado pela correlação reduzida com os enterramentos nativos. As predições realizadas com a representação diferencial são muito sensíveis a erros, o que faz com que erros pontuais nas atribuições das direções reduzam drasticamente a qualidade das predições. Isso se deve ao fato de que a construção do traço é feita de maneira sequencial e os níveis de enterramento de todos os resíduos além do primeiro são função do nível do resíduo anterior. Sendo assim, a escolha da direção para um dado resíduo afeta a atribuição dos níveis de enterramento para todos os resíduos subsequentes na cadeia (Figura 3.6).

Podemos evidenciar este efeito acontecendo no conjunto inteiro de proteínas se filtrarmos os resultados removendo aquelas cadeias cujas predições têm baixa correlação com o sinal de enterramento nativo. Na Tabela 3.5 mostramos os resultados para as cadeias em que o sinal predito e o sinal real apresentam correlação linear superior a 0.3

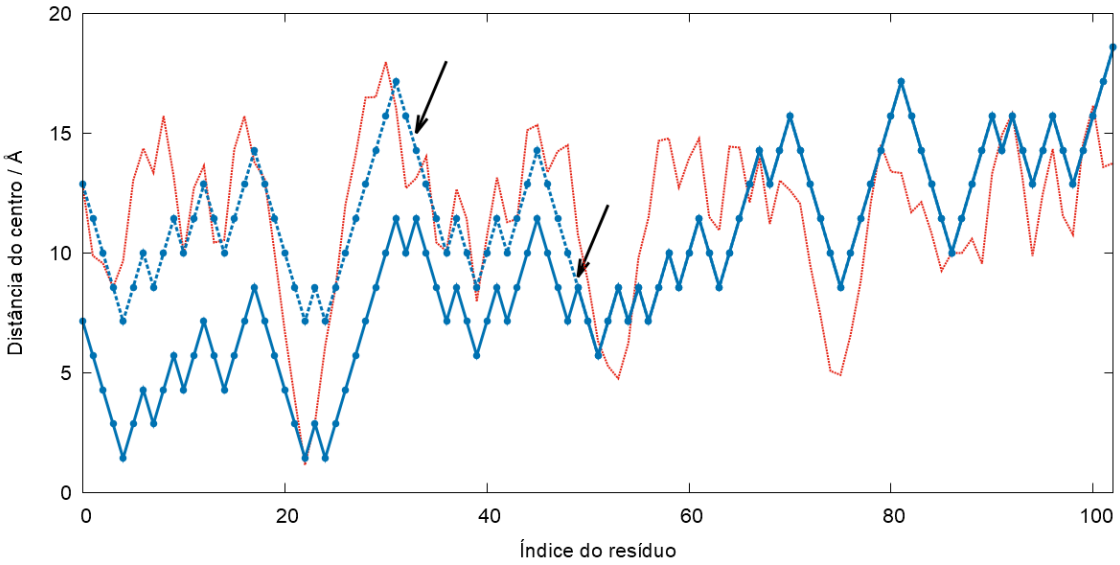

Figura 3.6. Sinal recuperado a partir das direções de maior probabilidade (linha contínua) para a proteína 2i9h-A, apresenta correlação com os enterramentos nativos (linha vermelha) de 0.38 e RMSD de 5.42Å. Se a predição tivesse atribuído corretamente as direções indicadas pelas setas, obteríamos um sinal com correlação de 0.60 e RMSD de 3.25Å (linha tracejada).

simultaneamente na predição diferencial de 3 símbolos e na predição com 4 camadas e direção da cadeia lateral. O valor de 0.3 foi escolhido de modo que aproximadamente metade do conjunto de proteínas permanecesse na análise.

Tabela 3.5. RMSD entre os enterramentos reais e o traço obtido da predição para os melhores alfabetos de enterramento considerados. Aqui foram analisadas apenas as cadeias cuja correlação entre o sinal predito e o sinal real foi superior a 0.3 em ambos os métodos simultaneamente.

Representação dos enterramentos		RMSD / Å	
		Sem filtro	Com filtro
50 a 120 resíduos	Diferencial (3 símbolos)	4.53 ± 0.06	3.91 ± 0.06
	4 camadas + direção C_β	4.68 ± 0.04	4.46 ± 0.04
120 a 250 resíduos	Diferencial (3 símbolos)	6.7 ± 0.1	5.22 ± 0.09
	4 camadas + direção C_β	6.34 ± 0.04	5.97 ± 0.07

Aqui fica claro que o filtro afeta positivamente mais as predições da representação diferencial do que as predições por camadas, reforçando a ideia de que esta representação é especialmente sensível a erros. A representação diferencial apresenta, no entanto, maior potencial para a obtenção de um sinal mais preciso caso possamos tratar adequadamente estes erros pontuais, como evidenciado pela melhor resolução, das predições de boa qualidade. Surge, portanto, a necessidade de que a atribuição das direções seja compatível com toda a distribuição de probabilidades preditas, não só em cada resíduo, mas na

sequência inteira. Neste contexto, a recuperação do sinal a partir da sequência de diferenças passa a ser um problema de minimização que pode ser tratado por meio de um algoritmo de otimização numérica, cujos resultados serão apresentados na Seção 3.3.

3.3 Otimização das predições

A recuperação do sinal de enterramento é um passo sensível às atribuições de direções realizadas ao longo da cadeia, dada a sua natureza recursiva, e um erro em uma determinada posição na sequência se propaga para todos os resíduos subsequentes. A utilização de um algoritmo de otimização matemática nos possibilitará buscar no espaço de traços de enterramento um sinal compatível com toda a distribuição de probabilidades obtida das predições, diminuindo a arbitrariedade na atribuição das direções preditas. Além disto, o uso de um algoritmo deste tipo também permitirá combinarmos resultados de diferentes predições, devido ao potencial efeito sinergético entre descritores distintos, conforme observado nos resultados de predições de 2 camadas com orientação do C_β (Tabelas 3.2 e 3.4).

Implementamos para este fim um método de Monte Carlo que permitirá explorar este espaço de traços de enterramento através de simulações rápidas. Utilizamos como termos de energia a ser minimizada funções dependentes das distribuições de probabilidades obtidas diretamente das predições apresentadas. O programa foi implementado com a possibilidade de aceitar mais de uma função de energia simultaneamente, permitindo que seja possível combinar o resultado de predições diferentes para uma dada proteína.

O seguinte algoritmo foi utilizado: iniciamos com um estado inicial do sinal dado por $\boldsymbol{S} = (s_1, \ldots, s_n)$. Este estado possui uma sequência de diferenças, ou direções, dada por $\boldsymbol{D} = (d_1, \ldots, d_{n-1}) = (s_2 - s_1, \ldots, s_n - s_{n-1}), d_i \in \mathcal{D}$. A cada iteração, escolhemos ao acaso 3 posições w_1, w_2 e w_3, $1 \leq w < n$, tal que $\boldsymbol{D} = (\ldots, d_{w_1}, \ldots, d_{w_2}, \ldots)$. Então geramos uma sequência $\boldsymbol{D'}$ segundo os passos

$$\boldsymbol{D'} = (d'_1, \ldots, d'_{n-1}) \leftarrow \boldsymbol{D} \qquad \text{inicialmente } \boldsymbol{D'} = \boldsymbol{D} \tag{3.11}$$

$$d'_{w_1} \leftarrow d_{w_2} \qquad \text{atribuímos } d_{w_2} \text{ para } d'_{w_1} \tag{3.12}$$

$$d'_{w_2} \leftarrow d_{w_1} \qquad \text{atribuímos } d_{w_1} \text{ para } d'_{w_2} \tag{3.13}$$

$$d'_{w_3} \leftarrow \delta \qquad \text{com } \delta \in \mathcal{D} \text{ escolhido ao acaso} \tag{3.14}$$

essencialmente trocando os valores entre duas posições na sequência de diferenças e então modificando uma terceira posição para um valor aleatório, conforme ilustrado na Figura 3.7. Todas as escolhas ao acaso seguem a distribuição uniforme.

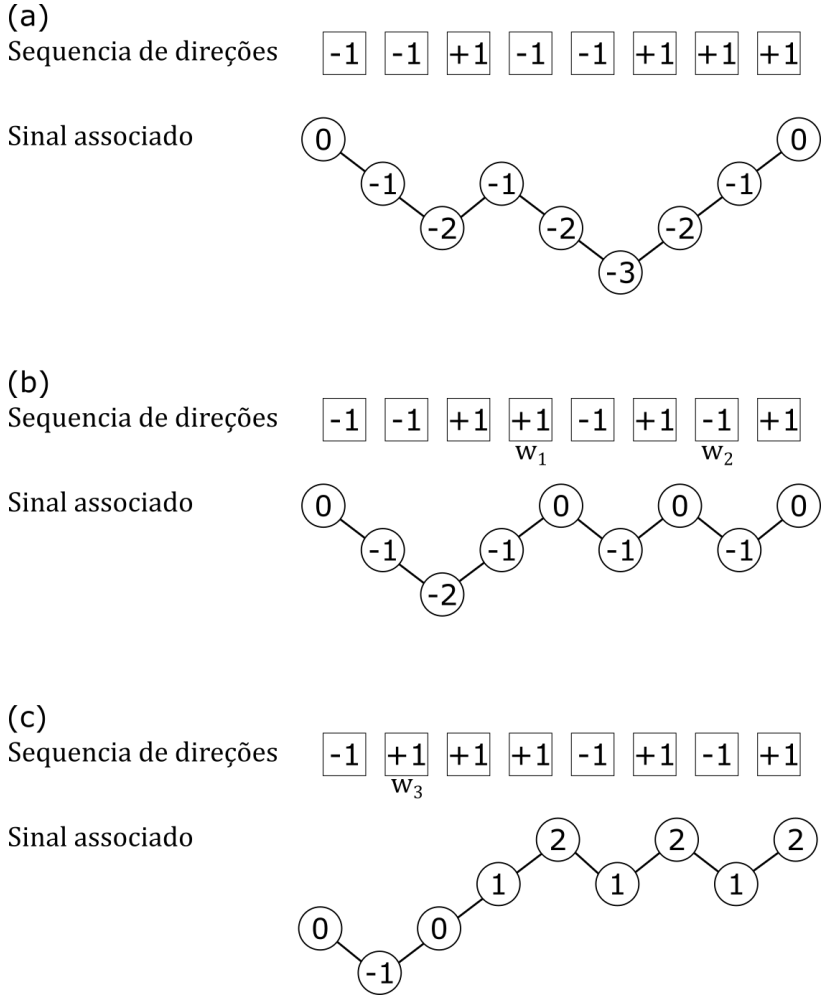

Figura 3.7. (a) Estado inicial, proveniente da predição pelo HMM. (b) Primeira modificação da sequencia de diferenças: duas posições ao acaso, w_1 e w_2 têm seus valores trocados entre si (Equações 3.12 e 3.13). Este movimento não altera a proporção de símbolos no sinal, uma vez que se as posições escolhidas tiverem o mesmo símbolo, nenhuma modificação será feita. (c) Segunda modificação da sequencia de diferenças: uma terceira posição w_3 tem seu símbolo trocado ao acaso (Equaçãos 3.14). Este movimento altera a porporção de símbolos na sequência de diferenças. O sinal associado a (c) é o sinal submetido aos cálculos de energia (Equação 3.17 e Equações 3.18, 3.19 e 3.20) e pode ou não ser aceito para a próxima iteração do algoritmo.

Cada sequência de diferenças modificada também possui um sinal associado

$$\boldsymbol{S'} = (s'_1, s'_2, \cdots, s'_n) \tag{3.15}$$

$$= (0, s'_1 + d'_1, \ldots, s'_{n-1} + d'_{n-1}) \tag{3.16}$$

que é submetido aos cálculos de energia após todos os passos de modificação da sequência de diferenças. A Figura 3.7 ilustra como o algoritmo é aplicado sobre uma sequencia de diferenças e como isso se reflete no sinal de enterramento associado a ela. Calculamos então a energia dos estados, onde a energia total de cada estado é dada pela soma de todos os termos de energia (utilizados na simulação) daquele estado

$$E = E_1 + E_2 + \cdots \tag{3.17}$$

Três termos de energia foram desenvolvidos para este trabalho, dois dos quais apresentamos a seguir:

$$E_l = -\sum_{i=1}^{n} \ln(p_{prd}(s_i)) \qquad \text{energia das camadas} \tag{3.18}$$

$$E_d = -\sum_{i=1}^{n} \ln(p_{prd}(d_i)) \qquad \text{energia das direções} \tag{3.19}$$

onde p_{prd} é a probabilidade computada pelo HMM para um dado símbolo, e um terceiro que será apresentado posteriormente. Por fim, aplicamos o critério de Metropolis (57) para aceitar ou rejeitar o novo estado:

- se $E' \leq E$, então $\boldsymbol{S} \leftarrow \boldsymbol{S'}$
- se $E' > E$, então $\boldsymbol{S} \leftarrow \boldsymbol{S'}$ com probabilidade $e^{-\frac{(E'-E)}{T}}$, e T é um fator de temperatura que pode ser controlado.

Selecionamos algumas cadeias com as quais já trabalhamos anteriormente para submetermos às simulações Monte Carlo, de forma que diferentes tamanhos e classes estruturais estejam representados. As cadeias escolhidas foram (código PDB - cadeia): 1c9o-A, 1ifr-A, 1oz9-A, 2i9h-A, 3fil-A, 3tim-A. Otimizamos as probabilidades das predições diferenciais com 2 direções e das predições com 2 camadas com direção de C_β, uma vez que estas foram as predições que apresentaram a maior informação relativa, indicando que o conjunto de probabilidades nestas condições deve descrever de maneira mais adequada os enterramentos nativos. Também testamos a aplicação de um termo adicional que procura aproximar a distribuição de enterramentos do sinal à distribuição global dos enterramentos atômicos no banco de dados conforme publicado por Gomes et al. (37). A energia deste termo global é proporcional à divergência de Kullback-Leibler (58, 52) entre a distribuição

global descrita pela função de densidade de probabilidade dos enterramentos atômicos (Equação 1.2, Figura 1.2) e a distribuição dos níveis no sinal em uma determinada iteração:

$$E_g = n \cdot D_{KL}(L||P) \tag{3.20}$$

$$= n \cdot \sum_{r=\min(S)}^{\max(S)+1} l(r) \ln \frac{l(r)}{g(r)} \tag{3.21}$$

onde L é a distribuição de frequências $l(r)$ de níveis de enterramento r em um sinal $\boldsymbol{S}$ de tamanho n. Esta distribuição de frequências é inicialmente aquela que provém diretamente da predição por HMM, mas pode mudar ao longo das iterações. P é a distribuição global dos enterramentos atômicos (Equação 1.2, Figura 1.2) e $p(r)$ é a probabilidade do nível r obtida a partir dela:

$$p(r) = \int_{\frac{2r}{\max(\boldsymbol{S})+1}}^{\frac{2(r+1)}{\max(\boldsymbol{S})+1}} P(r)dr \tag{3.22}$$

As cadeias foram submetidas a 20 simulações em cada uma das configurações, com temperatura $T = 1$. O número de iterações nas simulações foi de $2500n$, e como sinal predito foi considerada a média dos decis de menor energia obtidos para cada simulação. Apresentamos os coeficientes de correlação entre o sinal predito e os enterramentos nativos na Tabela 3.6. Nesta tabela também apresentamos, para fins de comparação, os coeficientes das predições de maior correlação nos bancos de dados sem o uso do método Monte Carlo, que são as predições de 4 camadas (Tabelas 3.2 e 3.2).

Tabela 3.6. Correlação entre os enterramentos reais e o traço obtido das simulações Monte Carlo para as diferentes proteínas avaliadas usando diferentes combinações de funções de energia minimizadas. TG: termo global dos enterramentos atômicos; D: termo proveniente das predições diferenciais de 2 direções; L: termo proveniente das predições de 2 camadas; as três últimas colunas contem as correlações com as predições não otimizadas, obtidas diretamente do HMM conforme apresentado na Seção 3.2. 2D: diferencial com 2 direções; 2L+C_β: 2 camadas e orientação do C_β; 4L: 4 camadas.

Código	Coeficiente de correlação									
	Sem TG			Com TG			Somente	HMM		
	D+L	D	L	D+L	D	L	TG	2D	2L+C_β	4L
1c9o-A	0.69	0.54	0.65	0.66	0.64	0.62	0.10	0.62	0.69	0.74
1ifr-A	0.83	0.68	0.72	0.81	0.74	0.79	0.00	0.51	0.73	0.71
1oz9-A	0.81	0.74	0.67	0.79	0.77	0.69	0.08	0.56	0.66	0.73
2i9h-A	0.54	0.31	0.41	0.61	0.57	0.50	0.14	0.38	0.54	0.54
3fil-A	0.64	0.50	0.43	0.59	0.57	0.61	-0.02	0.32	0.53	0.40
3tim-A	0.35	0.34	0.12	0.47	0.31	0.43	-0.06	0.22	0.45	0.58

As simulações utilizando somente o termo global (TG) dos enterramentos foram realizadas como controle, a fim de se mostrar que este termo não pode ser independentemente

responsável por qualquer melhora nas predições. De fato, estas simulações resultaram em sinais não correlacionados com os enterramentos nativos. Este termo caracteriza uma restrição independente da sequência que ocorre naturalmente nos enterramentos nativos devido a restrições físicas relacionadas ao volume dos átomos e à globularidade das estruturas e sua aplicação pode ser útil na redução de ruído dos sinais preditos a partir da sequência, o que fica evidenciado pelo aumento correlações dos sinais de alguns grupos de predição, conforme mostrado na Tabela 3.6, especialmente naqueles cujo sinal predito também é pouco correlacionado com os enterramentos nativos.

Analisando, por outro lado, os dados referentes às predições da representação diferencial, observamos que, especialmente quando acopladas ao TG, a otimização resulta em sinais mais correlacionados do que aqueles obtidos diretamente do HMM que consideram apenas a direção de maior probabilidade. Estes dados são compatíveis com a hipótese de que podemos extrair mais informação das predições ao considerarmos todo o conjunto de probabilidades na obtenção do sinal, e que essa informação deve ser necessária na recuperação do sinal a partir das direções, dada a natureza recursiva do algoritmo. O mesmo padrão não foi observado nas predições de camadas para estas proteínas, exceto nos casos da 1ifr-A e da 3fil-A acoplados ao TG, onde o sinal otimizado de 2 camadas com direção se mostrou mais correlacionado com os enterramentos nativos até mesmo que na predição direta de 4 camadas.

Em todos os casos, contudo, verificamos um efeito sinergético na determinação do sinal ao acoplarmos as predições de diferenças com as predições de camadas, uma vez que o sinal otimizado a partir das predições em conjunto se mostrou mais correlacionado com os enterramentos nativos do que o sinal otimizado a partir das predições individuais para cada cadeia. Podemos atribuir este efeito ao fato de estarmos conjugando conjuntos de dados cuja informação não é redundante: um deles diz respeito à localização espacial dos átomos na cadeia, e o outro diz respeito à orientação dos átomos ao longo da cadeia. Apresentamos os sinais otimizados para a cadeia 1ifr-A e uma comparação com o sinal não otimizado decorrente da melhor predição anterior por camadas (4 camadas) na Figura 3.8. Apresentamos também uma comparação entre os sinais para as outras proteínas na Tabela 3.7.

Ressaltamos que mesmo nas situações em que o sinal otimizado é menos correla-

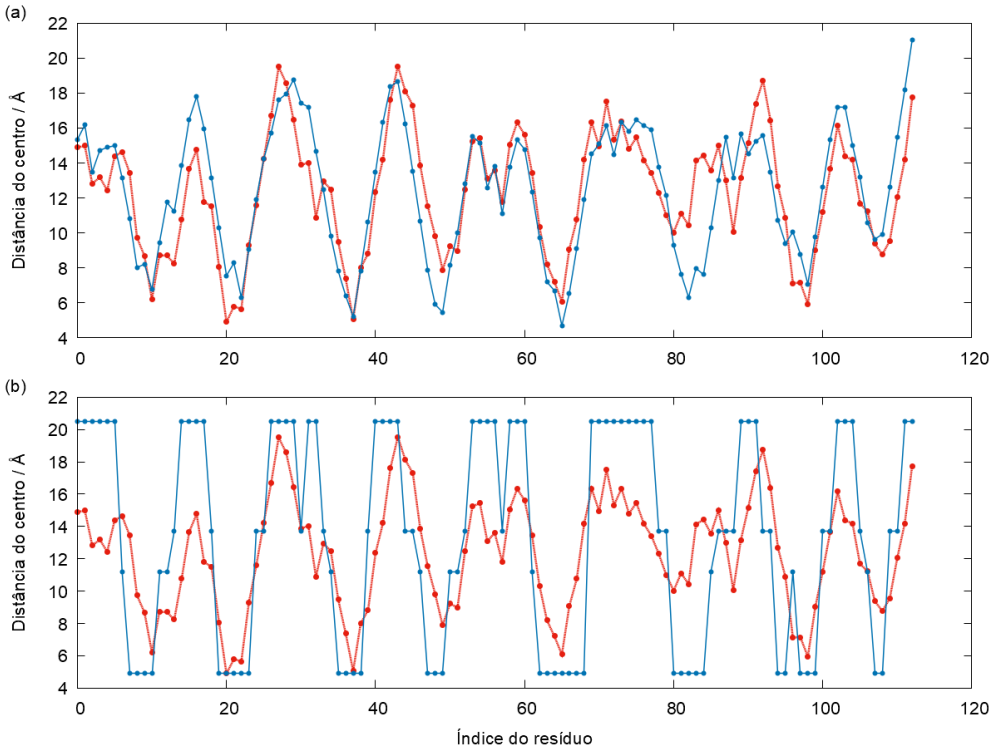

Figura 3.8. Proteína 1ifr-A. (a) Sinal otimizado com os potenciais provenientes da predição diferencial (2 direções), predição com 2 camadas e direção da cadeia lateral e o potencial da distribuição global dos enterramentos atômicos. (b) Sinal obtido a partir da predição direta de 4 camadas. As linhas vermelhas são referentes ao sinal real de enterramento e as linhas azuis são referentes às predições.

Tabela 3.7. RMSD e correlação dos enterramentos reais com os sinais otimizados a partir das simulações Monte Carlo combinando a predição pela representação diferencial de 2 direções, a predição por 2 camadas com direção da cadeia lateral e o termo global da distribuição dos enterramentos atômicos, para as diferentes proteínas avaliadas e uma comparação com os resultados obtidos a partir da predição direta de 4 camadas.

Código	Correlação linear		RMSD / Å	
	4 camadas	Sinal otimizado	4 camadas	Sinal otimizado
1c9o-A	0.74	0.66	3.81	2.87
1ifr-A	0.71	0.81	4.83	2.20
1oz9-A	0.73	0.79	4.64	2.75
2i9h-A	0.54	0.61	4.73	2.96
3fil-A	0.40	0.59	4.40	2.69
3tim-A	0.58	0.47	6.50	4.66

cionado com os enterramentos reais do que a predição por camadas, como é o caso das cadeias 1c9o-A e 3tim-A, obtemos um sinal de melhor resolução como evidenciado pelo menor RMSD desse sinal em todos os casos. Na cadeia 3tim-A, inclusive, o sinal otimzado apresentou um erro crítico entre os resíduos 60 a 90 (Figura 3.9). A predição por camadas

não comete este erro mas possui uma flutuação considerável nesta região, o que resultou na otimização ineficiente desta parte da cadeia. Contudo, ainda assim obtivemos uma melhora notável no RMSD desta predição em relação à predição anterior.

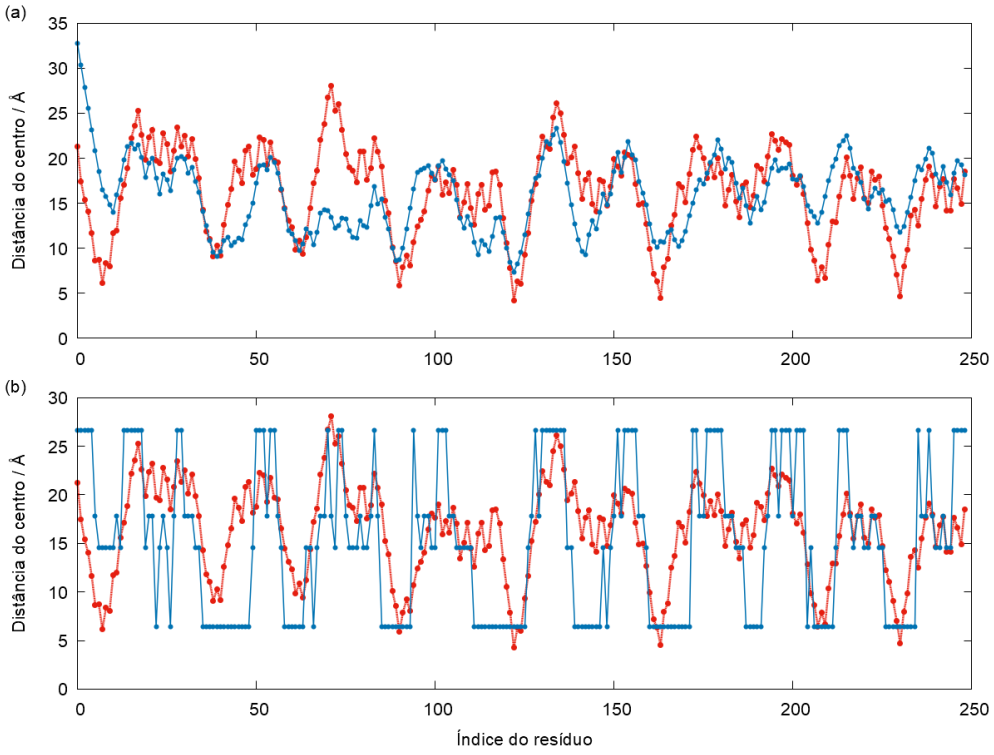

Figura 3.9. Proteína 3tim-A. (a) Sinal otimizado com os três potenciais desenvolvidos. Observamos que houve um erro neste sinal na região entre os resíduos 60 e 90, o que não ocorre no sinal obtido a partir da predição direta de 4 camadas (b). As linhas vermelhas são referentes ao sinal real de enterramento e as linhas azuis são referentes às predições.

Desta forma, o método aqui proposto se apresenta como uma técnica mais robusta e mais tolerante a falhas, capaz de explorar um espaço contínuo de distâncias, em contraste com uma quantidade fixa de camadas discretas, o que nos dá uma perspectiva de maior resolução no sinal das predições de enterramento também para outras cadeias aqui não abordadas. Além disto, vimos anteriormente que um número pequeno de camadas é insuficiente para descrevermos a estrutura tridimensional de proteínas grandes, implicando que a estrutura terciária de cadeias longas necessita de alfabetos de representação cada vez maiores (muitas camadas) para obtermos predições de camadas adequadas. O uso da representação diferencial, por sua vez, aliada aos métodos de otimização propostos nesta seção devem nos habilitar a desenvolver potenciais de enovelamento com resolução suficiente para a descrição da estrutura de proteínas globulares independentemente de seu tamanho. Por fim, ressaltamos que embora estas predições não tenham sido validadas

por simulações de dinâmica molecular, a obtenção de sinais de enterramento predito mais próximos dos sinais observados nas cadeias reais nos dá a perspectiva de predições de com maior qualidade para o uso em futuras simulações.

Capítulo 4

Conclusões e perspectivas

Neste trabalho estabelecemos algumas medidas que nos permitem estimar a quantidade de informação estrutural, em termos de enterramentos atômicos, que precisa ser fornecida para que uma estrutura acesse o estado nativo nas nossas simulações de dinâmica molecular. Obtivemos das análises realizadas no artigo que o mínimo de camadas necessário é de 3 a 4 camadas para descrevermos as cadeias pequenas (com até 66 resíduos) e 4 a 5 camadas para descrevermos as cadeias médias (113 e 104 resíduos). A proteína de 141 resíduos se mostrou estável com 6 camadas de enterramento e, embora tenha permanecido inacessível até com as simulações de 8 camadas, seu perfil energético destaca que esta pode ter sido uma questão do tempo total das simulações. Em todo caso, os resultados evidenciam a correlação entre o tamanho da cadeia com o número crescente de camadas necessário para a descrição de sua estrutura. Isto se traduz no fato de que há uma resolução mínima necessária, que pode ser independente do tamanho, para que os enterramentos caracterizem corretamente a estrutura. Em outras palavras, como o tamanho das camadas aumenta com o tamanho da sequência, precisamos de mais camadas para mantermos a resolução requerida em estruturas maiores.

Este princípio nos habilitou então a quantificar o erro tolerado pelas estruturas quando fornecemos informação incompleta ou errada acerca dos enterramentos atômicos. Caracterizamos, também em termos de estabilidade e de acessibilidade, o comportamento das estruturas ao serem simuladas com frações crescentes de erro nos potenciais de enovelamento fornecidos, expressas na forma de potenciais removidos ou potenciais inconsistentes com o enterramento nativo para um número crescente de átomos na cadeia. Estimamos que, para as proteínas pequenas analisadas, entre 75% a 80% da informação dos enterramentos

dos átomos, quando fornecida na forma de 4 camadas equiprováveis, é composta por informação redundante e que 20% a 25% dos enterramentos atômicos (cerca de 1 enterramento por resíduo) já é suficiente para que as estruturas nativas destas proteínas sejam acessíveis e mantenham-se estáveis. Em outras palavras, assumindo que toda a informação estrutural de uma proteína está contida em sua sequência primária, então para ser capaz de distinguir entre estruturas diferentes, o modelo efetivamente precisa de somente entre 20% a 25% da informação estrutural dependente da sequência primária, neste caso, fornecida na forma de enterramentos atômicos. O restante da informação estrutural dependente da sequência nas cadeias analisadas pode ser apagado, indicando que esta informação já está presente no sistema, através das restrições físicas independentes da sequência existentes no modelo, como volume de exclusão, ângulos diedrais, ligações covalentes e ligações de hidrogênio.

Nesta primeira parte do trabalho quantificamos, de forma geral, uma medida tanto da resolução necessária para uma representação de enterramentos atômicos como descritores estruturais quanto da tolerância do modelo quanto a erros na representação. Estas medidas nos permitem também avaliar de forma mais objetiva a adequação de predições estruturais de enterramento atômico a partir da sequência para o uso em simulações de enovelamento, uma vez que já estabelecemos um limiar para a quantidade de erros na classificação dos átomos provenientes destas predições.

No capítulo subsequente discutimos uma nova metodologia de representação dos enterramentos atômicos, não pela distância absoluta de cada átomo ao centro geométrico da estrutura, mas sim pelas distâncias relativas entre pares de átomos adjacentes ao longo da cadeia. Exploramos algumas maneiras de se codificar o sinal de enterramento atômico através desta representação, mostrando que podemos trivialmente decodificar esse sinal de tal forma que ele descreva adequadamente os enterramentos atômicos. A grande vantagem da representação diferencial dos enterramentos sobre a representação por camadas está em sua escalabilidade, uma vez que é possível descrevermos com os enterramentos de estruturas de qualquer tamanho utilizando alfabetos de apenas 2 ou 3 símbolos na codificação. Embora, no caso das predições utilizamos alfabetos com 4 e 9 símbolos a fim de mantermos a simetria da representação, os símbolos destes alfabetos são altamente correlacionados nas sequências produzidas: condicionalmente ao símbolo de um átomo na sequencia, o símbolo do átomo seguinte está restrito a apenas 2 ou 3 símbolos, respectivamente, resultando em uma sequência onde o alfabeto efetivo (dado

pela densidade de entropia por símbolo) da representação é de 2 ou 3 símbolos.

As predições realizadas a partir desta representação apresentaram, de forma geral, uma qualidade comparável às predições com camadas, mas a reconstituição do sinal de enterramento a partir das predições pode não ser tão direta, dado que uma variação pontual na direção da cadeia altera a forma geral do sinal em toda a sequencia. Propusemos então uma metodologia de otimização do sinal a partir das predições, que também permite combinarmos resultados de predições obtidas por representações diversas. Obtivemos assim, através da otimização das predições de direções combinadas um sinal cuja correlação com os enterramentos é superior à correlação obtida a partir das melhores predições diretas de camadas para 4 das 6 estruturas testadas. Os traços de enterramento atômico obtidos a partir desta otimização, além disso, nos permitem expressar a sequência de enterramentos dentro de um espaço contínuo de distâncias ao centro das cadeias, mesmo que as predições tenham sido feitas com alfabetos de apenas 2 símbolos. Isto confere maior resolução aos sinais preditos, tornando viável a predição de enterramentos para fins de simulação de enovelamento também em proteínas grandes, uma vez que não estamos mais limitados pelo número de camadas da predição. Vale ressaltar também que o uso da representação diferencial, por construção, confere às predições, além da informação dos enterramentos, parte da informação independente da sequência na forma das distâncias fixas entre os átomos denotando sua conectividade. Esta informação muitas vezes é perdida nas predições de camadas, o que resulta em resíduos adjacentes na cadeia serem preditos em camadas não consecutivas.

Assim, colocamos como perspectivas de aplicação destas predições a derivação de potenciais que possam ser utilizados propriamente em simulações de enovelamento proteico, da mesma forma como já foi realizado com predições puras de camadas em estudos anteriores. É interessante notar também que parte das limitações das predições de diferenças de enterramento se devem ao fato de que esta representação resulta na propagação de erros pontuais ao longo da cadeia predita. O HMM é um algoritmo capaz de capturar apenas correlações locais dentro de uma sequência mas é possível que existam correlações de longa distância em proteínas globulares que não são detectadas por um HMM e que possam contribuir para a correção de parte desses erros. Deste modo o uso da representação diferencial dos enterramentos atômicos pode ter também aplicações ainda mais impactantes em predições que sejam capazes de explorar essas possíveis correlações

de longa distância através de tecnologias mais sofisticadas de processamento de dados e análise de sequências.

Referências

1 CAMPBELL, Mary K; FARRELL, Shawn O. **Biochemistry**. 5. ed.: Thomson Learning, 2007. v. 1.

2 ALBERTS, Bruce et al. **Molecular Biology of the Cell**. 5. ed.: Garland Science, Taylor & Francis Group, 2008.

3 WATSON, J. D.; CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. **Nature**, v. 171, n. 4356, p. 737–738, abr. 1953.

4 NIRENBERG, M et al. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. **Proceedings of the National Academy of Sciences of the United States of America**, v. 53, n. 5, p. 1161–8, mai. 1965.

5 KENDREW, J. C. et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. **Nature**, Nature Publishing Group, v. 181, n. 4610, p. 662–666, mar. 1958.

6 PERUTZ, M. F. et al. Structure of Hæmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by X-ray analysis. **Nature**, v. 185, n. 4711, p. 416–422, 1960.

7 KENDREW, J. C. et al. Structure of myoglobin: A three-dimensional fourier synthesis at 2 . resolution. **Nature**, v. 185, n. 4711, p. 422–427, 1960.

8 PAULING, L.; COREY, R. B.; BRANSON, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. **Proceedings of the National Academy of Sciences**, v. 37, n. 4, p. 205–211, 1951.

9 DILL, Ken A. et al. The Protein Folding Problem. **Annual Review of Biophysics**, v. 37, n. 1, p. 289–316, jun. 2008.

10 DILL, Ken A.; MACCALLUM, Justin L. The Protein-Folding Problem, 50 Years On. **Science**, v. 338, n. 6110, p. 1042–1046, nov. 2012.

11 JANKOVIĆ, Brankica G; POLOVIĆ, Natalija D J. The protein folding problem. **Biologia Serbica**, v. 39, n. 1, p. 105–111, 2017.

12 ANFINSEN, C. B. et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proceedings of the National Academy of Sciences**, v. 47, n. 9, p. 1309–1314, set. 1961.

13 FINKELSTEIN, A V. 50+ Years of Protein Folding. **Biochemistry (Moscow)**, v. 83, S1, s3–s18, jan. 2018.

14 ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, American Association for the Advancement of Science, v. 181, n. 4096, p. 223–230, jul. 1973.

15 LEVINTHAL, Cyrus. How to Fold Graciously. In: _____. **Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois**. University of Illinois Press, 1969. p. 22–24.

16 DILL, Ken A. Theory for the Folding and Stability of Globular Proteins. **Biochemistry**, v. 24, n. 6, p. 1501–1509, 1985.

17 ZWANZIG, R.; SZABO, A.; BAGCHI, B. Levinthal's paradox. **Proceedings of the National Academy of Sciences**, v. 89, n. 1, p. 20–22, 1992.

18 DILL, Ken A.; CHAN, Hue Sun. From levinthal to pathways to funnels. **Nature Structural Biology**, v. 4, n. 1, p. 10–19, 1997.

19 LEOPOLD, P. E.; MONTAL, M.; ONUCHIC, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. **Proceedings of the National Academy of Sciences**, v. 89, n. 18, p. 8721–8725, 1992.

20 ONUCHIC, José Nelson; LUTHEY-SCHULTEN, Zaida; WOLYNES, Peter G. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. **Annual Review of Physical Chemistry**, v. 48, n. 1, p. 545–600, 1997.

21 LEVITT, Michael; WARSHEL, Arieh. Computer simulation of protein folding. **Nature**, Nature Publishing Group, v. 253, n. 5494, p. 694–698, fev. 1975.

22 TAKETOMI, Hiroshi; UEDA, Yuzo; GŌ, Nobuhiro. Studies on protein folding, unfolding and fluctuations by computer simulation. **International Journal of Peptide and Protein Research**, v. 7, n. 6, p. 445–459, jan. 1975.

23 GŌ, Nobuhiro; TAKETOMI, Hiroshi. Respective roles of short- and long-range interactions in protein folding. **Proceedings of the National Academy of Sciences**, v. 75, n. 2, p. 559–563, fev. 1978.

24 DILL, Ken A. et al. Principles of protein folding — A perspective from simple exact models. **Protein Science**, v. 4, n. 4, p. 561–602, 1995.

25 SHAKHNOVICH, Eugene; GUTIN, Alexander. Enumeration of all compact conformations of copolymers with random sequence of links. **The Journal of Chemical Physics**, v. 93, n. 8, p. 5967–5971, out. 1990.

26 ŠALI, Andrej; SHAKHNOVICH, Eugene; KARPLUS, Martin. How does a protein fold? **Nature**, v. 369, n. 6477, p. 248–251, mai. 1994.

27 BRYNGELSON, Joseph D. et al. Funnels, pathways, and the energy landscape of protein folding: A synthesis. **Proteins: Structure, Function, and Genetics**, v. 21, n. 3, p. 167–195, mar. 1995.

28 SHAKHNOVICH, Eugene. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. **Chemical reviews**, v. 106, n. 5, p. 1559–88, mai. 2006.

29 HAO, Ming Hong; SCHERAGA, Harold A. Molecular mechanisms for cooperative folding of proteins. **Journal of Molecular Biology**, v. 277, n. 4, p. 973–983, 1998.

30 KAUZMANN, W. Some Factors in the Interpretation of Protein Denaturation. In: ADVANCES in Protein Chemistry. [S.n.], 1959. v. 14. p. 1–63.

31 DILL, Ken A. Dominant forces in protein folding. **Biochemistry**, v. 29, n. 31, p. 7133–7155, ago. 1990.

32 LAU, Kit Fun; DILL, Ken A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. **Macromolecules**, v. 22, n. 10, p. 3986–3997, out. 1989.

33 PEREIRA DE ARAUJO, A. F. Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation. **Proceedings of the National Academy of Sciences**, v. 96, n. 22, p. 12482–12487, out. 1999.

34 NISHIMURA, Chiaki et al. Sequence Determinants of a Protein Folding Pathway. **Journal of Molecular Biology**, v. 351, n. 2, p. 383–392, ago. 2005.

35 DILL, Ken A. Polymer principles and protein folding. **Protein Science**, v. 8, n. 6, p. 1166–1180, 1999.

36 LINDEN, Marx Gomes van der et al. Ab initio protein folding simulations using atomic burials as informational intermediates between sequence and structure. **Proteins: Structure, Function, and Bioinformatics**, v. 82, n. 7, p. 1186–1199, jul. 2014.

37 GOMES, Antonio L C et al. Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. **Proteins: Structure, Function, and Bioinformatics**, v. 66, n. 2, p. 304–320, nov. 2006.

38 REIF, F. **Fundamentals of Statistical and Thermal Physics**. Waveland Press, 2009.

39 PEREIRA DE ARAÚJO, Antônio F. et al. 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**, v. 70, n. 3, p. 971–983, set. 2007.

40 PEREIRA DE ARAUJO, Antonio F.; ONUCHIC, José Nelson. 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**, National Academy of Sciences, v. 106, n. 45, p. 19001–19004, nov. 2009.

41 SHANNON, Claude Elwood. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3,4, p. 379–423, 623–656, 1948.

42 ROCHA, Juliana Ribeiro et al. Information-theoretic analysis and prediction of protein atomic burials: on the search for an informational intermediate between sequence and structure. **Bioinformatics**, v. 28, n. 21, p. 2755–2762, nov. 2012.

43 GRIEP, Sven; HOBOHM, Uwe. PDBselect 1992–2009 and PDBfilter-select. **Nucleic Acids Research**, v. 38, suppl_1, p. d318–d319, jan. 2010.

44 COVER, Thomas M.; THOMAS, Joy A. **Elements of Information Theory**. Hoboken, NJ, USA: John Wiley & Sons, Inc., set. 2005.

45 CROOKS, Gavin E; BRENNER, Steven E. Protein secondary structure: entropy, correlations and prediction. **Bioinformatics**, v. 20, n. 10, p. 1603–1611, 2004.

46 DURBIN, Richard et al. **Biological Sequence Analysis: probabilistic models of proteins and nucleic acids**. Cambridge University Press, 2002.

47 FERREIRA, Diogo César et al. Information and redundancy in the burial folding code of globular proteins within a wide range of shapes and sizes. **Proteins: Structure, Function, and Bioinformatics**, v. 84, n. 4, p. 515–531, abr. 2016.

48 WHITFORD, Paul C et al. Conformational transitions of adenylate kinase: switching by cracking. **Journal of molecular biology**, v. 366, n. 5, p. 1661–71, mar. 2007.

49 LINDEN, Marx Gomes van der. **Simulação do enovelamento de proteínas com potenciais de enterramentos atômicos dependentes da sequência**. 2013. Tese de Doutorado – Universidade de brasília.

50 YAMAGUCHI, Seiji et al. Preparation of bioactive Ti-15Zr-4Nb-4Ta alloy from HCl and heat treatments after an NaOH treatment. **Journal of biomedical materials research. Part A**, v. 97, n. 2, p. 135–44, mai. 2011.

51 CUTLER, Cassius Chapin. **Differential Quantization of Communication Signals**. [S.n.], 1952.

52 MACKAY, D. J. C. **Information Theory, Inference and Learning Algorithms**. Cambridge University Press, 2003.

53 BERMAN, Helen M et al. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 2000.

54 ALTSCHUL, Stephen F et al. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403–410, out. 1990.

55 MODDEMEIJER, R. The Distribution of Entropy Estimators based on Maximum Mean Log-likelihood. In: BIEMOND, J (Ed.). **21st Symp. on Information Theory in the Benelux**. Wassenaar (NL): Werkgemeenschap Informatie- en Communicatietheorie, Enschede (NL), 2000. p. 231–238.

56 EFRON, B; TIBSHIRANI, R J. **An Introduction to the Bootstrap**. Taylor & Francis, 1994. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).

57 METROPOLIS, Nicholas et al. Equation of State Calculations by Fast Computing Machines. **The Journal of Chemical Physics**, v. 21, n. 6, p. 1087–1092, 1953.

58 KULLBACK, S; LEIBLER, R A. On Information and Sufficiency. **The Annals of Mathematical Statistics**, The Institute of Mathematical Statistics, v. 22, n. 1, p. 79–86, mar. 1951.

Anexos

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 $\mathrm{H} = \{\mathrm{A, C, F, G, I, L, M, V, W, Y}\}$ and $\mathrm{P} = \{\mathrm{D, E, H, K, N, P, Q, R, S, T}\}$, respectively. In HPN a third, 'neutral', class includes residues from both HP groups, $\mathrm{N} = \{\mathrm{A, G, H, S, T}\}$. 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:

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

with limiting behavior

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

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.

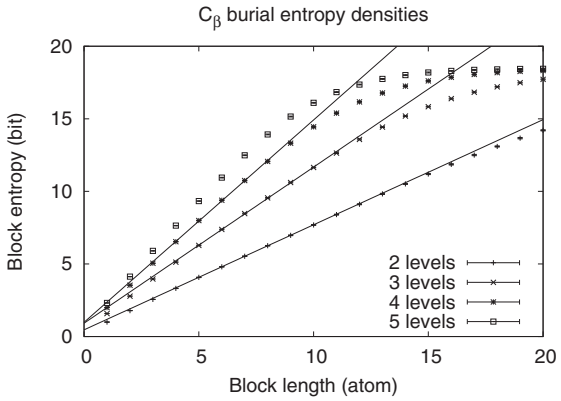

Fig. 1. N-block sequence entropy estimates as a function of block size N for different alphabets of C_β burial levels. Both the entropy density (inclination) and excess entropy (intersect with the ordinates) are obtained from straight lines fitted to the linear region, which is clearly identified for $L = 2$ 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

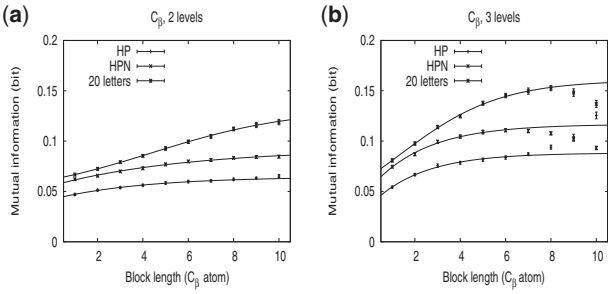

Fig. 2. Estimates for the mutual information, $I(Q_0; B^N)$, between a single central amino acid identity, Q_0, and N-blocks of burials, B^N, as a function of block size N, for two (**a**) and three (**b**) levels of C_β burials. Different sets of points correspond to different alphabets of amino acid 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

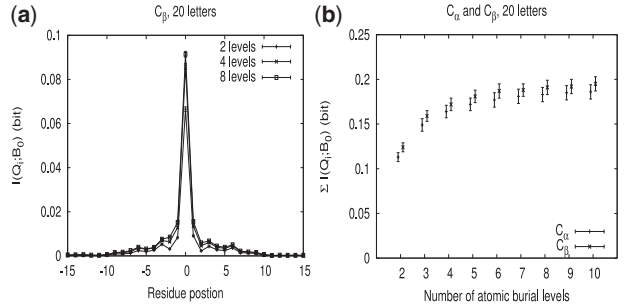

Fig. 3. Positional mutual information $I(Q_i; B_0)$ between amino acid identity at position i, Q_i, within the N-block of identities Q^N, and central C_β burial, B_0, for 20 amino acid letters and various numbers of burial levels (**a**) and limiting behavior for the sum of positional mutual information 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**)

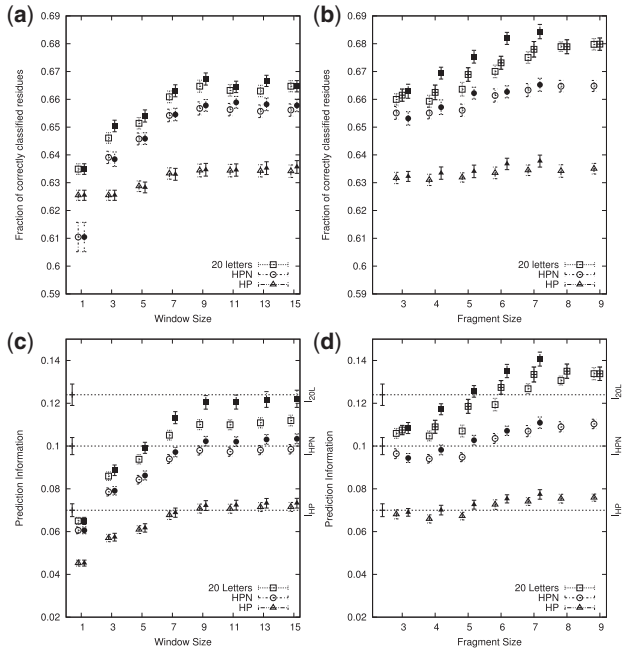

Fig. 4. Prediction accuracy A (**a**,**b**) and prediction information I_p (**c**,**d**) for two levels of C_β burials with different identity alphabets. Plots in the first column (a,c) show results for NBC predictions; the second column (b,d) refers to the HMM results. The NBC method is bounded, within 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

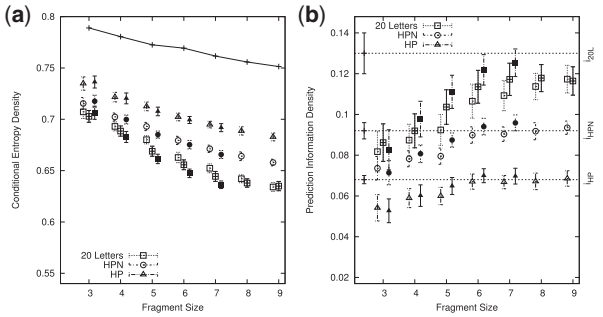

Fig. 5. For HMM results, the *density* of prediction information, i_p, can be calculated as the difference between an N-dependent estimate for the entropy density of burial levels, $h_N(B)$, Equation (S13) of the Supplementary Information (shown as a solid line in **a**), and an analogous estimate for 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

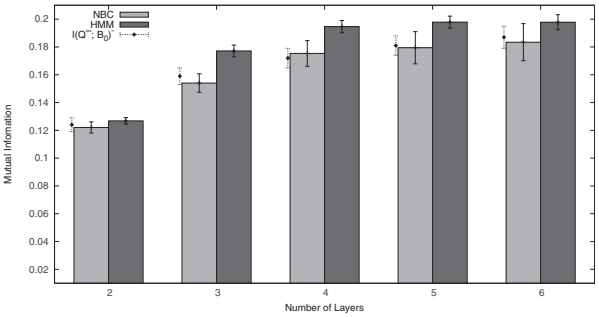

Fig. 6. As the number of discrete burial layers increases, the quality of prediction, as measured by the prediction information, I_p, also improves, at least up to four to five layers. The results are shown for NBC2 and HMM C_β predictions. Analogous results for C_α are shown in the Supplementary Information. 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

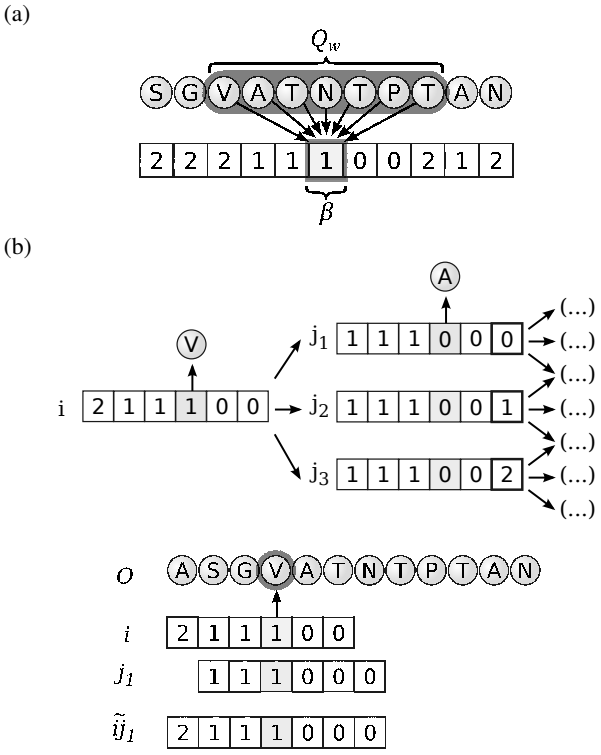

Fig. 1. (a) In the Naive Bayesian Classifier algorithm (NBC), $p(\beta|Q_w)$ defines the probability of having the burial β associated to the central residue in a window of residues Q_w. (b) In the Hidden Markov Model (HMM), each state corresponds to a fragment of $f-1$ hidden variables and has only 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

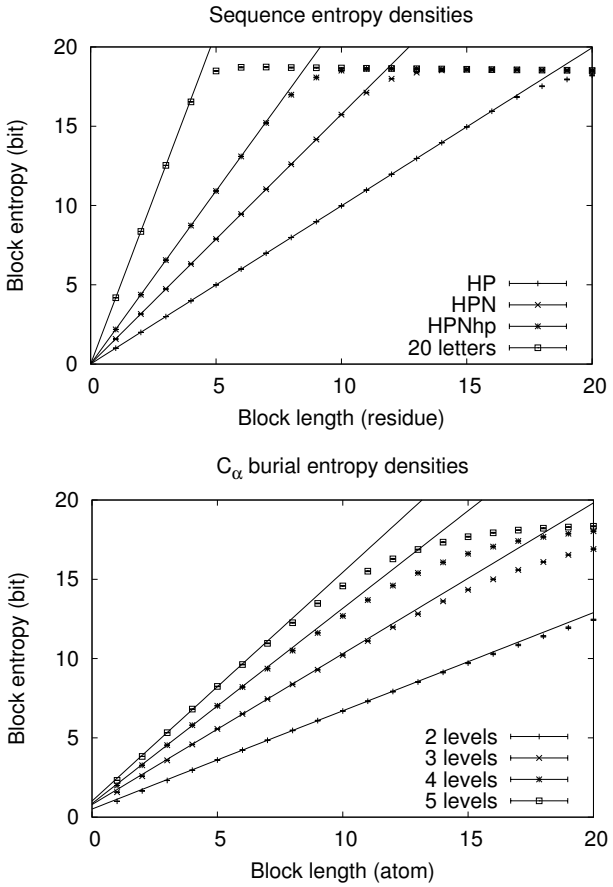

Fig. 2. Entropy for identities and C_α burials. N-block sequence entropy estimates as a function of block size N for different alphabets of C_α burial levels. Straight lines represent linear fits to the data from which the entropy density (inclination) and excess entropy (intersect with the ordinates)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

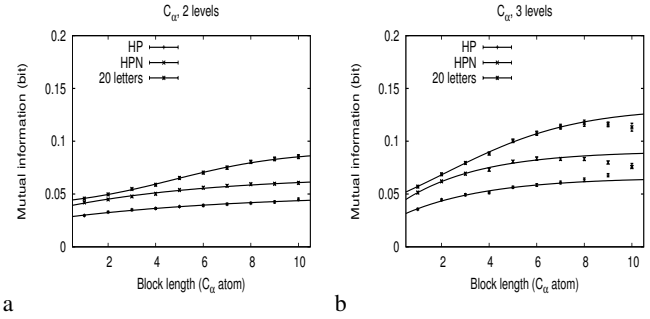

Fig. 3. Mutual information for C_α burials. Estimates for the mutual information, $I(Q_0; B^N)$, between a single central amino acid identity, Q_0, and N-blocks of burials, B^N, as a function of block size N, for 2 (a) and 3 (b) levels of C_α burials. Different sets of points correspond to different 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).

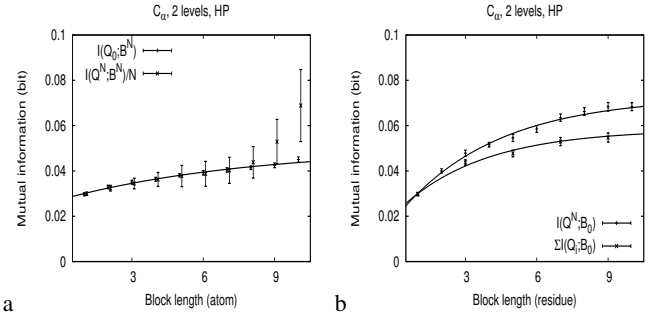

Fig. 4. Approximations in mutual information estimates. Comparison between $I(Q_0; B^N)$ and $I(Q^N; N^N)/N$, for the HP alphabet and 2 levels of C_α burials, reveals virtually coincident values before saturation (a), suggesting that Eq. 3 of the main article is a good approximation. Comparison between $I(Q^N; B_0)$ 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.

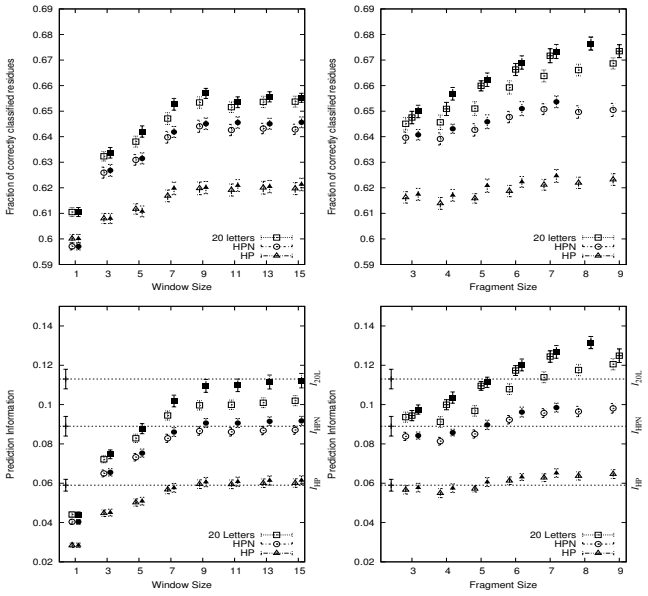

Fig. 5. Accuracy and prediction information for C_α burials. Prediction accuracy A (a,b) and prediction information I_p (c,d) for two levels of C_α burials with different identity alphabets. Plots in the first column (a,c) show results for NBC predictions; the second column (b,d) refers to the HMM results. 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.

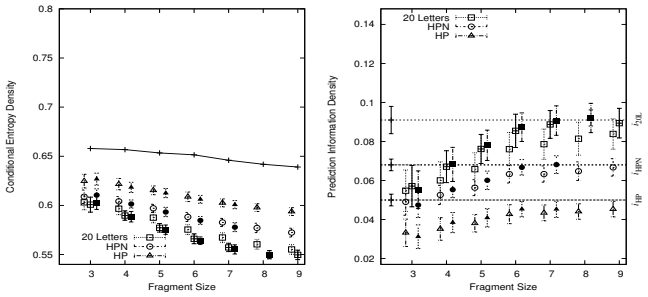

Fig. 6. Prediction information density for C_α burials. For HMM results, the *density* of prediction information, i_p, can be calculated as the difference between an N-dependent estimate for the entropy density of burial levels, $h_N(B)$, Eq. 13, and an analogous estimate for the entropy density conditional 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.

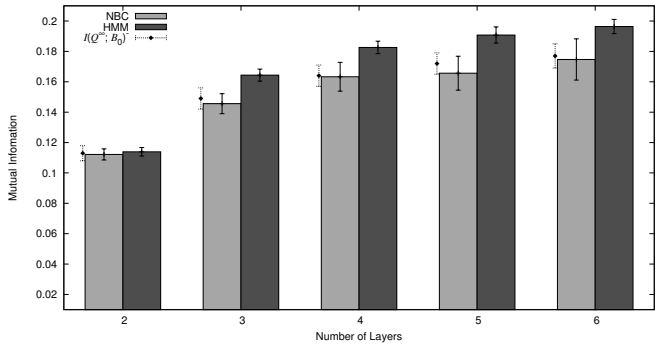

Fig. 7. Dependence of prediction information on the number of burial levels for C_α atoms. As the number of discrete burial layers increases, the quality of prediction, as measured by the prediction information, I_p, also improves, at least up to 4-5 layers. Window size of 15 and fragment size of 7 were 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 Biologia Teórica e Computacional, Universidade de Brasília, Brasília-DF 70910-900, Brazil

[2] Departamento de Física, Instituto de Biociências, Letras e Ciências Exatas, UNESP - Univ Estadual Paulista, São José do Rio Preto-SP, 15054-000, Brazil

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

ABSTRACT

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 2014; 82:1186–1199.

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 e Computacional, 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 26 November 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

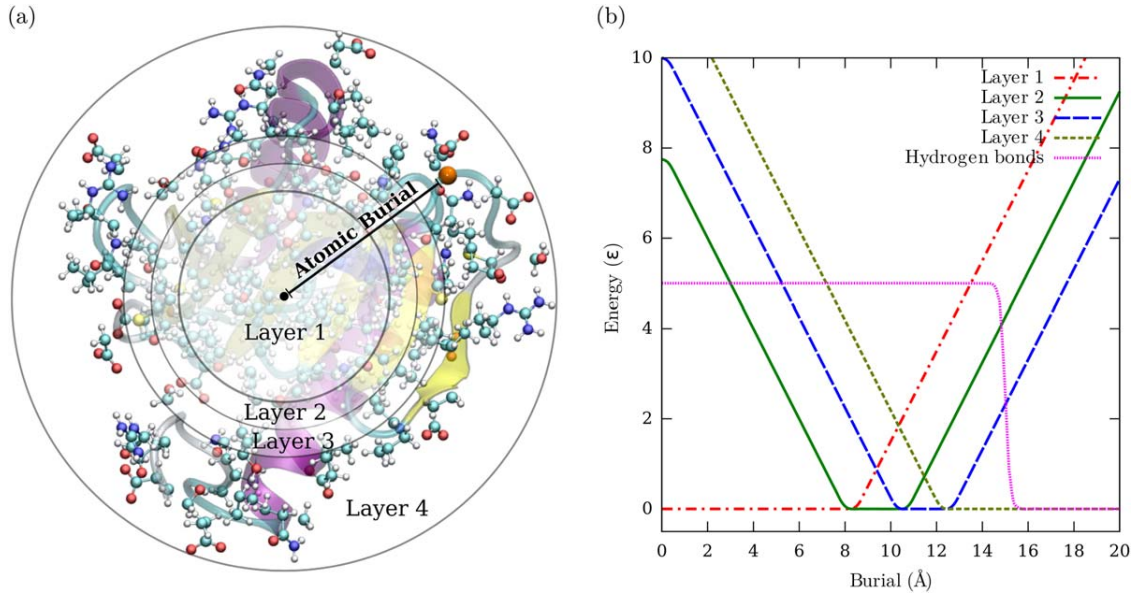

Figure 1

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.[16] (**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:

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

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:

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

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:

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

and

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

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 was 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

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

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,

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

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

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 $121<N_r\leq 160$, is 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

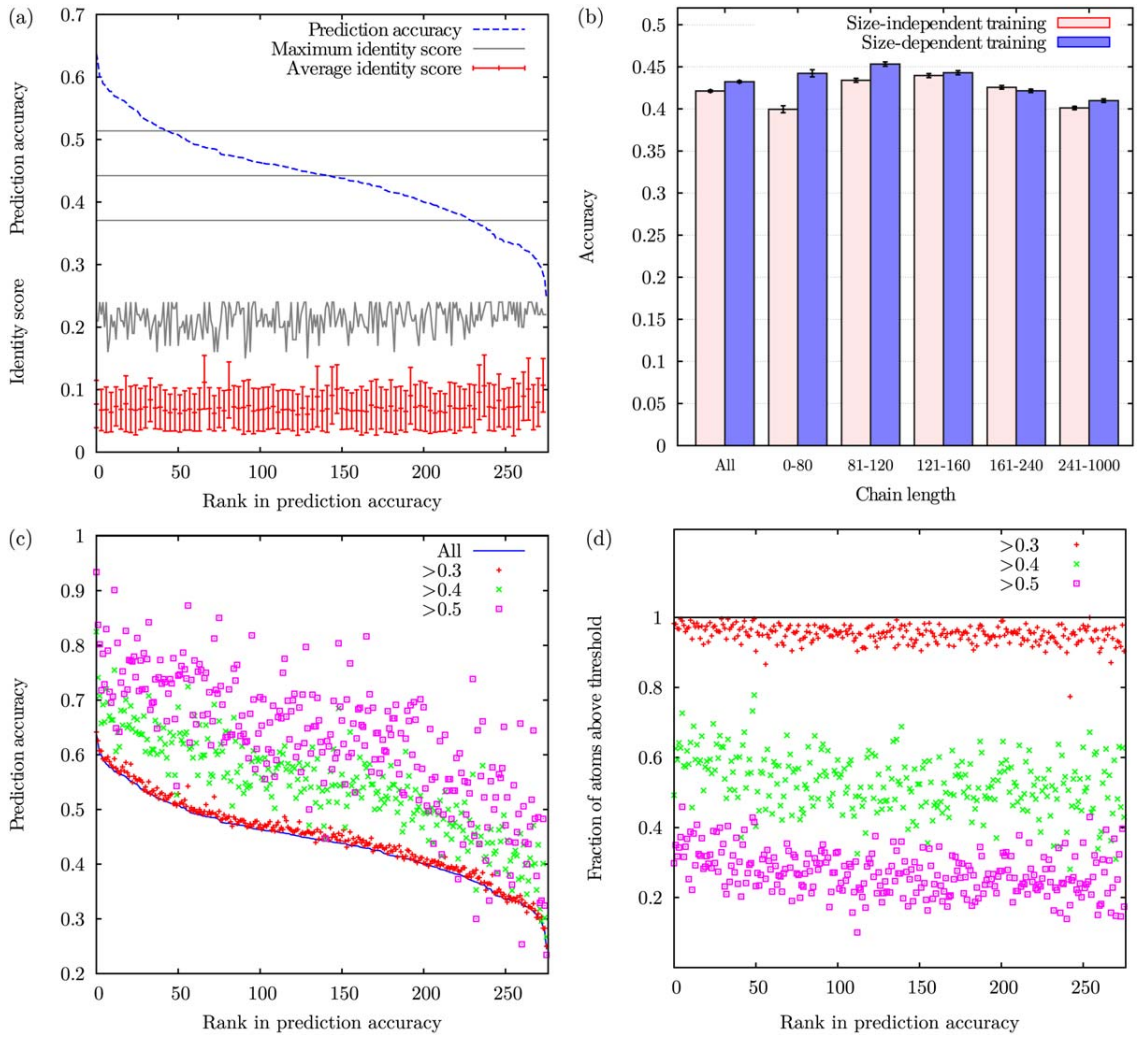

Figure 2

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,

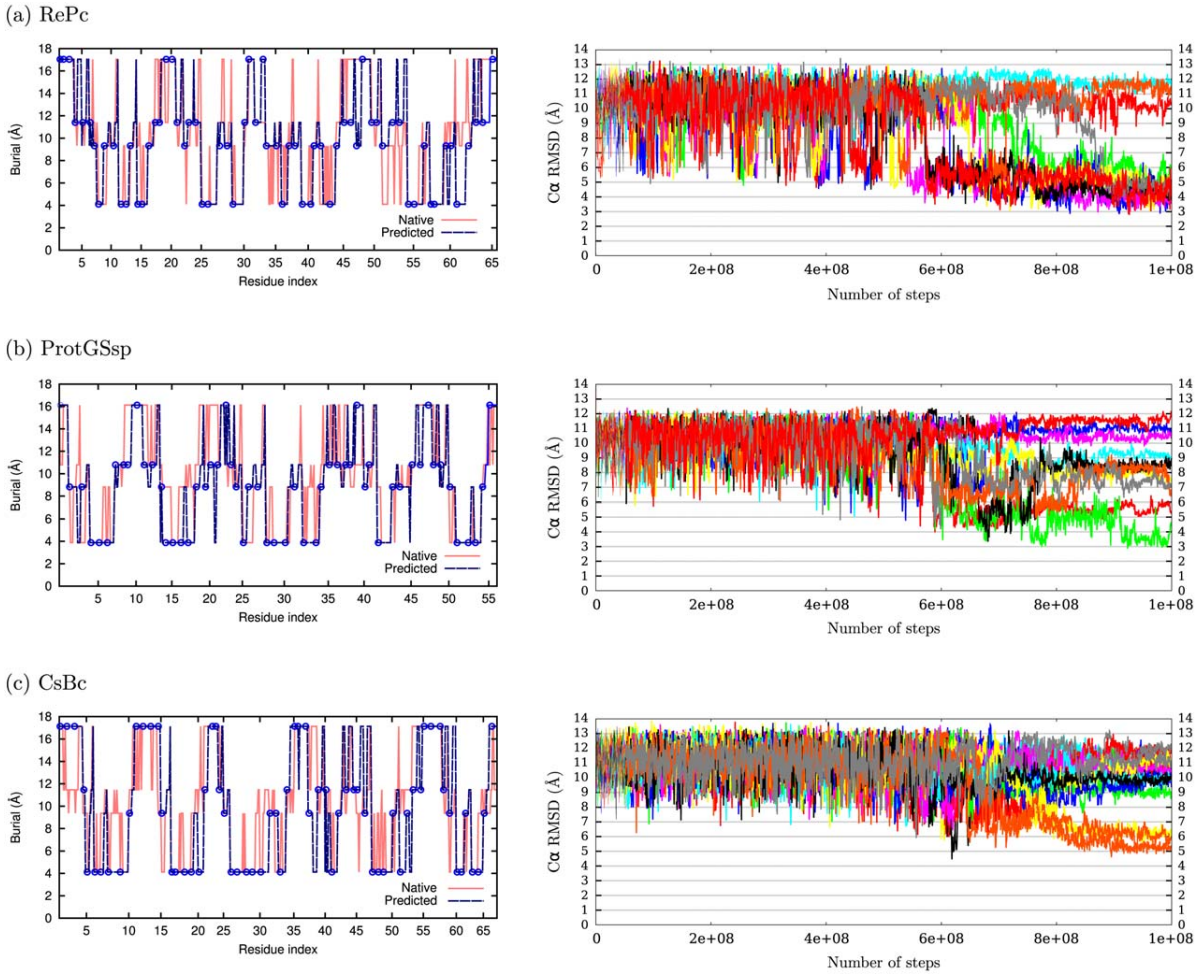

Figure 3

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 to the left. 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 on the right side of Figure 3, 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

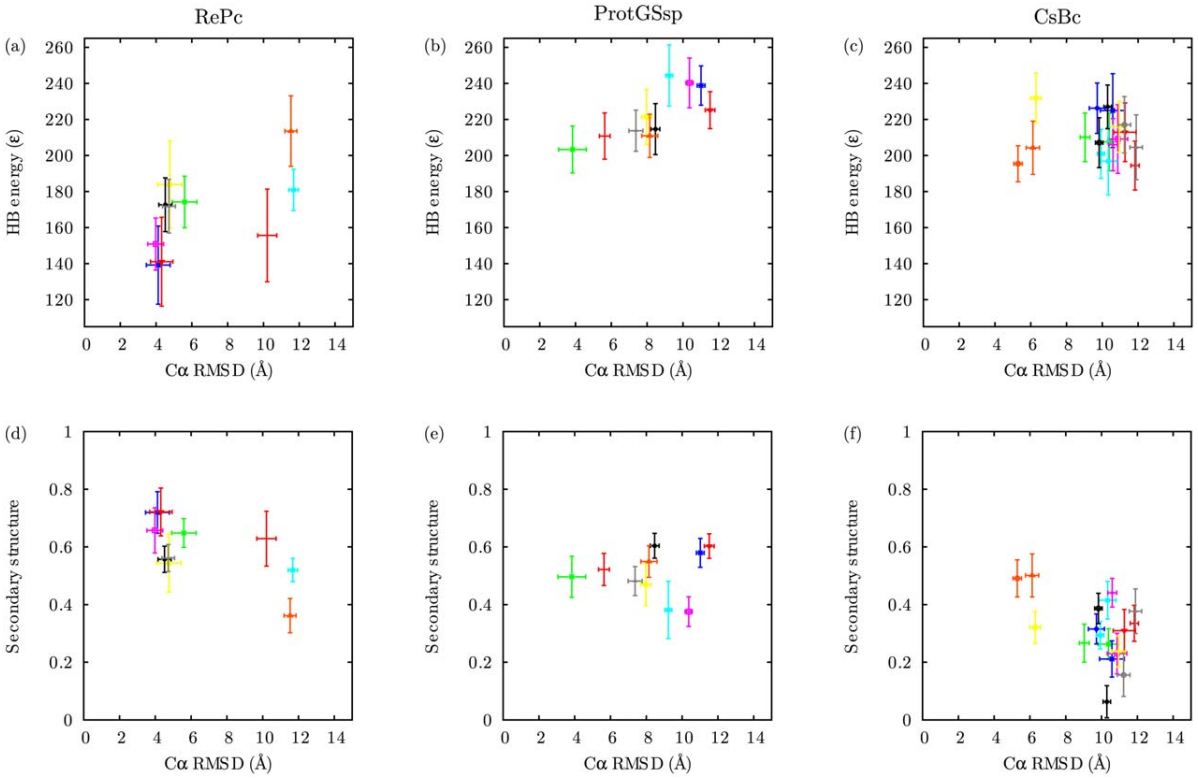

Figure 4

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.]

structure, 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

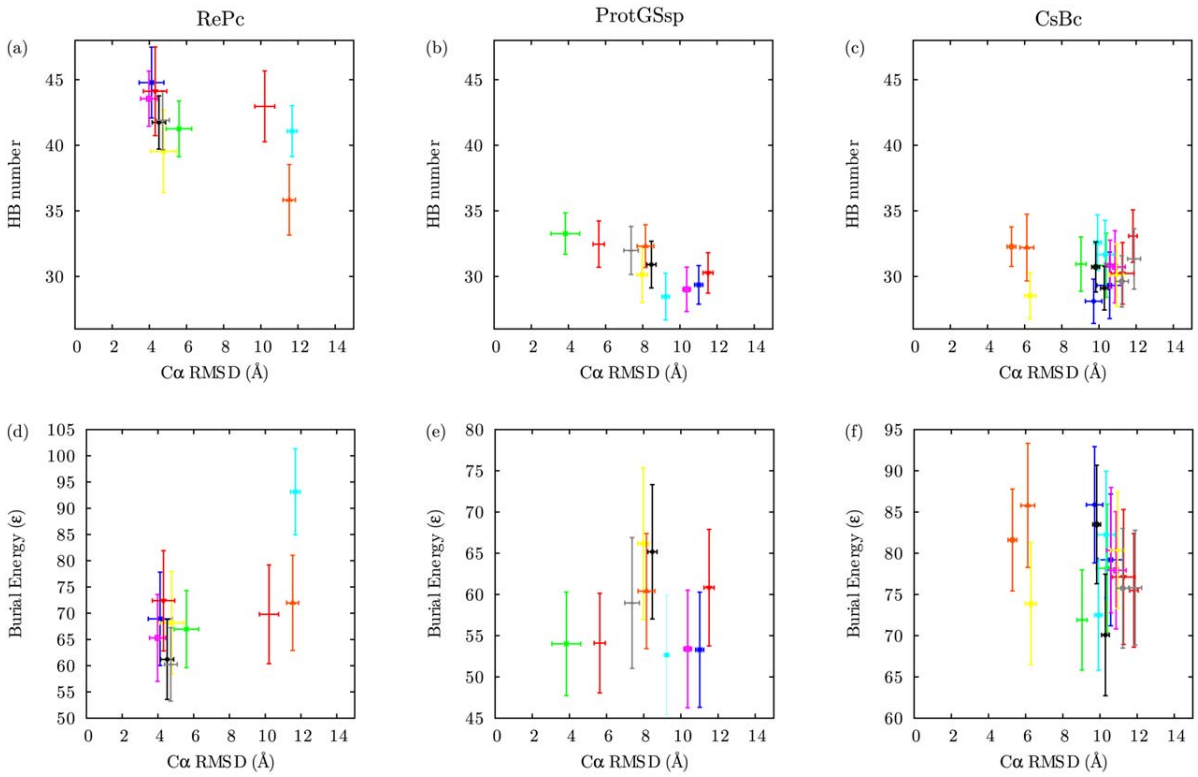

Figure 5

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

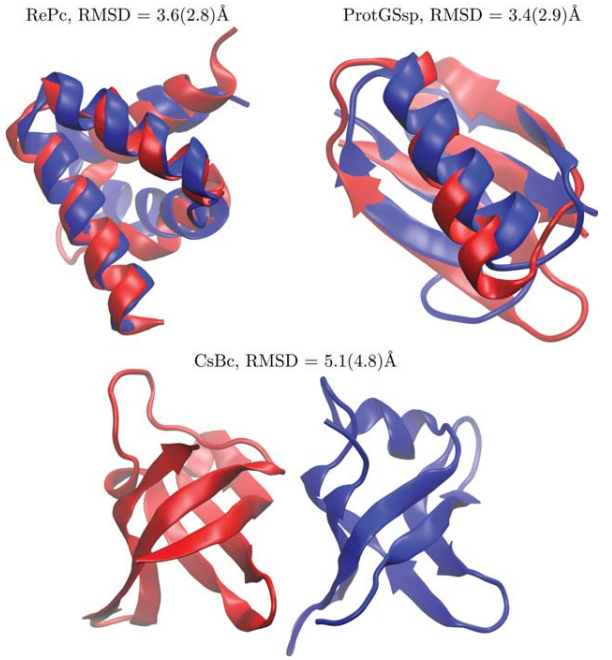

Figure 6

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 have 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.

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 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

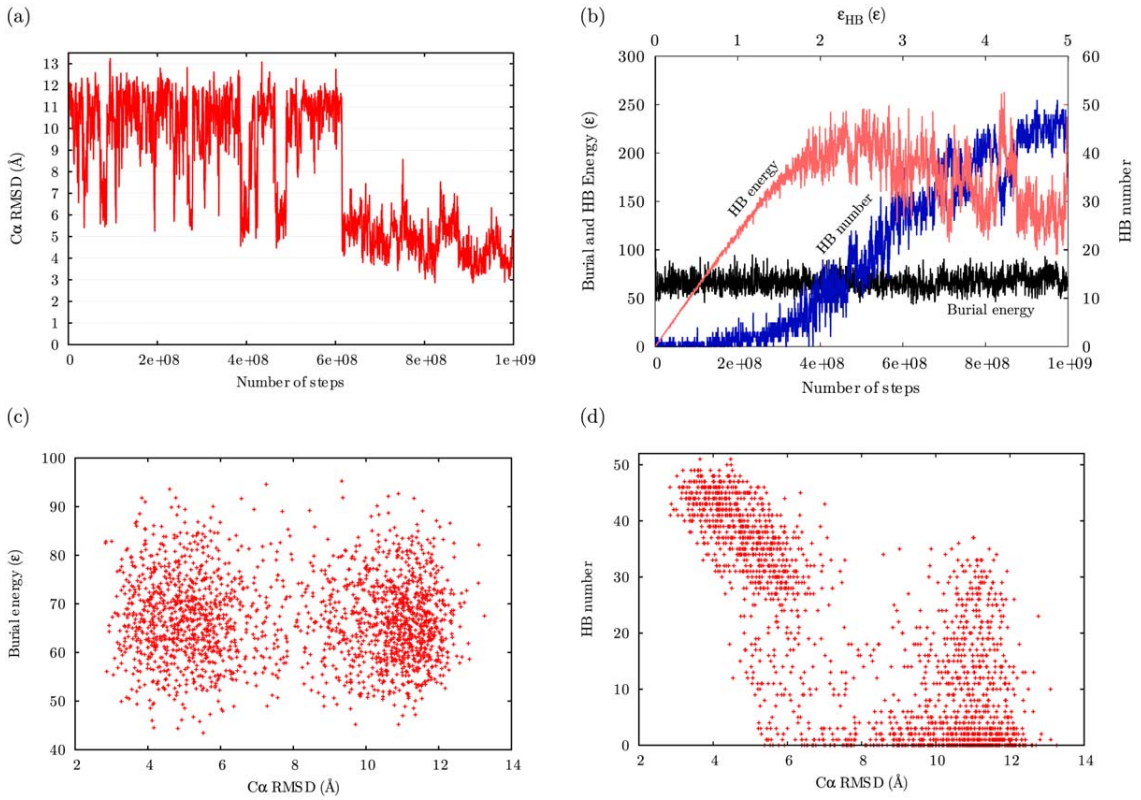

Figure 7

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,

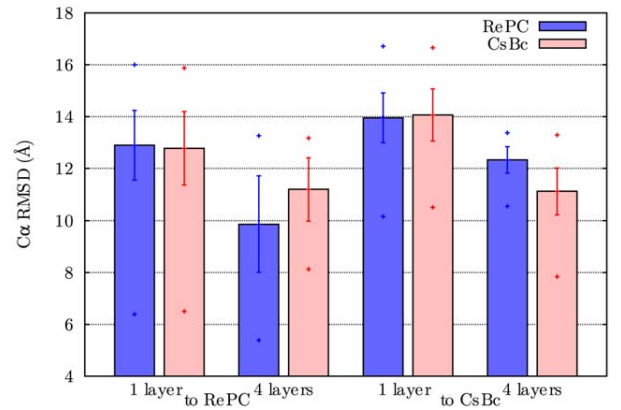

Figure 8

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 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. 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.
16. 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.
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 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.