Transmissão de informação entre seqüência primária, enterramentos atômicos e estrutura tridimensional de proteínas globulares

Orientando: Antonio Luiz Cruz Gomes
Orientador: Antônio Francisco P. de Araújo

Dissertação apresentada ao Programa de Pós-Graduação em Biologia Molecular da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Biologia Molecular.

Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular

Agradecimentos

Aos meus pais, por todas as condições que me deram para continuar os estudos.

À minha família, incluindo além dos meus pais, os irmãos, primos e tios, responsáveis por momentos de confraternização, conversa, apoio essenciais para o dia a dia e muito mais.

Aos amigos do meu semestre, da quadra, da turma do pirulito, das viagens, de toda a biologia que são muito importantes nos momentos de lazer e crises existenciais. Alguns mais próximos, alguns perdemos o contato, alguns se mudam, mas todos com um carinho especial que nos fazem sentir saudades.

Ao pessoal do futebol, nada melhor que uma boa pelada.

A todas as pessoas que passaram pelo laboratório, pelas ajudas, pelas conversas diárias ou simplesmente pela companhia na rotina de trabalho.

Aos funcionários da UnB essenciais para o funcionamento dessa instituição e importantes para minha conclusão de mestrado.

A todos os professores que de alguma forma me ajudaram ou transmitiram conhecimento.

Ao meu orientador, pelo ensinamento e boa convivência durante todo o mestrado e pelo incentivo ao aprendizado de matérias de base teórica, tais como matemática e física.

À CAPES, pelo auxílio financeiro.

A beleza da ciência está em enxergar o mundo cada vez mais complexo de modo cada vez mais simples.

Resumo

Nós investigamos a possibilidade de que enterramentos atômicos, definidos a partir das distâncias ao centro geométrico, sejam preditos a partir da seqüência e contenham informação suficiente para determinar a estrutura tridimensional de proteínas globulares. A segunda hipótese foi testada por meio de simulações de Monte Carlo para 4 proteínas pequenas ($1E0L$, $1IGD$, $1ENH$, $1ORC$), onde a energia de cada átomo diminuia à medida que sua distância ao centro se aproximava da distância nativa. Para todas as proteínas foi observada uma correlação entre essa energia artificial e o drms e as estruturas de energia mais baixas eram essencialmente nativas ($drms <\approx 1,5\mathring{A}$). Para investigar a primeira hipótese, utilizamos um banco de dados de 321 proteínas globulares não homólogas e encontramos que a distribuição de enterramentos atômicos para cada tipo de átomo pode ser descrita por uma função adaptada da distribuição de Fermi-Dirac. Um parâmetro ajustado nessa função é relacionado com a hidrofobicidade e mostrou correlação com dados presentes na literatura. Essa função fornece também um parâmetro que representa um potencial estatístico e mostra ser capaz de distinguir a estrutura nativa de estruturas compactas ao acaso para $1ORC$. Considerando seqüências de proteínas como sendo geradas por uma fonte de informação, e com base na teoria de Shannon, encontramos que seqüências de aminoácidos armazenam cerca de 4 $bits/aminoácido$. A discretização do enterramento permite quantificar sua entropia informacional da mesma forma. Como resultado, encontramos que uma precisão entre 10% e 20% do raio de giro deve ser um limite para discretização do enterramento atômico. Encontramos ainda que a informação possível de ser extraída a partir da seqüência de aminoácidos é maior quanto maior a distinção dos tipos atômicos.

Abstract

We investigated the possibility that atomic burials, defined by the distance from the protein center, might be predicted from the primary sequence and might have enough information to determine the three-dimensional structure of globular proteins. The second hypothesis was tested by Monte Carlo simulation of 4 small proteins ($1E0L$,$1IGD$, $1ENH$, $1ORC$), where the energy of each atom decreased as its distance approached its native distance. For all proteins, there was a correlation between this artificial energy and drms of the final conformations and the lowest energy conformations were essentially native (drms $< 1,5\mathring{A}$). To investigate the first hypothesis, we used a data bank of 321 non-homologous globular proteins and found that the distribution of atomic burials can be fitted by a Fermi-like function. In this function, one parameter is related to hydrophobicity and showed correlation with previous scales. Furthermore, this function also provides a statistical potential which was able to distinguish native from non-native random globular structures. Considering protein sequences as being generated by an information source, in the sense of the Shannon theory, we estimated their informational entropy as 4 $bits/aminoacid$. Discretization of atomic burials allows quantification of their informational entropy. As result, we obtained that a precision between 10% and 20% of the radius of gyration should be a limit to the discretization of atomic burials. We also found that we can extract more information of the aminoacid sequence the more specific we discriminate the atomic type.

Lista de variáveis e símbolos

$\epsilon'_\tau(r)$	Energia efetiva (equação 13).
h_τ	Parâmetro da energia efetiva linear $(\epsilon'_\tau(r)) = \epsilon_\tau(r) = h_\tau r$).
h^*_τ e α^*_τ	Parâmetros da energia efetiva não linear $(\epsilon'_\tau(r) = \epsilon^*_\tau(r) = h^*_\tau r^{\alpha^*_\tau}$).
λ_i	Eixo de inércia (equação 2).
$\rho^*(r)$	Densidade volumétrica de probabilidade normalizada $p(r)/(Ar^2)$.
Asf	Asfericidade (equação 2).
G_{HP}	Grupo onde os átomos são diferenciados em 2 tipos, Hidrofóbicos e Hidrofílicos.
G_{res}	Grupo onde os átomos são diferenciados em 20 letras, representada por cada tipo de resíduo.
G_{at}	Grupo onde cada átomo de cada resíduo é diferenciado.
G_{ph}	Grupo onde cada átomos é difenciado, e, no caso da cadeia principal são diferenciados se estão ou não envolvidos em ponte de hidrogênio.
k	Constante de compactação (equação 1).
n	Número de átomos.
$n(r)$	Distribuição da quantidade de átomos à r (equação 6).
N	Quantidade de resíduos.
$p(r)$	Densidade de probabilidade da distribuição de átomos em relação à r.
$p_\tau(r)$	Densidade de probabilidade da distribuição do átomo do tipo τ (equação 13).
r	Distância reduzida de um átomo ao centro da proteína (equação 6).
R	Distância de um átomo ao centro da proteína.
R_g	Raio de Giro (equação 3).

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O estudo do enovelamento de proteínas tem sua origem no começo do século XX. Em 1929 Wu foi o primeiro a sugerir que a desnaturação de uma proteína ocorria por causa de um efeito de desenovelamamento [1]. Nos anos seguintes, diversos autores estudaram a termodinâmica desse processo. Em 1931, Anson et al. publicaram um artigo descrevendo que a desnaturação de algumas proteínas era um processo reversível [1, 2]. Alguns anos depois começaram a ser considerados estudos relacionados à termodinâmica do dobramento de proteínas [3, 4]. Em 1951, foi publicado um artigo sobre o processo de desnaturação e renaturação de proteínas, onde os autores fizeram uma análise termodinâmica do enovelamento do quimotripsinogênio considerando o processo como sendo de dois estados [3]. Em 1962, Tanford publicou um artigo comparando a energia livre de desnaturação com a contribuição hidrofóbica esperada de cada resíduo [4]. Nos anos seguintes, a reversibilidade da desnaturação era conhecida para algumas proteínas [2,3,5] e foi consolidada a idéia de que a informação da estrutura de uma proteína está em sua seqüência [5–7]. Em 1973, Anfinsen mostrou que a nuclease de estafilococos sintetizada artificialmente também se enovela e mantém a atividade da proteína obtida *in vivo*. Essa mesma proteína pode ser quebrada em um sítio específico e os dois fragmentos gerados são capazes de se dobrar mantendo a atividade da estrutura nativa [7].

Concomitantemente à idéia de que a conformação está codificada na seqüência, houve o desenvolvimento de técnicas de difração de raios-X, sendo que, em 1958, a mioglobina foi a primeira proteína a ter sua estrutura determinada experimentalmente [8]. O conhecimento da estrutura de algumas proteínas possibilitou o aparecimento de novas estratégias para a tentativa de predição da estruturas a partir da seqüência de aminoácidos. Na década de 70, era comum a busca da predição de estruturas secundárias, para a partir de então buscar, eventualmente, a estrutura tridimensional [9, 10]. Em um artigo de 1974, Schulz mostra uma comparação entre 11 métodos desenvolvidos para a predição de estrutura secundária [10].

Ao final dos anos 60, foi levantada uma questão, conhecida como paradoxo de Le-

vinthal, de que uma cadeia peptídica não teria tempo suficiente de explorar o espaço conformacional e alcançar o mínimo termodinâmico, supostamente representado pela estrutura nativa [6]. Em um artigo de 1969, Zwanzig et al. concluiram que uma cadeia de 101 monômeros levaria cerca de 10^{27} anos para explorar aleatoriamente todo o espaço conformacional [11]. Porém, o tempo de renaturação de uma proteína ocorre na ordem de segundos. Para resolver essa questão, Levinthal sugeriu que o enovelamento de uma proteína fosse direcionado por um caminho específico até a estrutura nativa.

A idéia de caminho específico colocou em debate se a estrutura de uma proteína representava apenas um mínimo local de energia livre, possível de ser alcançado cineticamente, ou se representava o mínimo global. Atualmente, a solução mais aceita para o paradoxo de Levinthal é a representação do espaço conformacional como uma superfície de energia livre em forma de funil (figura 1 [12, 13]). Esse modelo satisfaz tanto a hipótese termodinâmica quanto a necessidade de acessibilidade cinética. Segundo essa representação, diversos caminhos direcionam o enovelamento ao mínimo global dessa superfície, que corresponde à estrutura nativa. A forma de funil é consistente com a hipótese de que a energia média dos contatos nativos, ou seja, contatos presentes na estrutura nativa, é menor do que a dos não nativos, de tal modo que a formação de contatos nativos implica em uma diminuição tanto da energia quanto da entropia à medida que o sistema evolui do estado desenovelado para o estado nativo [12]. O aumento da declividade desse funil energético com a profundidade é importante para caracterizar o processo de enovelamento como um processo de dois estados, e um mínimo de energia profundo em relação a estruturas aleatórias é necessário para a estabilidade da estrutura nativa [14]. A profundidade desse mínimo pode ser representada matematicamente pelo cálculo do valor-Z (tradução de "Z-score") da energia da estrutura nativa [15], ou seja, a diferença em unidades de desvio padrão que a energia da estrutura nativa está em relação à energia média.

Segundo Privalov, quatro tipos de interações não covalentes intramoleculares podem existir em uma proteína: interações de van de Waals, pontes salinas, pontes de hidrogênio e as "interações hidrofóbicas" [16]. Exceto pelas interações hidrofóbicas, todas as interações interiores às proteínas também ocorrem entre a proteína e o solvente. A priori,

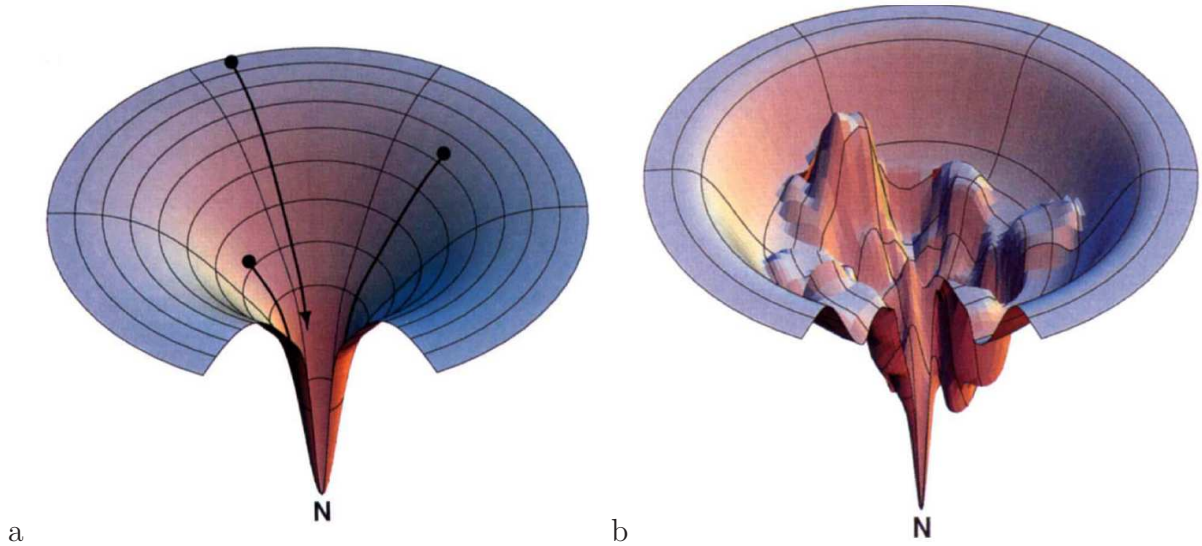


Figura 1: Esquema representando um funil de energia. A área desse gráfico representa os espaços conformacionais, relacionados à entropia. Nota-se que a estrutura nativa encontra-se em um mínimo profundo, acessível cineticamente por causa do funil energético. A energia livre representada inclui o efeito entrópico da água. Em a) temos um funil energético perfeito e em b) um funil energético com rugosidades, caracterizando algumas armadilhas cinéticas. (retirado de [12])

não existe nenhuma razão física que indique que essas interações sejam mais favoráveis para a estrutura nativa. Em conseqüência, torna-se razoável concluir que as interações hidrofóbicas representam a principal causa responsável para a estabilidade da estrutura nativa, mesmo após o estudo de Privalov ter concluído, por meio de dados experimentais e demonstração matemática, que a diferença entre as forças das interações proteína-solvente e proteína-proteína é essencial para as propriedades termodinâmicas de uma proteína.

A explicação para o efeito hidrofóbico é descrita por Dill et al. em [17]. Segundo essa explicação, grupos apolares, ao fazerem contato com um solvente aquoso, organizam as moléculas de água ao seu redor, criando uma camada de solvatação relativamente rígida. Essa organização da água diminui a entropia do sistema, desfavorecendo a mistura. Por esse princípio, quanto maior a hidrofobicidade de um grupo atômico, maior será sua tendência em se esconder da água. Essa conclusão está de acordo com resultados experimentais e é reforçada por resultados de um modelo em duas dimensões. Nesse modelo, as moléculas de água são representadas como um disco contendo três eixos capazes de interagir por pontes de hidrogênio. Os resultados desse modelo reproduzem propriedades anômalas da água e corroboram a idéia anterior descrita para o efeito hidrofóbico [17].

O efeito hidrofóbico em uma proteína ocorre por causa da presença de grupos apolares ao longo da seqüência. É a cadeia lateral que diferencia a hidrofobicidade dos resíduos de aminoácidos que compõem as proteínas. Fauchère & Pliska descrevem experimentalmente a hidrofobicidade de cada aminoácido a partir de sua energia livre de transferência do octanol para água [18]. Miller et al. mostraram que existe correlação entre as hidrofobicidades dos resíduos calculadas por Fauchère & Pliska e suas exposições médias ao solvente, observadas em estruturas de proteínas obtidas por cristalografia [19].

Evidências experimentais mostram que a desnaturação de uma proteína ocorre como uma transição de fase de 1ª ordem, ou seja, a desnaturação ocorre em um processo de tudo ou nada. Em [16], Privalov caracteriza a desnaturação de uma proteína como uma transição de fase e ilustra o fato com experimentos que mostram o processo de desnaturação da Lisozima (desnaturação por temperatura, pH e GdmCl). Em alguns modelos computacionais, a transição de fase pode ser visualizada a partir de uma distribuição bi-

modal, representada pelo estado nativo e pelo estado desnaturado [20,21]. A transição de fase facilita o estudo termodinâmico do processo de enovelamento ao permitir diferenciá-lo em dois estados.

A cadeia principal de uma proteína é formada por grupos polares, representados por um átomo de oxigênio e um de hidrogênio ligado a nitrogênio, capazes de interagir por pontes de hidrogênio. Como conseqüência ao enterramento de regiões hidrofóbicas, há o enterramento de grupos polares da cadeia principal. Ao se enterrar, um grupo polar precisa desfazer interações fortes com o solvente. A compensação do rompimento dessas interações ocorre com a formação de pontes de hidrogênio intramoleculares. Esse fato, que restringe a possibilidade de conformações nativas, é evidenciado pelas estruturas secundárias α-hélices e β-folhas. Essas estruturas secundárias são comuns em proteínas e se formam por padrões de interação de ponte de hidrogênio entre os átomos da cadeia principal. A restrição de ponte de hidrogênio limita as conformações acessíveis e, por essa razão, se torna um fator importante para ser considerado em potenciais estatísticos. Hoang et al. fizeram um modelo considerando pontes de hidrogênio e interações hidrofóbicas. Nesse modelo, as seqüências eram diferenciadas entre resíduos H e P, e somente os resíduos H contribuíam para as interações hidrofóbicas. Simulações de Monte Carlo utilizando esse modelo mostraram que o peso considerado para as pontes de hidrogênio é um fator importante não apenas para formação de estruturas secundárias, como também para a determinação do tipo de estrutura secundária predominante [22].

Modelos com características de proteína são criados para desenvolver e testar teorias. Nos modelos mais simples, o espaço conformacional se limita a posições discretas, em duas ou três dimensões, acessíveis aos monômeros. Algumas características observadas experimentalmente podem ser obtidas através de modelos minimalistas. O pressuposto de que a evolução selecionou interações minimamente frustradas, ou seja, que contatos favoráveis se formam na direção da estrutura nativa, inspira a consideração de que a cinética do enovelamento é guiada por seus contatos nativos. Uma aplicação dessa visão é a descrição da cinética e termodinâmica de uma proteína a partir da função de energia *GO*. Na função *GO*, apenas contatos nativos contribuem energeticamente para a energia

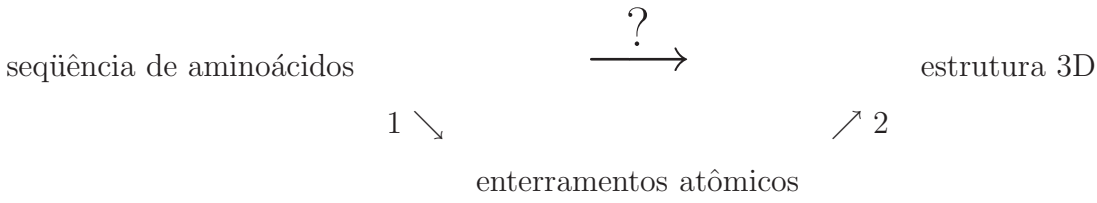
da estrutura. Onuchic et al assume esse fato como exemplo para construção de um funil de energia [13] e ilustra esse processo para o repressor Arc, em [23]. Porém, em 2006, Garcia et Al. mostrou, para um modelo tridimensional, que a cinética de dobramento pode depender da função de energia utilizada [24].

O modelo *GO* não é um modelo relacionado à predição de estrutura de proteínas, visto que a função de energia desse modelo necessita da informação da estrutura nativa para ser estabelecida. Um modelo para predição de estrutura deve ter sua função de energia estabelecida a partir da seqüência primária. Um modelo minimalista de duas dimensões foi usado por Lan & Dill em 1989 [25]. Para esse modelo existiam dois tipos de monômeros, H e P, representando hidrofóbicos e polares, dos quais o contato HH contribuía negativamente para a energia da estrutura. Nesse trabalho, os autores obtiveram algumas seqüências que tinham poucas conformações acessíveis. Na década de 80 e 90, era comum considerar a estrutura nativa de modelos protéicos como maximamente compacta. Um modelo da década de 90 permitia uma estrutura nativa compacta, porém, esse modelo levava em consideração uma potencial de contato, dependente de uma combinação, 2 a 2, de 20 letras [15, 26].

Ao final da década de 90 [20], Araújo obteve um método de desenhar seqüências capazes de se enovelar para uma estrutura nativa única a partir de um potencial simples dependente apenas da hidrofobicidade e do número de contatos de cada monômero. De acordo com esse modelo, estruturas segregadas teriam maior probabilidade de ser potencialmente nativa. Uma estrutura é considerada maximamente segregada se metade de seus monômeros se encontrarem enterrados e a outra metade expostos. Esse modelo foi eficiente para enovelar, em 2 estados, modelos de 2 e 3 dimensões [20, 27] e mostra a possibilidade de se reproduzir comportamento de proteína diferenciando os monômeros apenas pela sua hidrofobicidade.

Pelo princípio do efeito hidrofóbico, grupos apolares tendem a se esconder do solvente. Por esse motivo surge a expressão "enterramento", referente ao quanto um átomo ou resíduo está escondido do solvente. Diversas definições podem ser utilizadas para representar o enterramento em uma proteína. O enterramento de um átomo é comumente

definido pela fração de sua superfície acessível ao solvente ou pelo número de contatos. Por causa de sua relação com a hidrofobicidade, uma boa definição de enterramento é importante para um modelo de predição de estrutura. Se, de alguma forma, o efeito hidrofóbico está codificado na seqüência de aminoácidos, é possível a predição de enterramentos atômicos preferenciais e, a partir de então, a estrutura nativa, conforme o seguinte esquema:



Neste trabalho, definimos o enterramento atômico como a distância de cada átomo ao centro da proteína. Buscamos verificar, a partir dessa definição, se os enterramentos atômicos contém informação suficiente para determinar a estrutura de proteínas globulares. Por causa da simetria da estrutura de proteínas globulares, espera-se que o enterramento de um átomo esteja relacionado à hidrofobicidade do grupo atômico a que pertence. Dessa forma, enterramentos preferenciais poderiam ser estabelecidos a partir da seqüência de aminoácidos.

2 Objetivos

2.1 Objetivo geral

- Contribuir para o entendimento do enovelamento proteico e eventual predição de estrutura de proteínas globulares a partir da seqüência de aminoácidos via enterramentos atômicos.

2.2 Objetivos específicos

- Analisar a distribuição de enterramentos atômicos em proteínas globulares.

- Predizer enterramentos atômicos preferenciais a partir da seqüência de aminoácidos.

- Verificar se enterramentos atômicos contêm informação suficiente para enovelar proteínas globulares.

- Otimizar a quantidade de informação a ser extraída da seqüência de aminoácidos para obter enterramentos atômicos.

3 Descrição de enterramentos atômicos em proteínas globulares

3.1 Características das proteínas do banco de dados

O enterramento de um átomo está relacionado com sua exposição ao solvente. Em proteínas globulares, a simetria radial permite inferir que essa exposição está relacionada à distância de um átomo ao centro da proteína. Por esse motivo selecionamos no banco de dados apenas proteínas globulares e representaremos o enterramento dos átomos pela sua distância ao centro da proteína.

A seleção do banco de dados foi feita a partir de uma lista, contendo 731 cadeias de proteínas obtidas por cristalografia de raio-X, com identidade menor que 25% e resolução melhor que 2.5 $\mathring{A}$, de acordo com os critérios descritos pelo pdbselect [28]. Dessas 731 cadeias de proteínas, foram selecionadas apenas aquelas que que satisfizeram o critério de globularidade, resultando em um total de 321 proteínas.

Definimos como globular uma proteína que satisfaça a condição de ser compacta e esférica. Para isso, ela deve satisfazer as condições $k \leq 2,9$ e $Asf \leq 0,1$. Onde,

$$k = \frac{R_g}{N^{1/3}} \tag{1}$$

é uma constante positiva, chamada de constante de compactação. Quanto menor for o seu valor, mais compacta será a proteína e

$$Asf = \frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}{2(\lambda_1 + \lambda_2 + \lambda_3)^2} \tag{2}$$

a asfericidade. O valor de Asf pode variar de 0 a 1. A asfericidade de uma esfera é zero [29] . R_g é o raio de giro da proteína e é calculado a partir das distâncias R_i de cada átomo i ao centro geométrico de uma proteína, conforme a equação:

$$R_g = \sqrt{\frac{\sum_{i=1}^{n} R_i^2}{n}} \tag{3}$$

e, λ_1, λ_2 e λ_3 são chamados de eixo de inércia de uma proteína, calculado pelas raízes do determinante da matriz

$$\begin{pmatrix} \sum (x_i - \overline{x})^2 - \lambda & \sum (x_i - \overline{x})(y_i - \overline{y}) & \sum (x_i - \overline{x})(z_i - \overline{z}) \\ \sum (y_i - \overline{y})(x_i - \overline{x}) & \sum (y_i - \overline{y})^2 - \lambda & \sum (y_i - \overline{y})(z_i - \overline{z}) \\ \sum (z_i - \overline{z})(x_i - \overline{x}) & \sum (z_i - \overline{z})(y_i - \overline{y}) & \sum (z_i - \overline{z})^2 - \lambda \end{pmatrix}$$

onde x_i, y_i e z_i são as coordenadas do átomo i, sendo os somatórios em cada termo realizados de $i = 1$ até $i = n$, enquanto $\overline{x}$, $\overline{y}$ e $\overline{z}$ representam as médias das coordenadas dos n átomos, ou seja, as coordenadas do centro geométrico da proteína. A distribuição de proteínas deacordo com sua asfericidade e constante de compactação é mostrada na figura 2.

Selecionado o banco de dados, definimos 5 conjuntos disjuntos, que separam as cadeias de proteínas de acordo com o tamanho. $N \leq 100$ (55), $100 \leq N < 150$ resíduos (73), $150 \leq N < 200$ (66) , $200 \leq N < 300$ (74) , $300 < N$ (53), onde N representa o tamanho em quantidade de resíduos de aminoácidos e os valores entre parênteses indicam a quantidade de cadeias em cada conjunto. Para esses diferentes grupos, representamos por $n(R) = (1/N_p)(\delta n(R)/\delta R)$, a distribuição média da quantidade de átomos relativa à distância R do seu centro geométrico. Onde N_p representa o número de proteínas para um conjunto p e $\delta n(R)$ representa a quantidade de átomos encontrados em uma casca esférica de raio R, definida como o intervalo $R - \delta R/2 \leq R < R + \delta R/2$ (figura 3).

Conforme mostrado na figura 3, para todos os grupos de proteínas, a quantidade de átomos cresce quadraticamente para R pequeno, atinge um máximo a uma distância R_{max} dependente do tamanho e em seguida sofre uma queda até se anular. Esse comportamento pode ser observado na figura $3a$. Em $3b$ verificamos os mesmos dados em uma escala log-log. Uma reta de inclinação 2 evidencia o crescimento quadrático para raios pequenos (crescimento proporcional ao volume de uma casca esférica de raio R). Na figura $3c$ a representação monolog mostra a queda quase exponencial para R grande. Em $3d$ é mostrada a densidade volumétrica de átomos em relação a R, ou seja $\frac{n(R)}{4\pi R^2}$. Esses dados sugerem que a densidade média em uma região central de uma proteína globular é próximo de uma constante, seguido de uma queda abrupta e são consistentes com a idéia de um núcleo compacto e uma região menos densa, hidrofílica, acessível ao solvente.

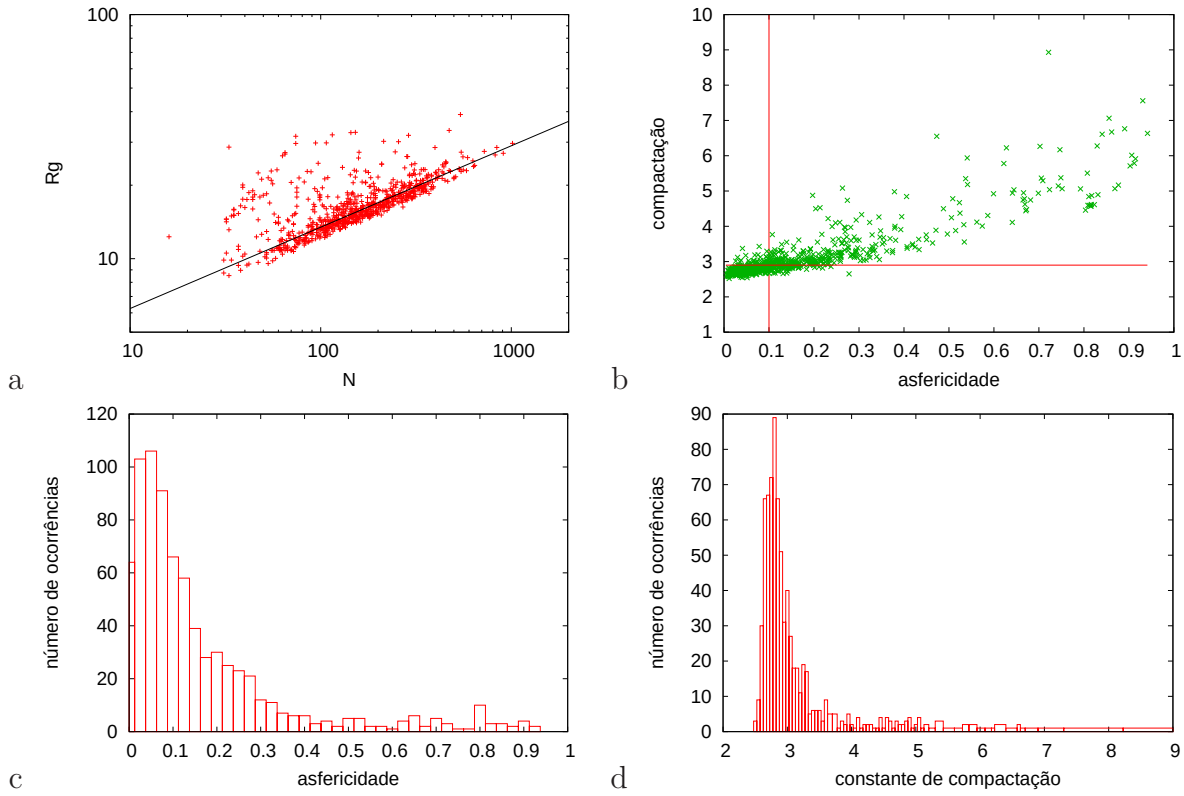


Figura 2: Critérios para seleção de proteínas globulares. a) R_g em função da quantidade de aminoácidos N. A linha que corta o gráfico representa $k = 2.9$. Há 448 proteínas abaixo dessa linha. Em b) acrescentamos o critério de asfericidade para a seleção do banco, e encontramos 321 proteínas que se adequaram a essas condições. Em c) a distribuição em histograma das proteínas de todo o banco segundo a asfericidade. Em d) a distribuição em histograma das proteínas segundo a constante de compactação.

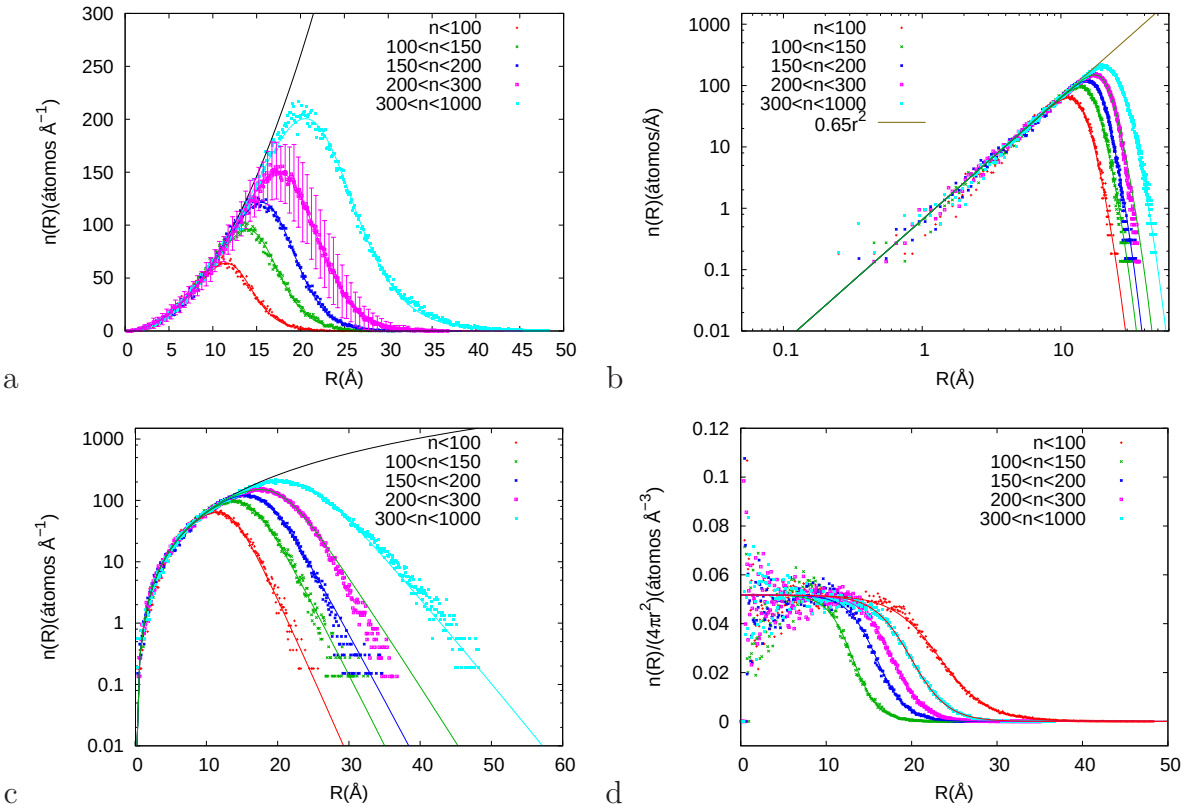


Figura 3: O padrão de distribuição da quantidade de átomos de acordo com a distância ao centro geométrico ($n(R)$) é semelhante para proteínas de diferentes tamanhos. Todos os grupos de proteínas mostram um crescimento quadrático para distâncias pequenas e atingem um máximo seguido por uma queda até se anular para distâncias grandes. O raio que representa o ponto de máximo é maior para proteínas maiores. Todas essas características podem ser visualizadas em a. A representação em escala loglog facilita a visualização do crescimento quadrático em R pequeno (b) e a representação monolog facilita a visualização de uma queda quase exponencial para R grande (c). Em d mostramos a densidade volumétrica de átomos em função de R. Para R pequeno, a densidade é máxima, próximo de uma constante. Em seguida ocorre uma queda abrupta até a densidade se anular. Matematicamente, o erro para R pequeno diverge e esse fato justifica a dispersão dos pontos nessa região.

De modo interessante, a densidade volumétrica de átomos mostrada em $3d$ comporta-se de modo semelhante à distribuição de Fermi-Dirac, mostrada na figura 4. Combinando a função que descreve a distribuição de Fermi-Dirac

$$F(x) = \frac{e^{-\beta(x-\mu)}}{1 + e^{-\beta(x-\mu)}} \tag{4}$$

com o crescimento do volume de uma casca esférica, conseguimos uma função contínua que descreve a distribuição dos átomos em relação R, conforme a equação 5 a seguir:

$$n(x) = AF(x)x^2 = \frac{Ax^2 e^{-\beta(x-\mu)}}{1 + e^{-\beta(x-\mu)}}. \tag{5}$$

onde A, β e μ são constantes.

O ajuste dessa função foi feito para cada grupo de proteínas e é mostrado na figura 3 por linhas contínuas. As barras de erro foram representadas apenas para o grupo cujas proteínas possuem entre 200 e 300 resíduos de aminoácidos. Percebe-se pelo gráfico $3d$ que a densidade volumétrica de átomos $\rho(R) = n(R)/(4\pi R^2)$ é semelhante para todos os grupos de proteínas, com a principal diferença na variável μ, relacionada à distância onde a densidade $\rho(R)$ cai para $\rho_0/2$, em que $\rho_0 = \rho(0) = A/(4\pi)$ representa a densidade máxima de átomos. $\rho_0 = 0.052$ $\acute{a}tomos/\mathring{A}^3$ está relacionado à densidade de átomos por volume, indicando que cada átomo ocuparia um volume médio de 19.3 $\mathring{A}^3$. Esse resultado é condizente com o resultado obtido por Tsai et al, onde o volume ocupado por um átomo varia entre $14,4$ e $22,25$ $\mathring{A}$ [30].

A distribuição de átomos em relação a R é semelhante para proteínas de diferentes tamanhos, tendo como diferença evidente a posição do R_{max}. Definimos a variável distância reduzida

$$r = \frac{R}{R_g} \tag{6}$$

da mesma forma que Meirovitch et al. em [31, 32]. Ao utilizar a distância reduzida como variável, as distribuições dos grupos de proteínas de diferentes tamanhos, mostradas na figura 3, convergem para um distribuição comum. A figura 5 mostra essa convergência

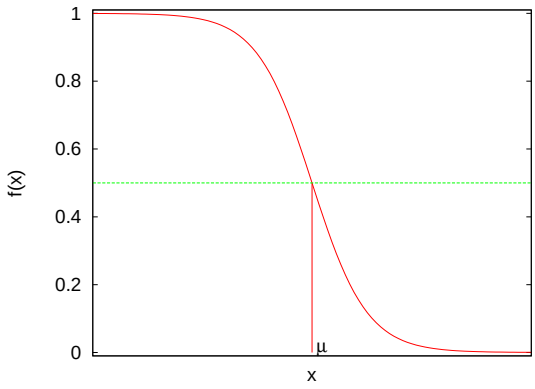


Figura 4: Distribuição de Fermi-Dirac. O parâmetro β está relacionado a abruptalidade da queda da curva e o parâmetro μ está relacionado com o ponto de inflexação da curva.

através da distribuição normalizada, $p(r)$, definida na equação

$$p(r) = \frac{n(r)}{n_o} \tag{7}$$

onde $n(x)$ está definido na equação 5 e n_0 representa a quantidade de átomos amostradas. No mesmo gráfico está representado $P(r)$ e $\rho^*(r) = p(r)/(Ar^2)$, que representam a freqüência relativa de átomos com distância reduzida menor que r e a densidade volumétrica de probabilidade, normalizada de tal modo que $\rho^*(0) = 1$.

De acordo com um modelo proposto por Araújo [20], o valor-Z da energia da estrutura nativa de uma proteína é máximo quando essa estrutura é segregada, ou seja, os resíduos hidrofóbicos maximizam sua condição "enterrada" e os hidrofílicos maximizam sua exposição. O padrão de distribuição dos átomos mostra uma curva "simétrica", que pode ser discretizada de modo que cerca de metade dos átomos se encontram na região de alta densidade e metade na região de baixa densidade. Essa observação pode estar relacionada com a segregação estrutural e ser essencial para uma estrutura nativa. De acordo com o modelo proposto por Araújo, estruturas segregadas têm maior probabilidade de serem potencialmente nativa. Esse modelo leva em consideração um potencial simples, não-específico, capaz de enovelar cadeias em dois estados para modelos de rede em duas e três dimensões [20, 27].

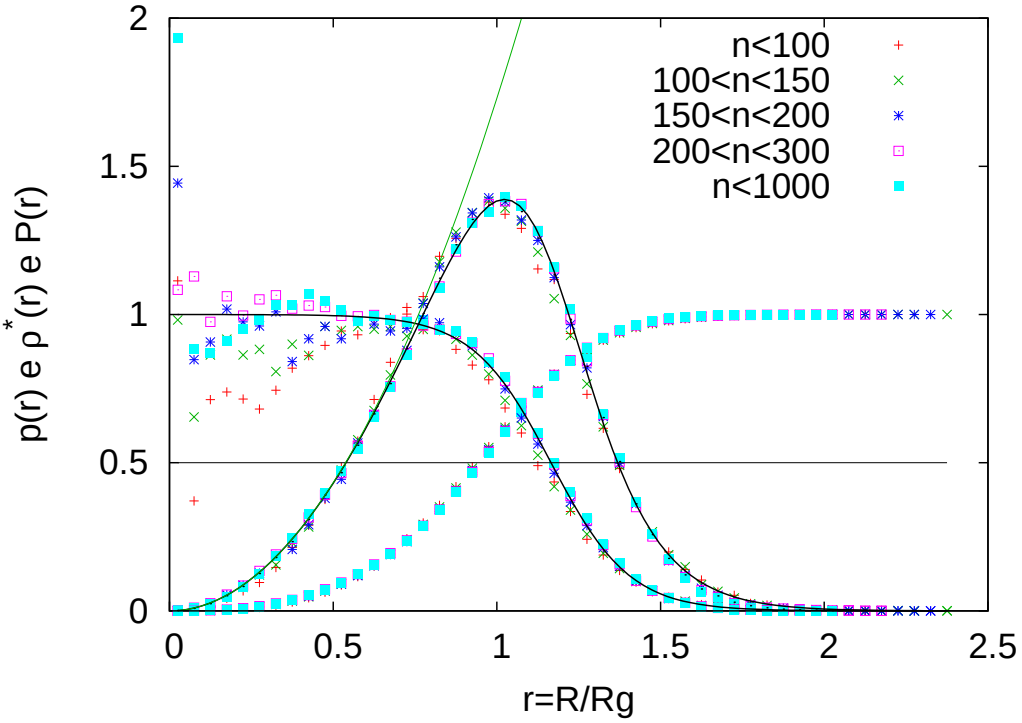


Figura 5: A distribuição da probabilidade $p(r)$ é semelhante para todos os 5 grupos de proteínas. A densidade volumétrica de probabilidade mostra um comportamento semelhante à distribuição Fermi-Dirac e um ajuste da curva Fermi-Dirac foi feito, obtendo $\beta = 8,37$ e $\mu = 1,166$. Essa densidade é máxima quando $r = 0$, onde $\rho^*(r) = 1$. Cerca de metade dos átomos estão em uma região com densidade menor que 0,87 e a outra metade em uma região com densidade maior que 0,87; o raio para esse corte ocorre quando $P(r) = 0,5$, ou seja $r \approx 0,94$

3.2 Segregação estrutural

Para o modelo de rede criado por Araújo [20], a segregação de uma estrutura seria máxima se metade de seus monômeros fizessem o máximo de contatos e a outra metade não fizesse nenhum contato. Em nosso trabalho, a segregação é relacionada à densidade de átomos por volume. Desse modo, um raio de corte separa a região cuja densidade atômica estaria relacionada a uma região hidrofóbica daquela em que os átomos estariam supostamente acessível ao solvente (densidade mais baixa).

Matematicamente, a distribuição de Fermi-Dirac é caracterizada por uma região de densidade quase constante, $\rho(r) \approx \rho_0$ para $r \leq \mu - 1/\beta$ e uma queda de ρ_0 para quase 0 no intervalo $\mu - 1/\beta \leq r \leq \mu + 1/\beta$ ou $r_0 < r < r_0 + 2/\beta$. A quantidade de átomos $N(r)$ pertencente a um raio menor ou igual a r é definida por $N(r) = \int_0^r n(r')dr' = \int_0^r Ar'^2F(r')dr'$, onde $n(r')$ segue a definição dada pela equação 5. Desse modo, temos que:

$$N_{r_0} = \int_0^{r_0} Ar'^2F(r')dr' \approx \frac{Ar_0^3}{3} \tag{8}$$

representa a quantidade de átomos contida no interior de uma esfera de raio r_0 e

$$N_t = \int_0^{\infty} Ar'^2F(r')dr' \approx \frac{A\mu^3}{3} + \frac{A\pi^2}{3\beta^2}\mu. \tag{9}$$

representa a quantidade total de átomos. Para resolver a integração acima, usamos a aproximação:

$$\int_0^{\infty} \phi(x)F(x)dx \approx \int_0^{\mu} \phi(x)dx + \frac{\pi^2}{6\beta^2}\phi'(x)|_{(x=\mu)} \tag{10}$$

descrita em [33].

Ao supor que a estrutura de uma proteína é maximamente segregada quando $N_t = 2N_{r_0}$, então:

$$\mu^3 + \frac{\pi^2}{\beta^2}\mu = 2(\mu - 1/\beta)^3 \tag{11}$$

Multiplicando todos os termos por β^3, essa igualdade é rearranjada e é obtido o polinômio de terceiro grau $G(\beta\mu)$ descrito a seguir:

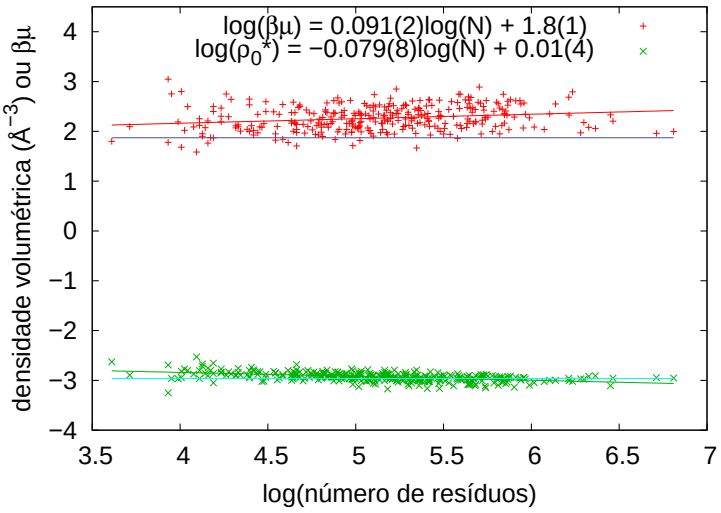


Figura 6: O produto $\beta\mu$ cresce muito devagar com o aumento do número de resíduo, mantendo-se em torno de 10 ± 3 (a). Pela definição do raio de corte em r_0, o produto $\beta\mu$ deveria ser constante e próximo de $6,63$, porém uma outra definição para r_0 poderia melhorar esse resultado (detalhes no texto). A densidade máxima de átomos, no centro da proteína, manteve-se próximo de $0,052/\mathring{A}^3$ para todas as proteínas.

$$G(\beta\mu) = (\beta\mu)^3 - 6(\beta\mu)^2 + (6 - \pi^2)\beta\mu - 2 = 0 \tag{12}$$

A solução dessa equação ocorre quando $\beta\mu \approx 6,63$. Por essa definição, a segregação estrutural é máxima quando o produto $\beta\mu$ se encontra próximo de $6,63$. Para verificar essa previsão teórica, ajustamos a equação 5 para cada proteína individualmente e encontramos que o produto $\beta\mu$ estaria em torno de 10 ± 3. Pela figura 6, o produto $\beta\mu$ estaria espalhado em torno de um valor constante. Apesar desse número ser bem acima daquele obtido na predição teórica, ele não é absolutamente inconsistente e uma definição diferente para r_0 pode melhorar a compatibilidade desse resultado.

A representação da segregação estrutural pode ser visualizada na figura 5. Nessa figura temos a representação da densidade volumétrica normalizada da distribuição de probabilidade, $\rho(r)^*$, em relação à distância reduzida. Juntamente com $\rho(r)^*$, temos a

função $P(r)$ (probabilidade de encontrar um átomo em um raio menor ou igual à r). Visualizando pelo gráfico, temos que $P(r) = 0,5$ implica em $r = 0,94$ e $\rho^*(r) \approx 0,87$.

O ajuste de $\rho^*(r)$ gerou aos parâmetros β, μ os valores $8,37 \pm 0,07$ e $1,166 \pm 0,002$ respectivamente. De acordo com esse ajuste e a definição de r_0 utilizada como raio de corte para definir segregação, teríamos $r_0 = 1,04$. Supostamente, $P(r_0) = 0,5$, porém, graficamente observamos que $P(0.94) \approx 0.5$, ou seja, o raio de corte definido para determinar o valor ideal do produto $\beta\mu$ é um pouco acima do raio observado experimentalmente. Esse fato poderia justificar a diferença entre o valor esperado 6.63 e o valor obtido 10 ± 3 para o produto $\beta\mu$ estimado pela distribuição de cada proteína. O fato do valor de r_0 ser um pouco menor que o definido previamente implica em uma queda mais abrupta da densidade de átomos e, dessa forma, um aumento do valor de β. Como conseqüência desse fato, o produto $\beta\mu$ esperado se torna um pouco maior que o valor descrito previamente.

3.3 Descrição da distribuição de átomos para grupos atômicos específicos

A informação da distância reduzida de cada átomo parece ser suficiente para determinar a estrutura nativa, pelo menos para algumas proteínas pequenas (figura 15). Esse resultado reforça a busca por um método de predição das distâncias reduzidas a partir de sua seqüência primária de aminoácidos. Neste trabalho analisamos apenas proteínas globulares, imposição que relaciona a distância reduzida diretamente à exposição ao solvente. Conseqüentemente, distâncias reduzidas preferenciais podem ser estimadas a partir da hidrofobicidade de cada grupo atômico.

O modelo mais simples para predição de enterramentos preferenciais é diferenciando os átomos apenas por pertencerem a um resíduo hidrofóbico ou hidrofílico. Definimos como G_{HP} um conjunto de dois elementos, H, representando os resíduos hidrofóbicos (A, H, Y, M, L, W, C, V, I, F) e P, representando os resíduos hidrofílicos (K, E, D, Q, N, P, R, S, G, T). A distribuição específica dos átomos pertencentes a esses dois grupos pode ser visualizada na figura 7. Nessa figura é evidente a preferência dos átomos hidrofóbicos

à distâncias menores, quando comparado aos átomos hidrofílicos. Uma adaptação da equação 5, mostrada na equação 13, foi utilizada para ajustar essa distribuição e está representada na figura 7 por linhas contínuas.

$$p_\tau(r) = Ar^2 \frac{e^{-\beta(\epsilon'_\tau(r)-\mu_\tau)}}{1+e^{-\beta(r-\mu)}} \tag{13}$$

Nessa equação, $p_\tau(r)$ representa a densidade de probabilidade de um átomo τ em relação à distância reduzida r. O ajuste foi feito para duas formas do parâmetro $\epsilon'_\tau(r)$. Na forma mais simples, a energia efetiva é considerada como se variasse linearmente com r, de forma que $\epsilon'_\tau(r) = \epsilon_\tau(r) = h_\tau r$ (figura 7-a). Nesse caso, o parâmetro h_τ está diretamente relacionado com a hidrofobicidade do grupo atômico ajustado e os valores $h_\tau = 1,138 \pm 0,008$ para os resíduos hidrofóbicos e $h_\tau = 0.879 \pm 0,005$ para os resíduos hidrofílicos corroboram essa idéia. O ajuste foi aprimorado ao se acrescentar mais um parâmetro para a energia efetiva. De forma que $\epsilon'_\tau(r) = \epsilon^*_\tau(r) = h^*_\tau r^{\alpha_\tau}(r)$. O resultado desse ajuste é mostrado em 7b.

Os valores de μ e β foram obtidos do ajuste de $p(r)$, pertencente à equação 7. $\epsilon'_\tau(r)$ é chamado de energia efetiva e está relacionado com a hidrofobicidade do átomo do tipo τ. μ_τ é chamado de potencial químico e seu valor pode ser calculado ao saber $\epsilon'_\tau(r)$, de tal modo que $\int_0^\infty p_\tau(r) = 1$.

Modelos mais detalhados são possíveis ao aumentar a especificidade que diferencia os tipos de átomos. Definimos o conjunto G_{res}, contendo 20 letras. Nesse conjunto, cada átomo é classificado de acordo com o tipo de resíduo ao qual pertence. De mesmo modo a G_{HP}, ajustamos a equação 13 para a distribuição de cada um dos resíduos e obtivemos a hidrofobicidade h_τ de cada resíduo. A distribuição específica desse ajuste é comparada com a distribuição padrão $p(r)$ e é mostrada nas figuras 8 e 9. Com o resultado dos ajustes, encontramos uma classificação dos resíduos de acordo com a hidrofobicidade h_τ. Essa escala mostra a Lisina como o resíduo mais hidrofílico ($h_K = 0,79$) e a Fenilalanina como o mais hidrofóbico ($h_F = 1,19$). Na tabela 1 está representado os valores de h_τ, obtido para cada um dos resíduos.

No início da década de 80, Meirovitch et al. introduziram o conceito de distância

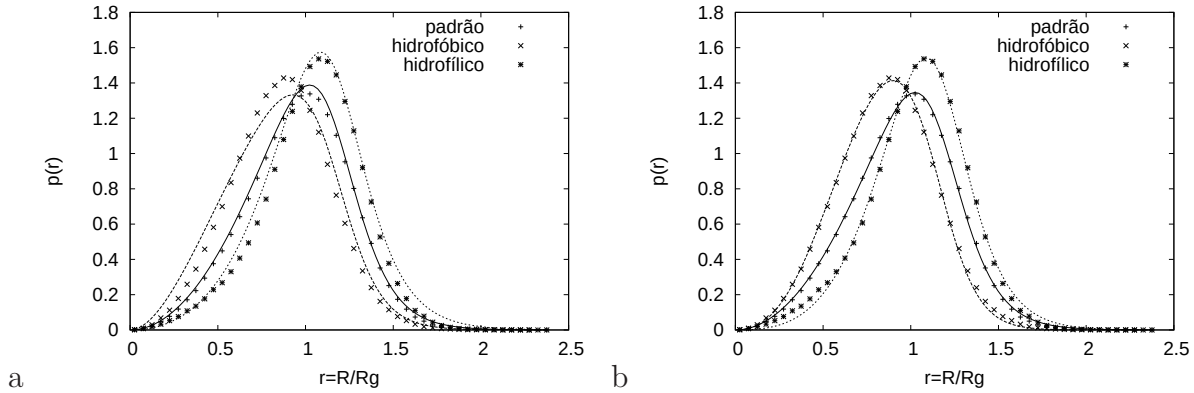


Figura 7: A distribuição dos átomos hidrofóbicos é deslocada para a esquerda enquanto a dos hidrofílicos é deslocada para a direita. As linhas contínuas ajustadas em (a) são referentes à equação 13, considerando a energia efetiva linear ($\epsilon(r) = h_\tau r$). Em (b), a energia efetiva foi considerada não linear ($\epsilon^*(r) = h^*_\tau r^{\alpha_\tau}$). A melhoria do ajuste ao se acrescentar o parâmetro α_τ para a energia efetiva pode ser observada ao se comparar os gráficos de a e b. Onde a soma dos quadrados da diferença entre a curva ajustada e os pontos distribuídos foi de $0,19$ e $0,07$ para o caso linear e de $0,02$ e $0,03$ para o caso não linear para o grupo hidrofóbico e hidrofílico respectivamente.

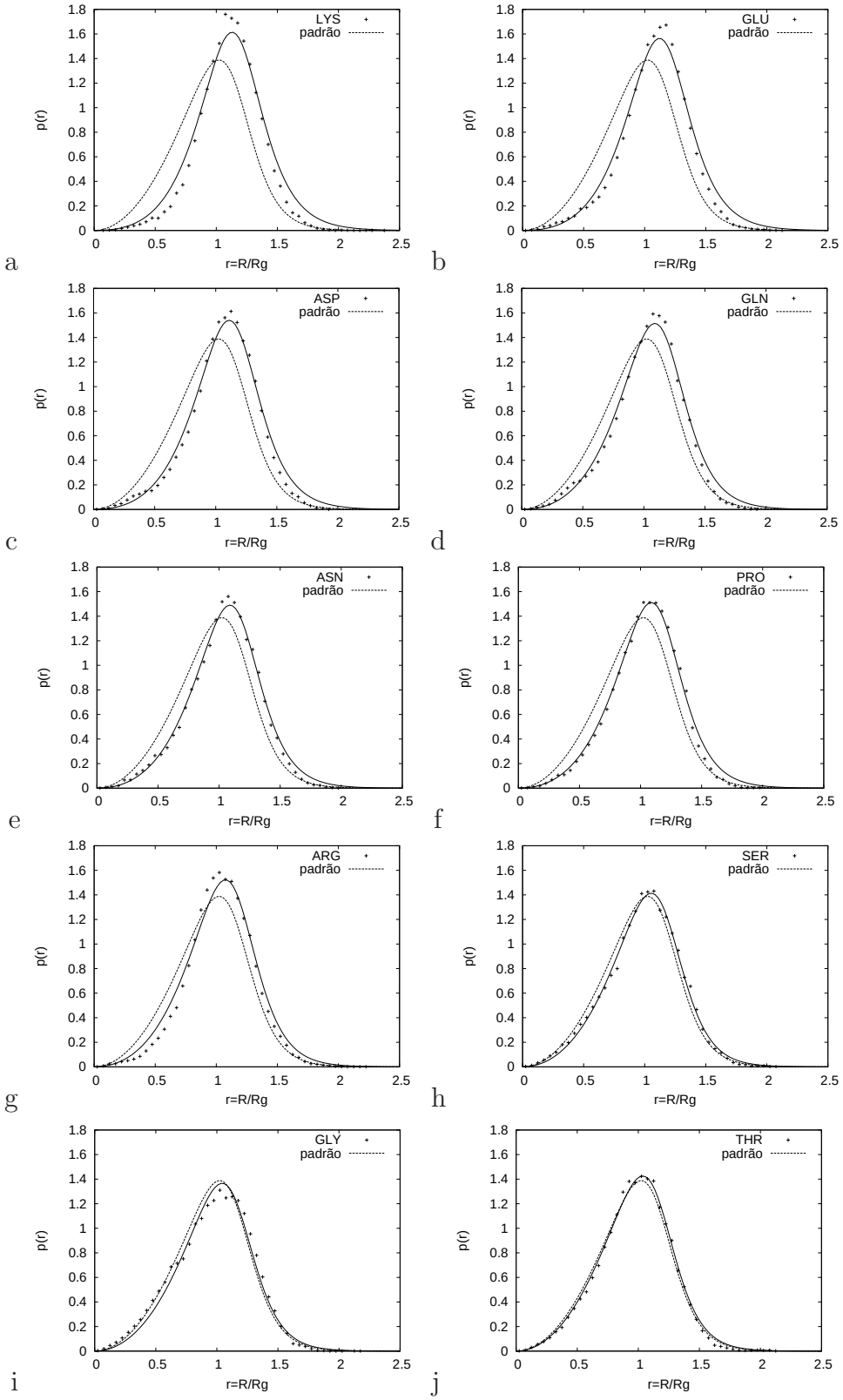


Figura 8: Ajuste da equação 13, com energia efetiva linear, para a distribuição dos resíduos polares.

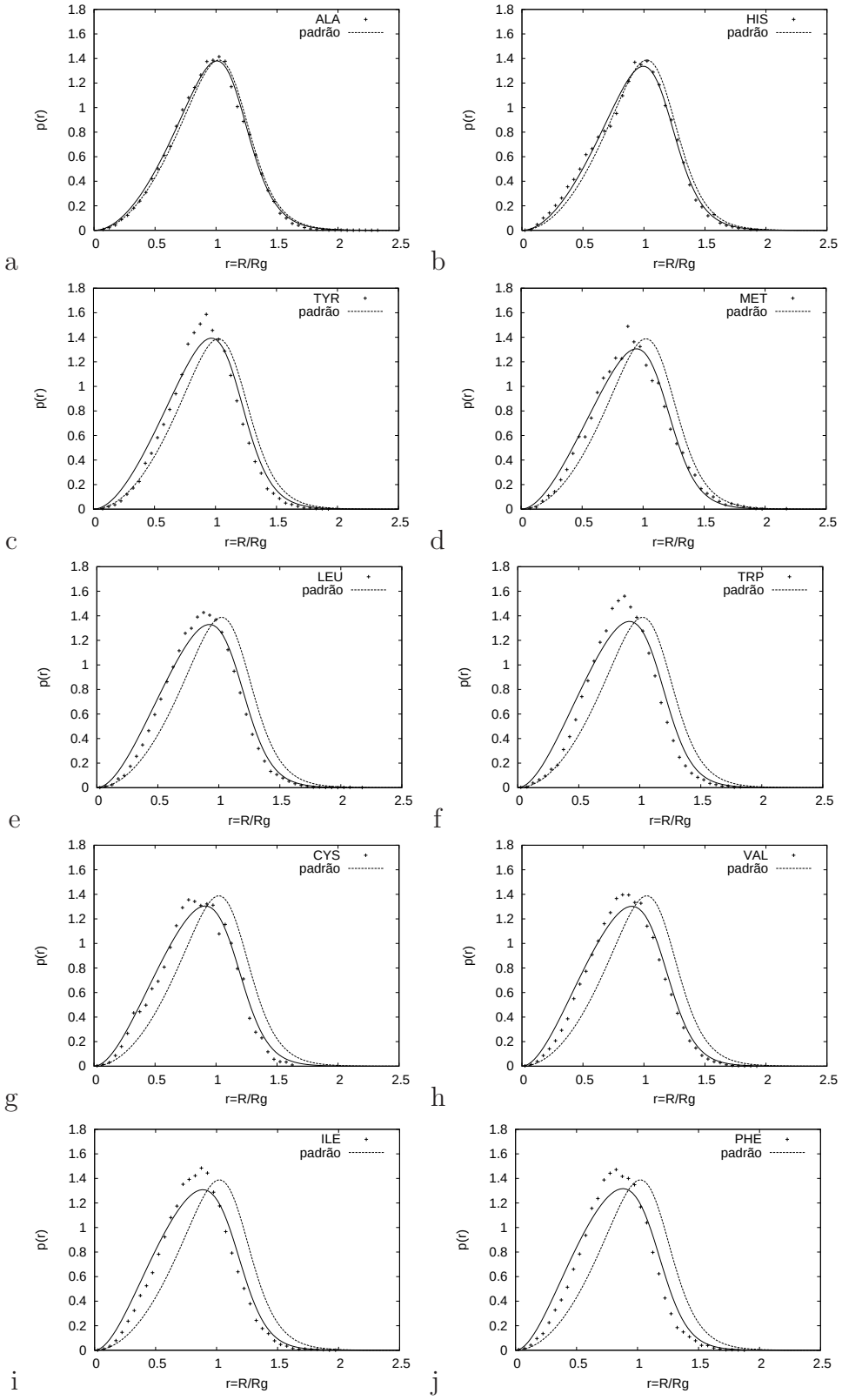


Figura 9: Ajuste da equação 13, com energia efetiva linear, para a distribuição dos resíduos apolares.

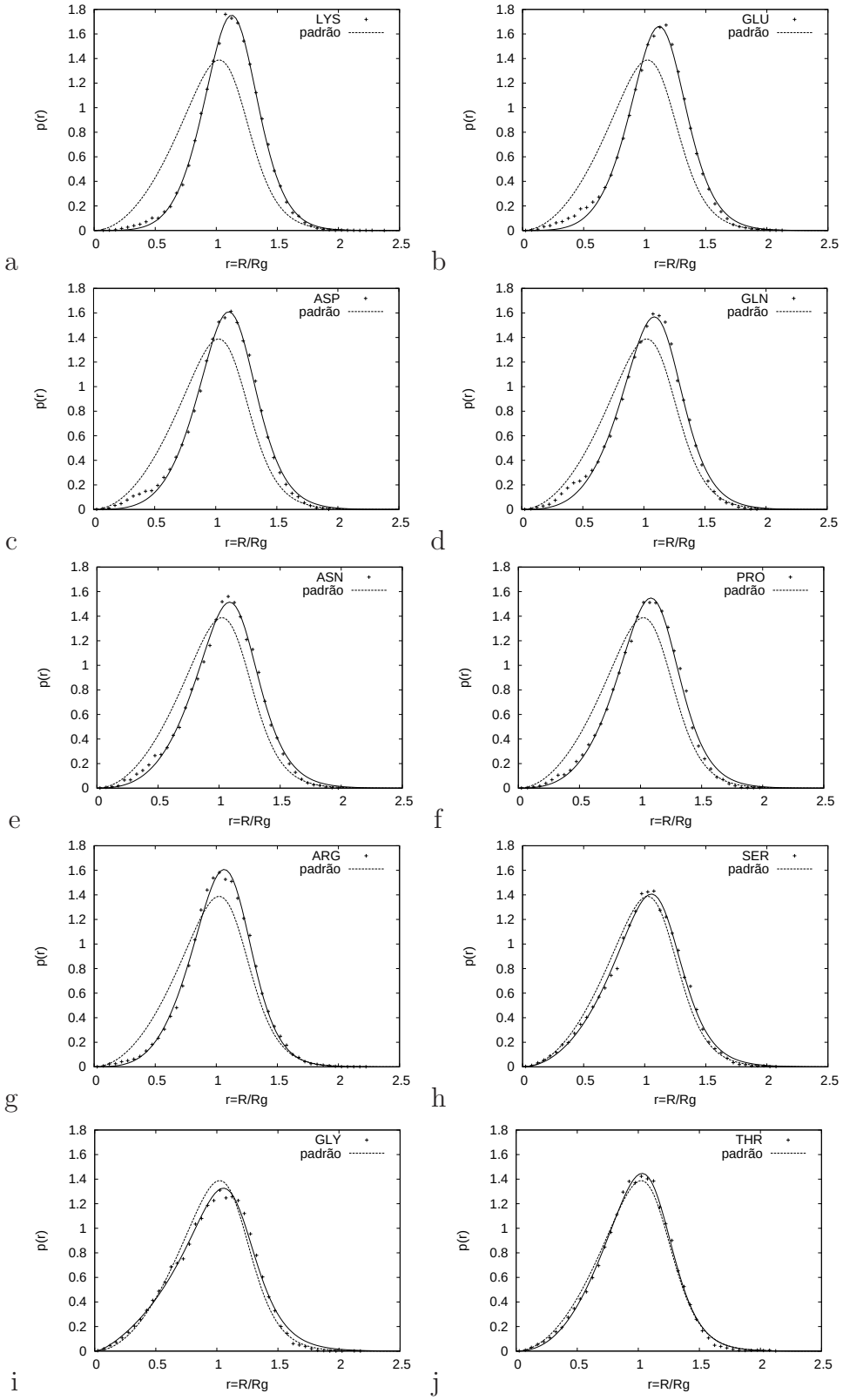


Figura 10: Ajuste da equação 13, com energia efetiva não linear, para a distribuição dos resíduos polares.

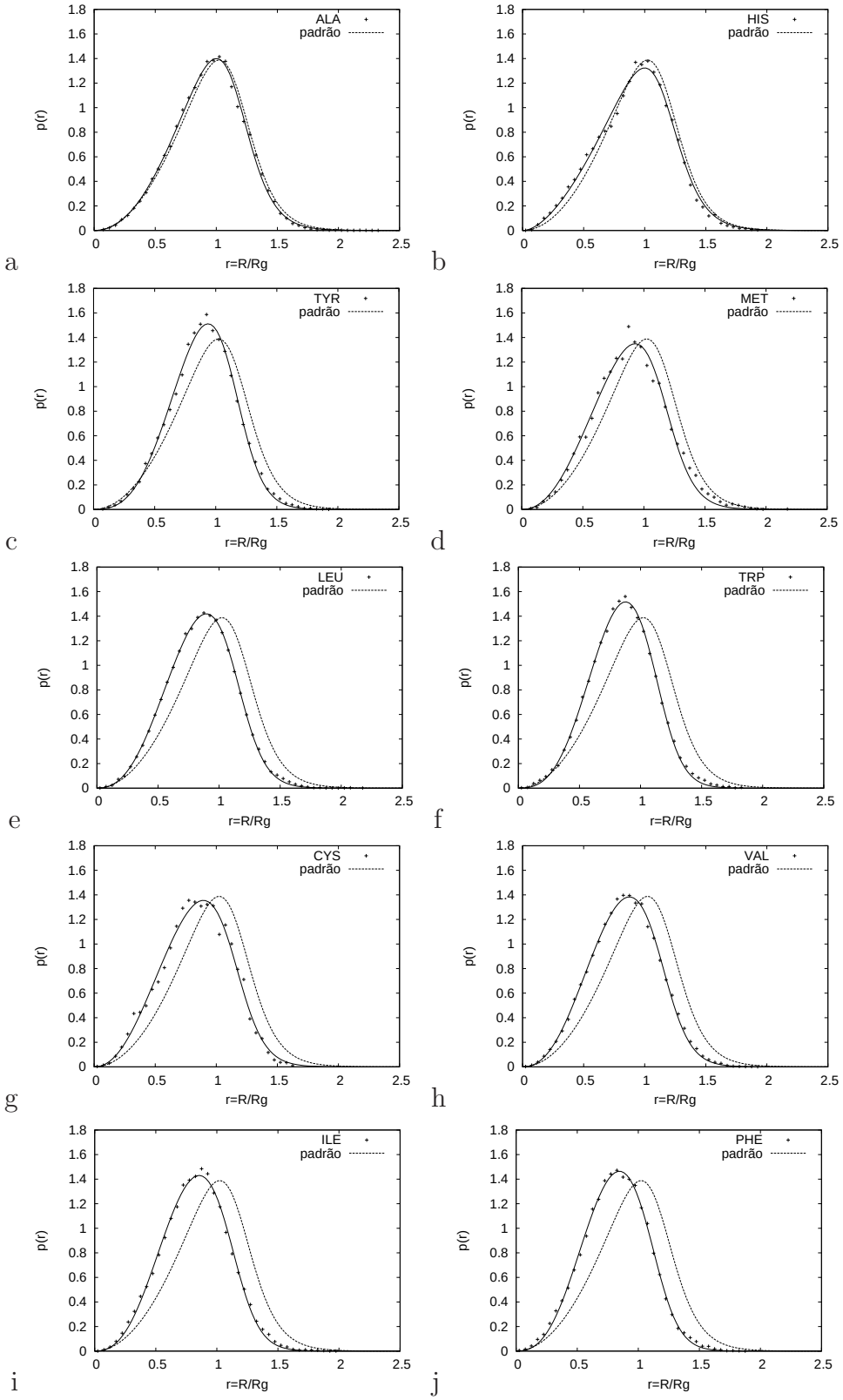


Figura 11: Ajuste da equação 13, com energia efetiva não linear, para a distribuição dos resíduos apolares.

reduzida. Em seus trabalhos, os autores utilizaram um banco de dados de 19 proteínas e classificaram a hidrofobicidade dos resíduos de acordo com sua "distância reduzida média" [31, 32]. A figura 12a mostra a correlação entre as hidrofobicidades h_τ obtida neste trabalho com a "distância reduzida média" obtida no trabalho de Meirovitch. A correlação desses resultados foi de $C = -0,93$. Esse fato indica que quanto maior a hidrofobicidade, menor será a distância média que um resíduo se encontrará do centro da proteína.

Comparamos também h_τ com outras duas classificações obtidas na literatura. A primeira delas foi publicada em 1983. Nessa escala, a hidrofobicidade dos aminoácidos foi obtida experimentalmente pela energia livre de transferência do octanol para a água [18]. A outra escala foi calculada de estruturas de proteínas obtidas por cristalografia, baseada na energia livre de transferência de um resíduo do interior ao exterior de uma proteína. Nessa escala, um resíduo foi considerado "interior" caso a fração de sua superfície acessível fosse menor que 5% [19]. A comparação de h_τ com essas escalas de hidrofobicidades é mostrada na figura 12 b e c e em ambos os casos o coeficiente de correlação de Pearson foi maior que 0,9.

Os primeiros ajustes para G_{res} foram obtidos ao considerar a energia efetiva variando linearmente com r, $\epsilon'_\tau(r) = \epsilon_\tau(r)$. Ao substituir $\epsilon_\tau(r)$ por $\epsilon^*_\tau(r) = h^*_\tau r^{\alpha_\tau}$ uma melhora é observada ao ajuste teórico para a distribuição e a hidrofobicidade de cada resíduo pode ser representada pelo produto $h^*_\tau\alpha_\tau$. A correlação linear entre $h^*_\tau\alpha_\tau$ e h_τ é de 0,99 e pode ser visualizada na figura 12 d.

A definição de energia efetiva variando linearmente com r ajustou bem a distribuição dos resíduos em região próxima ao centro, mas não tão bem nas extremidades. A diferença entre o ajuste e a distribuição aumentava conforme o valor da hidrofobicidade variasse em relação ao padrão (quanto maior o valor absoluto de $\Delta h_\tau = h_\tau - 1$). Assumindo a energia efetiva da forma $\epsilon^*_\tau(r)$, verificou-se uma melhora no ajuste. Esse fato é facilmente visível, tanto no agrupamento G_{HP} (figura 7) quanto para os vinte resíduos (G_{res}) figuras 10 e 11.

De modo inesperado, os parâmetros (h^*_τ e α_τ) mostraram uma informação além da

τ	P_τ	h_τ	μ_τ	$\Delta G_{in\rightarrow out}$	$\Delta G_{oct\rightarrow wat}$
K	0,07	0,79	0,95	-2,00	-1,35
E	0,07	0,80	0,97	-1,09	-0,87
D	0,06	0,83	1,00	-0,72	-1,05
Q	0,04	0,87	1,04	-0,74	-0,30
N	0,05	0,87	1,04	-0,69	-0,82
P	0,04	0,88	1,04	-0,44	0.,98
R	0,06	0,90	1,07	-1,34	-1,37
S	0,05	0,95	1,11	-0,34	-0,05
G	0,04	0,96	1,12	0,06	0
T	0,05	0,98	1,15	-0,26	0,35
A	0,05	1,03	1,19	0,20	0,42
H	0,03	1,05	1,20	0,04	0,18
Y	0,06	1,09	1,25	-0,21	1,31
M	0,02	1,12	1,27	0,71	1,68
L	0,09	1,14	1,29	0,65	2,32
W	0,03	1,15	1,31	0,45	3,07
C	0,01	1,16	1,31	0,67	1,34
V	0,06	1,17	1,32	0,61	1,66
I	0,06	1,18	1,33	0,74	2,46
F	0,06	1,19	1,34	0,67	2,44

Tabela 1: Os tipos de aminoácidos, τ, incluem todos os seus átomos e estão representados pelo código de uma letra. Suas respectivas freqüências no banco de dados, P_τ, hidrofobicidades e potencial químico, h_τ e μ_τ, obtidos dos ajuste das figuras 8 e 9, uma escala de hidrofobicidades, em Kcal/mol, derivada da freqüência de diferentes resíduos na região considerada interior ou exterior de proteínas, definido a partir da área de sua superfície acessível em um banco de proteínas obtidas por cristalografia e uma escala experimental, também em Kcal/mol, obtida da energia livre de transferência de cada resíduo da água para o octanol.

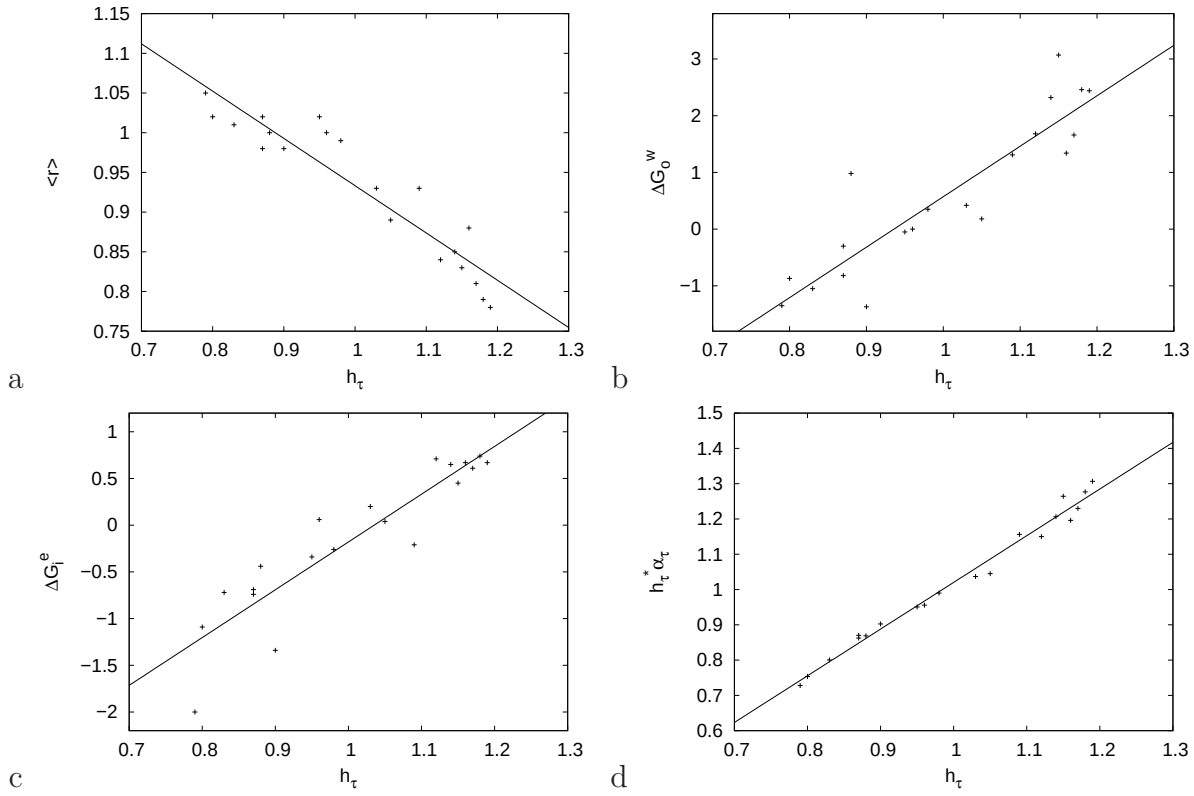


Figura 12: Comparação de h_τ com outras escalas de hidrofobicidade. a) A correlação de h_τ com a distância reduzida média obtida em [31, 32] foi de $-0,93$. Em b) h_τ foi comparada com a energia livre de transferência do octanol para água ΔG_o^w [18] e em c) com a energia livre de transferência do interior ao exterior ΔG_i^e [19], cujos coeficientes de Pearson foram de $0,9$ e $0,91$ respectivamente. A correlação em d) é acima de $0,99$ e compara h_τ com $h_\tau^* \alpha_\tau$.

própria hidrofobicidade dos resíduos. Um gráfico, representando h_τ vs α_τ para os 20 resíduos de aminoácidos, agrupou-os de acordo com suas peculiaridades químicas. Na figura 13, as regiões de isohidrofobicidades C são mostradas por diferentes curvas, do tipo $h_\tau \alpha_\tau = C$, com C variando de $0,7$ a $1,3$, sendo que $C = 1$ (hidrofobicidade neutra) foi mostrado em destaque. Nessa figura, os resíduos hidrofílicos puderam ser distinguidos como carregados (K,R,E,D), ou sem-carga (Q,P,N) pela condição $\alpha_\tau > 1,35$ ou $\alpha_\tau < 1,35$ respectivamente. Para os resíduos hidrofóbicos, $\alpha = 1,35$ serviu de corte para diferenciar os arômáticos (F,W,Y) dos alifáticos (I,L,V,M,C). Os resíduos S,H,G ($\alpha_\tau \leq 1$) e A,T (quase neutros) foram considerados indiferentes a essa classificação.

Os resultados mostrados nesta sessão estão publicados em [34] e uma cópia segue em anexo.

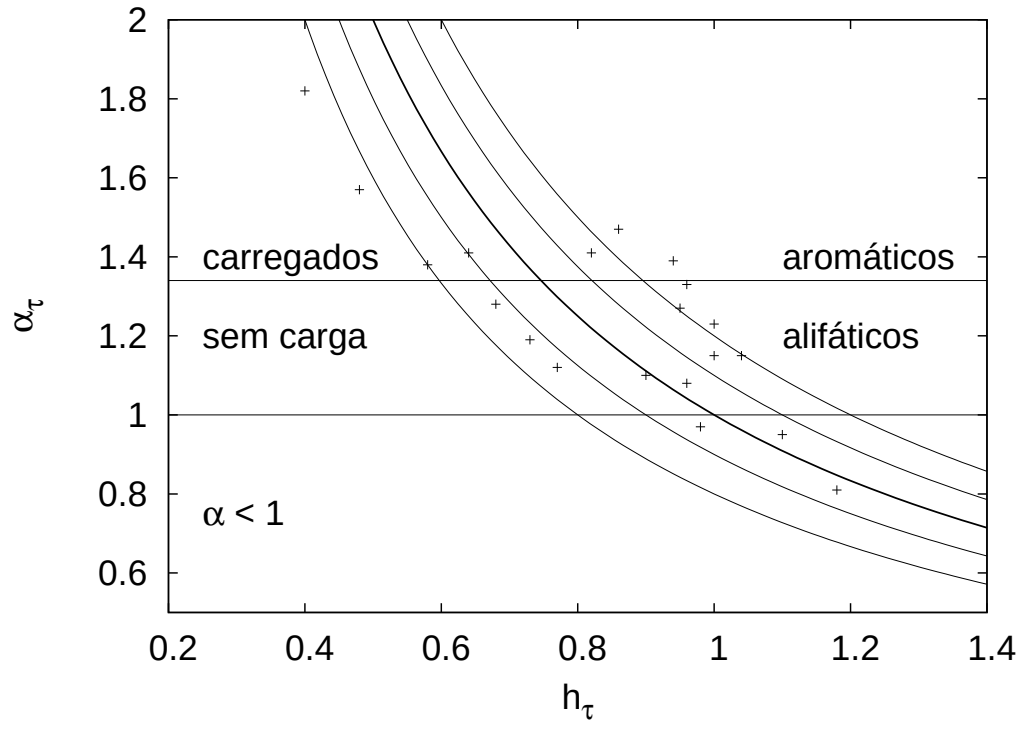


Figura 13: Representação dos resíduos em relação à h_τ e α_τ. As curvas representam regiões de isohidrofobicidades, com destaque para a condição neutra ($h_\tau \alpha_\tau = 1$). Características distintas são agrupadas em diferentes regiões, conforme destacado na figura e descrito no texto.

4 Os enterramentos atômicos contém informação suficiente para dobrar proteínas globulares pequenas

4.1 A informação do enterramento é suficiente para dobrar proteínas globulares pequenas

A busca da predição de distâncias reduzidas (r_i) a partir da seqüência de aminoácidos se torna mais interessante se for confirmada que essa variável contém informação suficiente para dobrar proteínas globulares. Se esse for o caso, podemos definir uma função de energia, com essa variável como parâmetro, capaz de alcançar a estrutura nativa. Chamamos de ED uma função de energia semelhante à função GO [13], porém, utilizamos como referência a distância nativa ideal ao invés de contatos nativos. Para esse caso, definimos a energia E de uma conformação conforme a equação a seguir:

$$E = \sum_i = (|R_i - R_i^*|) \tag{14}$$

onde $R_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$ representa a distância de um átomo i ao centro da conformação e R_i^* a distância "ideal", definida a partir da distância real, encontrada na estrutura nativa e $R_i = R_g \times r_i$. Por definição, o mínimo de energia ocorre quando cada átomo i se encontrar na distância R_i^* e, nesse caso, a energia da estrutura é zero.

Dois tipos de função de energia ED foi utilizada. Na forma mais simples, o efeito de ponte de hidrogênio não é relevante e a distância R_i^* de cada átomo é simplesmente o valor encontrado para a estrutura nativa. No segundo caso, acrescentamos uma distância ideal para os átomos N e O da cadeia peptídica quando, na estrutura nativa, estiverem envolvidos em pontes de hidrogênio. Dessa forma, para esses átomos definimos duas distâncias ideais. Durante uma simulação, a distância ideal para esses átomos é a mesma distância encontrada na estrutura nativa somente se esses átomos estiverem envolvidos em ponte de hidrogênio, caso contrário, sua distância ideal sobe para 14 $\mathring{A}$. O valor 14 $\mathring{A}$ escolhido foi arbitrário e é maior que a maior distância em que foi encontrado ponte de hidrogênio para as proteínas teste. A interação de ponte de hidrogênio foi considerada de

forma inespecífica. Ao considerar essa interação, estamos relevando sua importância física, que é essencial para a compensação entálpica de grupos hidrofílicos que são escondidos na conformação nativa.

Neste trabalho, para a interação entre um par N e O ser considerada ponte de hidrogênio, a distância entre esses dois átomos deve estar contida no intervalo $[2,6:3,2]$ $\mathring{A}$ e a orientação das cadeias serem tal que os ângulos θ_1, θ_2 satisfazem as condições $\theta_1 < 20°$ e $\theta_2 < 60°$. Sendo θ_1 o ângulo entre os vetores $\overrightarrow{NH}$ e $\overrightarrow{HO}$ e θ_2 entre $\overrightarrow{CO}$ e $\overrightarrow{OH}$.

Testamos, por meio de simulação, verificar se a função de energia ED, acrescida de um termo infinito para representar volume excluído, tinha informação suficiente para alcançar a estrutura nativa. Escolhemos para este teste 4 proteínas pequenas e globulares (código pdb: 1E0L, 1IGD, 1ENH e 1ORC), de modo a representar, respectivamente, os 4 grupos de proteínas conhecidos: toda β, α/β, toda α e $\alpha+\beta$, de acordo com a classificação de Levitt and Chothia [35].

O processo de simulação foi feito a partir de um programa criado por Shimada et al. [36] em que apenas os átomos pesados são representados e os ângulos e os comprimentos das ligações covalentes são rígido. A cada passo da simulação, uma mudança de conformação ocorre ao acaso, por uma pequena variação de um dos ângulos diedrais Φ, Ψ e e a aceitação da mudança de conformação segue o critério de Metropolis [37]. Segundo esse critério, uma mudança de conformação de um estado com energia E_1 para outro de energia E_2 é aceito caso $e^{\frac{(E_1-E_2)}{kT}}$ seja maior que um número aleatório entre 0 e 1, onde k representa a constante de Boltzmann e T a temperatura.

Em termos matemático, o valor de T está relacionado com a aceitação de passos com aumento de energia. Quanto maior o valor de T, maior a probabilidade de aceitação de um passo. No algorítmo ultilizado nas simulações, a proteína sofria um processo de "resfriamento", ou seja, T diminui com o número de passos. Esse resfriamento foi representado ao multiplicar a temperatura por um fator de $0,99$ após um certo número de passos. Esse método é conhecido na literatura por "simulated annealing" e permite explorar todo o espaço conformacional em temperaturas altas, direcionando as conformações para estruturas com menor energia à medida que a temperatura abaixa.

As conformações iniciais foram geradas por meio de uma simulação de desenovelamento a uma temperatura extremamente alta ($T = 100000$) durante 10 milhões de passos. Para o enovelamento, as simulações iniciaram de uma temperatura alta, onde praticamente todas as tentativas de movimentos eram aceitas, até atingir uma temperatura de congelamento, a qual a cadeia efetivamente congela. Para 1E0L, na condição sem restrição de pontes de hidrogênios, as simulações decorreram por 2 bilhões de passos, iniciando em uma temperatura $T = 1000$, decrescendo a cada 2 milhões de passos. No caso com restrição de ponte de hidrogênio, as simulações decorreram por 1 bilhão de passos, iniciando à temperatura $T = 200$, decrescendo a cada 1 milhão de passos. Para as outras três proteínas, as simulações decorreram por 2 bilhões de passos, iniciando a uma temperatura $T = 1000$, com decréscimo a cada 2,5 milhões de passos. A similaridade da estrutura alcançada com a estrutura nativa foi calculada a partir do drms dos C_α, definido como

$$drms = \sqrt{\frac{\sum_i^n \sum_{j=i+1}^n (d_{ij} - d_{ij}^*)^2}{N}}$$

onde d_{ij} representa a distância entre os C_α i e j na estrutura obtida e ${d_{ij}^*}^2$ a distância entre os C_α i e j na estrutura nativa.

Para 1E0L, foram realizadas 10 simulações sem considerar a restrição de ponte de hidrogênio e 15 restritas. Ao final, o drms das conformações alcançadas para as simulações sem restrição variou em torno de 2 e 4 $\mathring{A}$, enquanto ao considerar o efeito das pontes de hidrogênio no potencial, as conformações finais mostraram um drms variando de $0,7$ a 2 $\mathring{A}$. Uma correlação forte entre drms e energia foi encontrada para as simulações com potencial restrito (coeficiente de correlação de Pearson $C = 0,76$), mas não foi encontrada para o potencial sem restrição ($C = 0,37$ figura 14a). Além de melhorar a qualidade da conformação alcançada, a restrição de ponte de hidrogênio facilita a formação de estruturas secundárias. Esse resultado pode ser visto comparativamente na figura 15b e d.

Para as outras três estruturas 1IGD, 1ENH e 1ORC foram feitas apenas simulações onde o potencial é dependente da ponte hidrogênio. A correlação entre energia/átomo e drms para a conformação final das 12 trajetórias corridas para 1IGD foi de $0,82$ e a

estrutura de menor energia foi considerada nativa, mostrando um drms de $0,67$ (figura 14b). As diferentes conformações finais foram classificadas visualmente em 4 letras. "A" representa as trajetórias que alcançaram a conformação nativa, "B" conformações de baixa energia e baixo drms, com formação de estruturas secundárias, classificadas como próximas da nativa, "C" são estruturas com energia alta e drms baixo, fato ocorrido por não formar as estruturas secundárias e 'D" representa as estruturas mal dobradas, com a conformação das estruturas secundárias formadas em orientação diferente da nativa. No caso da 1IGD, a folha β formada entre a extremidade N-terminal e C-terminal estava na conformação incorreta anti-paralela. Ainda na figura 14b, nota-se que as conformações finais classificadas como D se localizaram na região de maior energia e drms. A figura 15e-h compara as diferentes conformações alcançadas com a conformação nativa.

Para 1ENH, 11 de 12 trajetórias alcançaram uma conformação semelhante à nativa e a correlação de Pearson entre drms e energia para essas 11 conformações foi $C = 0,78$. Se considerar as 12 trajetórias, a correlação cai para $C = 0,47$. A trajetória que alcançou a conformação diferente da nativa foi classificada como "D". Nessa estrutura, todas as hélices foram formadas, porém elas se organizaram como uma imagem especular da estrutura nativa (figura 15l), fato que resultou em um drms alto e energia intermediária (figura 14c). Para 1ORC, a relação entre drms e energia é mostrada na figura 14d, cujo coeficiente de Peason foi de $C = 0,56$. Entre as 10 trajetórias, 4 foram classificadas como mal dobradas. Nessas trajetórias, a conformação final alcançada formava todas as estruturas secundárias, porém a região de α hélices estava com uma organização diferente da nativa. Um exemplo dessas conformações é mostrado na figura 15p. Se desconsiderarmos essas 4 conformações mal dobradas, o coeficiente de correlação entre drms e energia sobe para $C = 0,92$.

Em sua revisão de 1990 [1], Dill descreve o efeito hidrofóbico como o principal fator responsável pela estabilidade da estrutura nativa de uma proteína, mas ressalta a ponte de hidrogênio como importante para a organização interna. O resultado obtido para 1E0L corroboram essa idéia. A informação do enterramento dos aminoácidos foi suficiente para alcançar a estrutura nativa, na qual um drms abaixo de 2 $\mathring{A}$ foi obtido para a melhor

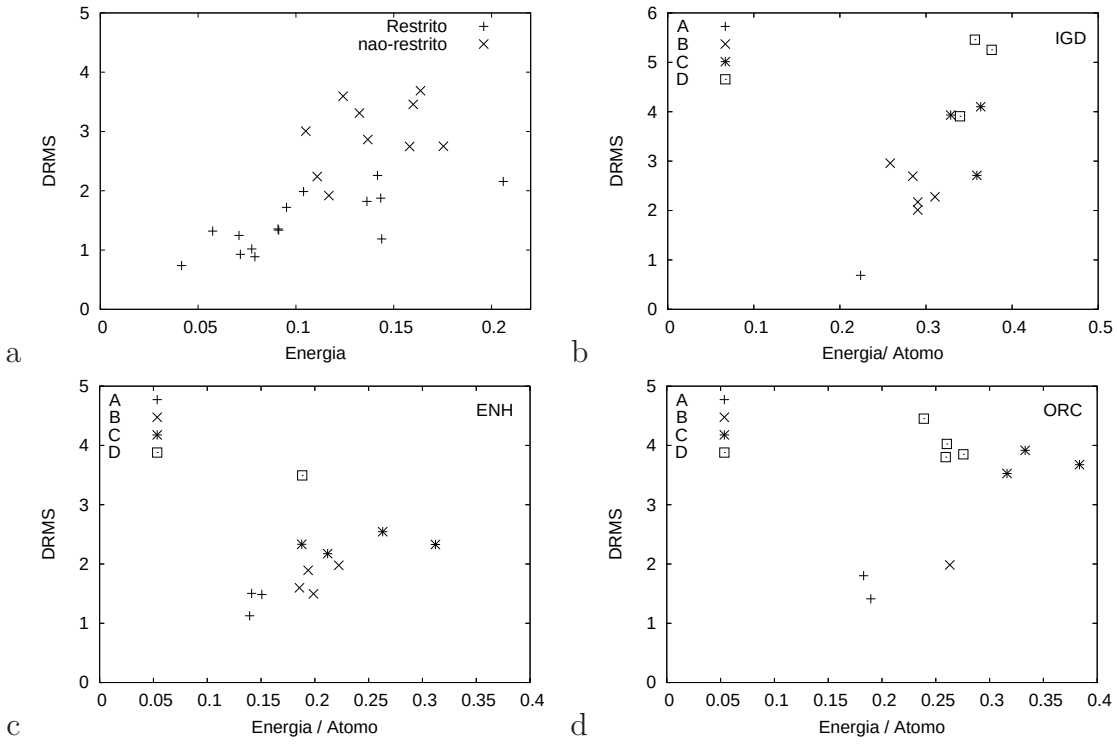


Figura 14: Representação da correlação entre drms e energia para as conformações finais obtidas em diferentes trajetórias de 4 diferentes proteínas utilizando o potencial de energia ED . a) Comparamos as conformações de 1E0L obtidas para potenciais restrito ou não à ponte de hidrogênio. As simulações que discriminavam ponte de hidrogênio congelaram em conformações com energia e drms menores. b),c),d) drms vs energia para conformações finais obtidas para as proteínas 1IGD,1ENH e 1ORC, respectivamente. 4 letras foram utilizadas para comparar as estruturas finais. "A" são estruturas classificadas como nativa, "B" representa conformações quase nativa, de baixa energia e drms, "C" representa estruturas de topologia nativa, com energia alta, por não terem formado as estruturas secundárias porém drms baixo e "D" foi utilizado para representar as estruturas mal dobradas, com topologia diferente da nativa (maiores detalhes na figura 15).

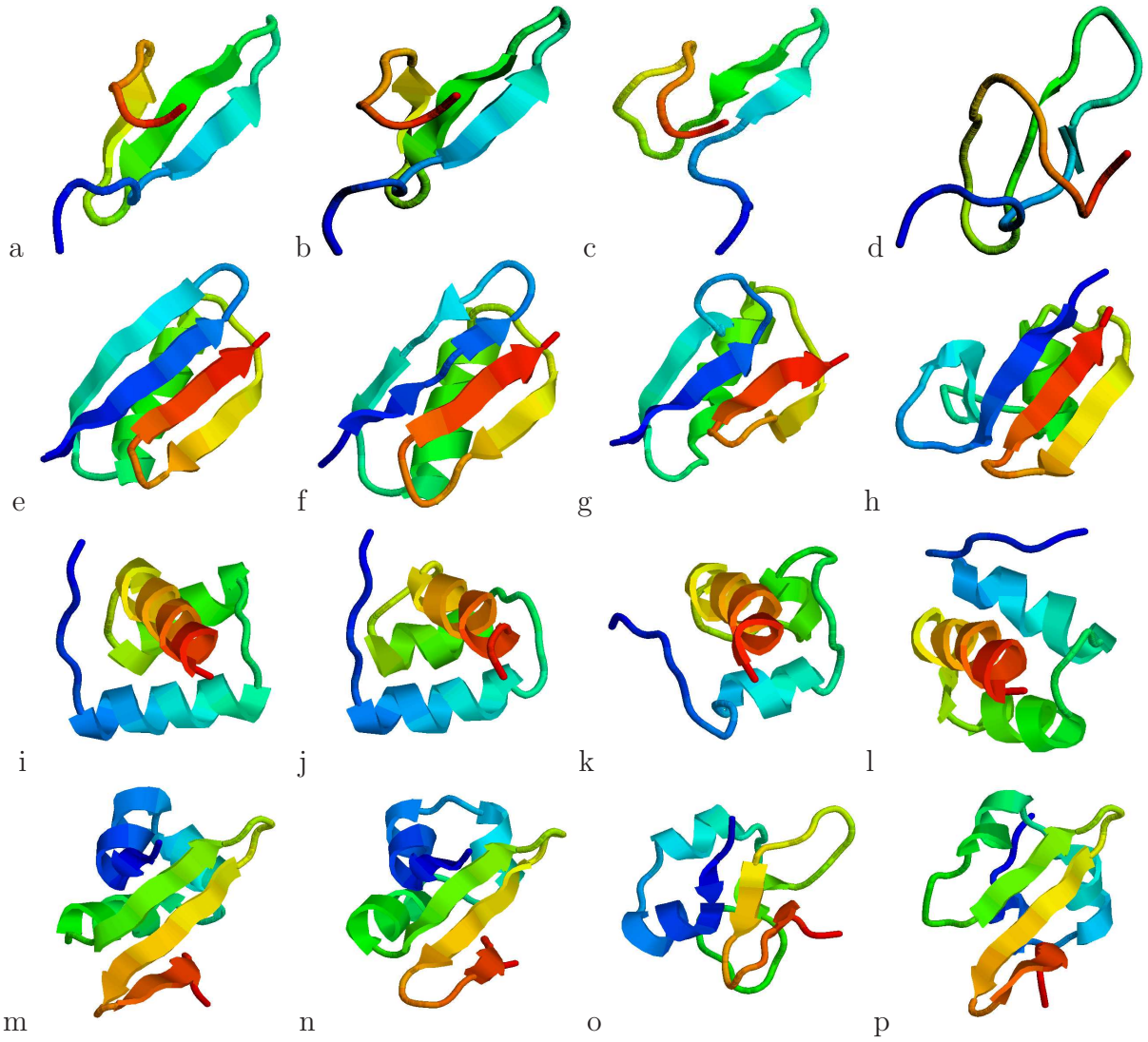


Figura 15: A primeira coluna representa a estrutura nativa e em cada linha está representada uma das proteínas discutidas no texto. 1E0L, 1IGD, 1ENH, 1ORC estão representados, respectivamente, na primeira, segunda, terceira e quarta linha. Em b) mostramos a melhor e c) a pior conformação alcançada utilizando o potencial restrito à ponte de hidrogênio. Em d) é mostrada a melhor conformação obtida para o caso do potencial não restrito. Para as outras proteínas, a 2ª coluna representa a melhor conformação obtida, classificada como "A". A terceira coluna representa uma estrutura classificada como "B" para 1IGD e classificada como "C" para 1ENH e 1ORC. Na quarta coluna, mostramos uma conformação considerada mal dobrada, classificada como "D".

simulação. Porém, considerando o efeito da ponte de hidrogênio, o drms das estruturas finais diminui (figura 14) e na melhor trajetória, o drms final ficou em torno de $0,7\ \mathring{A}$. Além disso, a relevância da ponte de hidrogênio mostrou ser importante para formação das estruturas secundárias (figura 15).

Para que a distância ao centro geométrico de cada átomo seja suficientemente informativa é essencial que não exista nenhuma conformação diferente da nativa com energia zero (mínimo por definição). Neste trabalho, nenhuma das simulações alcançou uma estrutura com o mínimo de energia possível para o potencial, porém o coeficiente de correlação mostra que o drms e a energia tendem a decrescer juntamente, sugerindo que a minimização de energia tende a convergir para a estrutura nativa e corroborando a hipótese de que a distância reduzida contém informação suficiente para alcançar a estrutura de uma proteína.

4.2 A energia efetiva pode ser utilizada com uma função de energia

Pode-se relacionar o caráter termodinâmico da equação 13 à distribuição dos átomos em relação à distância reduzida. Na distribuição de Fermi-Dirac, $\epsilon'_\tau(r) - \mu_\tau$ está relacionado com o peso de Boltzmann. Dessa forma, esse parâmetro é referente à um potencial estatístico e pode ser utilizado para calcular a "energia" de uma estrutura a partir da contribuição de cada átomo. A contribuição energética de cada átomo foi calculada para o caso em que $\epsilon'_\tau(r) = \epsilon^*_\tau(r)$, ou seja, a variação da energia efetiva não é linear em relação a r.

Uma função de energia ideal deve ser capaz de distinguir a estrutura nativa de todas estruturas mal dobradas. Um termo da equação 13 tem caráter de uma função de energia, que diferencia a posição preferencial de cada átomo. A partir do ajuste de $p_\tau(r)$, mostrado na equação 13, definimos $E_\tau(r)$ como função de energia característica para cada grupo atômico τ, conforme a equação a seguir:

$$E_\tau(r) = \Delta\epsilon^*_\tau(r) + \Delta\mu_\tau \tag{15}$$

onde

$$\Delta\epsilon^*_\tau(r) = \epsilon^*_\tau(r) - 1$$

e

$$\Delta\mu_\tau = \mu_\tau - \mu.$$

$E_\tau(r_0) \leq 0$ significa que o átomo τ é mais provável de ser encontrado na posição r_0 que um átomo aleatório. $E_\tau(r)$ é mostrado para alguns grupos atômicos representativos na figura 16. Em 16a, a função de energia diferencia os resíduos apenas em hidrofílicos e hidrofóbicos. Encontramos para os resíduos hidrofóbicos que $E_\tau(r)$ é negativo quando r pequeno. Para os resíduos hidrofílicos $E_\tau(r)$ é negativo em r grande. Em 16b compara-se $E_\tau(r)$ para Fenilalanina (F) e Lisina (K), segundo nossos dados, o resíduo mais hidrofóbico e mais hidrofílico, respectivamente. Nessa figura, a função de energia é mostrada para os átomos agrupados em 20 letras, conforme G_{res} e o efeito de diferenciar os átomos de cada resíduo. O grupo que diferencia cada átomo foi chamado de G_{at}. Pelo gráfico, percebe-se a diferença na função de energia ao especificar o tipo de átomo de cada resíduo.

O efeito da ponte de hidrogênio é mostrado para a Alanina, na figura 16c. Chamamos de G_{ph} o grupo no qual os átomos da cadeia principal são diferenciados por estarem envolvidos ou não em ponte de hidrogênio. Consideramos que C_α e C estão envolvido em ponte de hidrogênio caso N e O, do mesmo resíduo, respectivamente, façam ponte de hidrogênio. A Alanina é o resíduo com comportamento mais próximo do neutro. Ao se considerar todos os átomos da Alanina como um grupo (em G_{res}), o mínimo da função de energia é próximo do centro, tendo pouca variação em relação ao padrão. A função de energia representativa apenas do O da Alanina mostra uma preferência um pouco maior para raios maiores. Ao se levar em consideração o efeito de ponte de hidrogênio, o O livre diminui a energia para r expostos, enquanto o O envolvido em ponte de hidrogênio tem a energia reduzida para r enterrados.

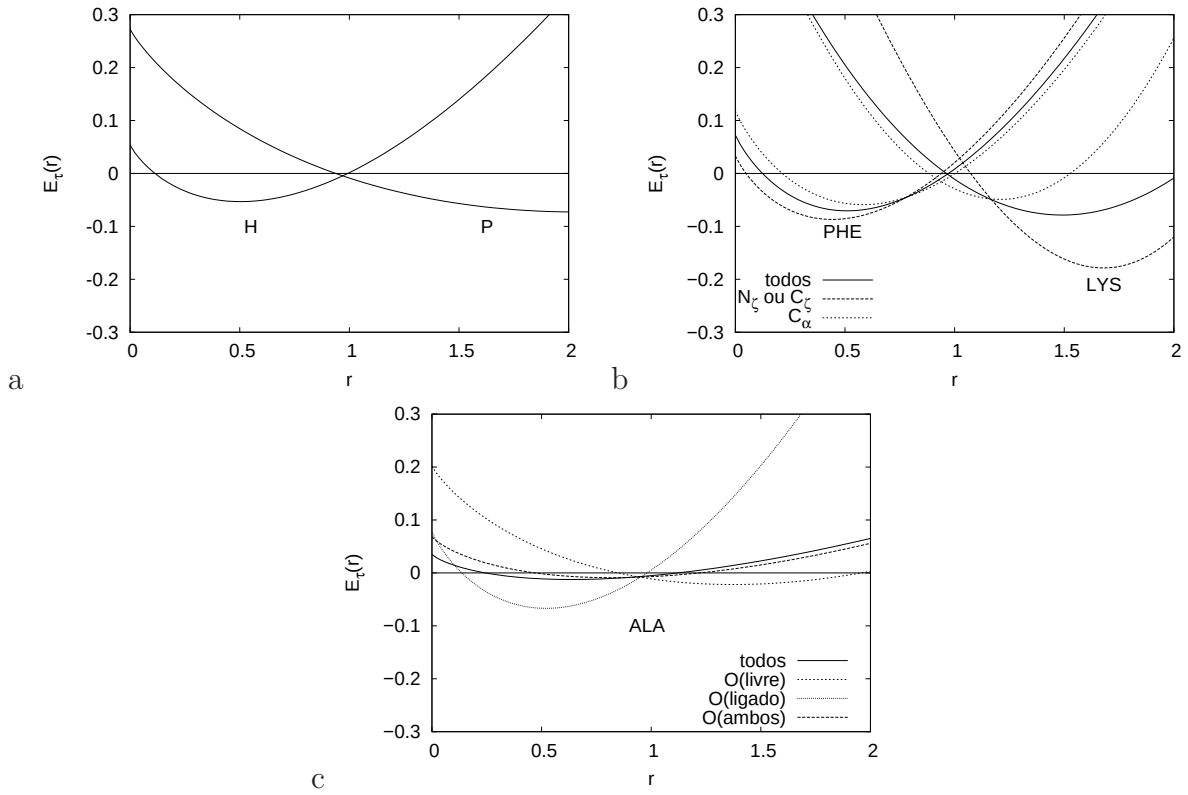


Figura 16: Representação do potencial estatístico em relação a r para diferentes grupos atômicos. a) A energia dos resíduos hidrofóbicos (H) é menor para regiões mais enterradas (r pequeno, no intervalo $[0:1]$), enquanto a energia para os resíduos hidrofílicos (P) é menor para a região mais exposta ao solvente (r grande, no intervalo $[1:2]$). b) O mesmo comportamento é visto para o resíduo mais hidrofóbico (PHE) e o resíduo mais hidrofílico (LYS) se comportam de maneira semelhante. A especificação do tipo de átomo torna a energia mais específica. c) Átomos envolvidos em ponte de hidrogênio "preferem" regiões enterradas e átomos livres de ponte de hidrogênio preferem regiões expostas.

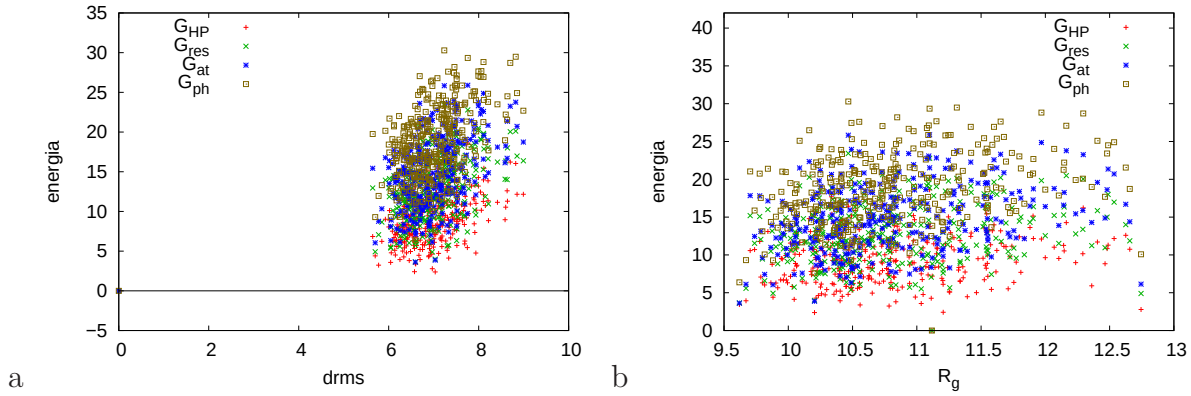


Figura 17: a) A segregação energética aumenta, ao se usar um potencial mais específico (de G_{HP} para G_{ph}). b) Não há correlação entre o potencial energético com o raio de giro da proteína.

Realizamos o teste de qualidade da função de energia para a proteína de código pdb 1ORC. Essa proteína é globular e pequena, com 61 resíduos de aminoácidos. Usamos como teste conformações compactas geradas ao acaso por meio de simulações de Monte Carlo baseada no programa desenvolvido por Shimada et al. [36]. Nesse caso, definimos a função de energia de forma que a energia de uma conformação é inversamente proporcional ao número de contatos. Das conformações obtidas, foram selecionadas apenas aquelas com raio de giro próximo ao da nativa, pertencendo ao intervalo [9:13] $\mathring{A}$ (R_g da nativa $\approx 11,1$ $\mathring{A}$), um total de 319 conformações.

A partir da função de energia definida na equação 15, calculamos a energia E_e de cada estrutura. Utilizando E_n, a energia da estrutura nativa, como referência definimos $\Delta E_e = E_e - E_n$. Os gráficos de ΔE_e versus drms e rg para para cada estrutura compacta não-nativa gerada são mostrados na figura 17. Nesses gráficos, observa-se que todas as amostras de conformações globulares obtidas ao acaso tiveram uma energia maior que a estrutura nativa. Ainda nesses gráficos, nota-se que, quanto maior a especificidade do grupo atômico, maior a profundidade da segregação energética da estrutura nativa.

A profundidade da segregação energética pode ser quantificada a partir do seu valor-Z. O valor-Z Z_e, da energia de uma estrutura e, está relacionado à "distância", em unidades

de desvio padrão, que sua energia encontra-se em relação à energia média ($Z_e = \frac{E_e - \overline{E}}{\sigma}$). Obtivemos para Z_n, o valor-Z da estrutura nativa, os valores $-3,0$, $-3,1$, $-3,3$ e $-4,1$ ao diferenciar os grupos atômicos de acordo com G_{HP}, G_{res}, G_{at} e G_{ph}, respectivamente. O aumento da segregação pode ser visualizado na figura 18a, em que a densidade de "decoys" é plotada em relação a $\Delta Z_e = Z_e - Z_n$. A figura mostrou que quanto mais específica foi a função de energia, maior a distância do valor-Z das estruturas ao acaso ao se comparar com o valor-Z da estrutura nativa.

Tanto o potencial GO [38] quanto o potencial ED contém informação suficiente para dobrar uma proteína . O potencial GO é estabelecido a partir dos contatos da estrutura nativa. Nesse potencial, somente contatos nativos contribuem e negativamente, para a energia de uma estrutura. O potencial ED foi descrito na equação 14. Pela figura 18b, podemos concluir que o potencial GO possui mais informação que o necessário ($Z_n \approx 50$) para dobrar uma proteína. O valor-Z do potencial ED, $Z_n \approx 10$, serve de referência como um limite de informação a partir do qual uma função de energia torna-se suficientemente informativa para alcançar a estrutura nativa. De um modo otimista, podemos considerar o valor $Z_n \approx 7$ estimado por Bowie et al. [39] como suficiente para alcançar a estrutura nativa. Na melhor das hipóteses obtidas, diferenciando átomos conforme o grupo G_{ph}, o potencial encontrado resultou no valor-Z para a estrutura nativa em torno de $4,1$.

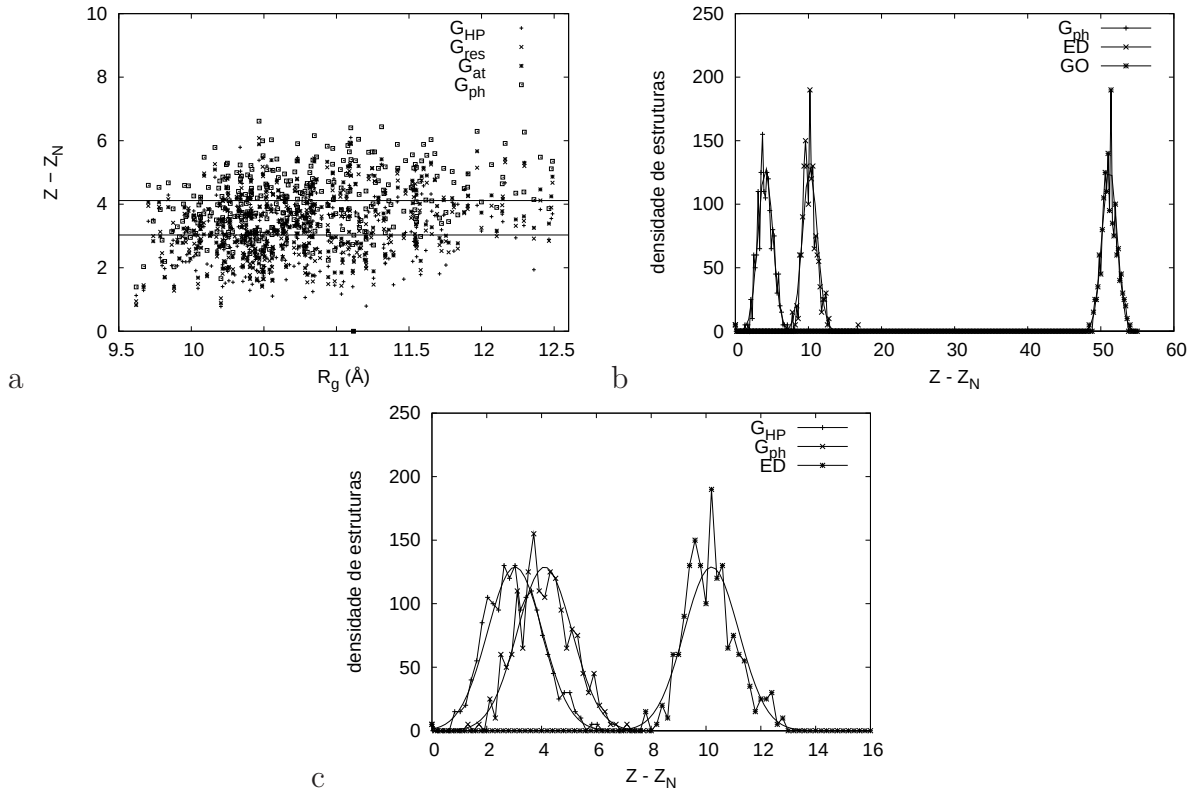


Figura 18: O aumento da segregação energética com o aumento da especificidade da função de energia é mostrado em a). As linhas retas representam $\overline{\Delta Z}$ para G_{HP} e G_{ph}. Não há relação entre R_g e ΔZ_e. O potencial GO parece ter mais informação que a necessária. A distribuição do valor-Z do potencial ED serve como referência da informação necessária para se dobrar um proteína b). Em c) compara o ganho de informação ao passar de G_{HP} para G_{ph}, almejando ter informação suficiente para atingir uma distribuição semelhante à de ED.

5 Análise da quantidade de informação da seqüência de aminoácidos e de enterramentos

5.1 Estimativa da quantidade de informação que pode ser extraída da seqüência de aminoácidos para a predição de enterramentos atômicos

A teoria da informação desenvolvida por Shannon tem suas raízes ligadas à explicação do fenômeno de comunicação por telégrafos. De acordo com essa teoria, uma "fonte", tal como um livro ou um palestrante, pode transmitir qualquer um entre uma vasto conjunto de sinais. Quanto maior a possibilidade de sinais, maior será a dúvida inicial para o receptor. Uma mensagem que representa uma possibilidade em 10 transmite uma menor informação comparada a uma mensagem em mil possíveis, uma vez que a diminuição da dúvida do receptor ao receber a mensagem é menor. Chamamos de entropia a medida dessa incerteza. A definição de entropia pela teoria da informação segue a mesma forma matemática daquela definida para a termodinâmica estatística e seu valor está relacionado à quantidade de possibilidades de mensagem que pode ser enviada por uma fonte ou a dúvida de um receptor. Quanto mais soubermos sobre a mensagem que uma fonte produzirá, menor será sua entropia [40]. No estudo de proteínas, a seqüência de aminoácidos pode ser considerada uma mensagem que é traduzida, pelo processo de enovelamento, para outra mensagem, escrita pelos estados conformacionais da estrutura 3D. De acordo com essa análise, de alguma forma, a informação da estrutura de uma proteína estaria codificada na seqüência de aminoácidos.

Se a seqüência primária representa uma mensagem, então os resíduo de aminoácidos representam a sua subunidade básica para formar essa mensagem, como letras que compõem as palavras e formam frases. De acordo com a teoria de Shannon, a entropia de uma mensagem pode ser estimada a partir da probabilidade de se encontrar as letras que a compõe [41]. O caso mais simples é considerar que as letras se combinam de forma independente. Se esse for o caso, a entropia H, em bits por letra, de uma mensagem composta

por n tipos de letra pode ser estimada simplesmente a partir da probabilidade P_i de cada uma das letras i, conforme a equação abaixo:

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{16}$$

Para um alfabeto de 20 letras, a entropia máxima possível é de $\log_2(20) \approx 4,32\, bits/aminoácido$. Os 20 resíduos de aminoácidos não são equiprováveis e com o banco de dados de proteínas globulares, descrito na seção 3.1, encontramos que a entropia de resíduos de aminoácidos estaria em torno de $4,20\, bits/aminoácido$.

Porém, já foi descrito na literatura que a combinação dos resíduos ao longo da seqüência primária não é independente. Em um trabalho de 2004, Betancourt et al. verificou a existência de "triplets" preferenciais e "triplets" raros para seqüências de proteínas [42]. Dessa forma, o valor de $4,20\, bits/aminoácido$ se torna um limite superior para a entropia da seqüência de aminoácidos. A estimativa da entropia torna-se cada vez mais exata ao se considerar que a combinação das letras não é independente. A equação 17 define H_m, a entropia de ordem m, com n possíveis tipos de letras e $P(x_{i1}, x_{i2}, ..., x_{im})$ a probabilidade da seqüência de letras $x_{i1}, x_{i2}, ..., x_{im}$ em um bloco contendo m letras.

$$H_m = -\frac{\sum_{i1=1}^{n}\sum_{i2=1}^{n}\cdots\sum_{im=1}^{n} P(x_{i1}, x_{i2}, ..., x_{im}) \log_2 P(x_{i1}, x_{i2}, ..., x_{im})}{m} \tag{17}$$

e limite $\lim_{m\to\infty} H_m$ representa a entropia "real" da seqüência de aminoácidos.

Os dados calculados a partir da equação 17 são mostrados na figura 19 como entropia, em $bits/aminoácido$, em função do tamanho do bloco de aminoácidos considerado. Para a construção dos gráficos, consideramos que a probabilidade de uma seqüência de resíduos equivale à freqüência relativa encontrada no banco de dados. Nossos dados são limitados estatisticamente e, por esse motivo, representamos no gráfico a curva de saturação, $H = \log_2(n)/m$, ou seja, quando cada bloco de resíduos aparece uma única vez, onde n é o número de possibilidades de blocos.

A entropia mostrou um decréscimo muito pequeno ao considerar blocos de resíduos enquanto o banco de dados não estava saturado (figura 19). Antes da saturação do banco de dados, com blocos de 4 resíduos de aminoácidos, a entropia encontrada estava em

tamanho do bloco	entropia	valor saturado
1	4.19665	17.0358
2	4.19264	8.5179
3	4.17517	5.6786
4	3.9806	4.25895
5	3.38718	3.40716
6	2.83807	2.8393
7	2.43349	2.43369
8	2.12936	2.12948
9	1.89278	1.89287
10	1.7035	1.70358

Tabela 2: entropia da seqüência primária, em *bits/aminoácido*, em função do tamanho do bloco.

torno de 4 *bits/aminoácido*. A tabela 2 mostra esses resultados na forma de valores absolutos. Para assegurar a idéia de informação constante em relação ao tamanho do bloco, representamos na mesma figura a entropia ao agrupar resíduos de aminoácidos com características semelhantes. Esse agrupamento reduz a quantidade de letras para 6, 5 ou 2 e permite amostras estatística suficiente para blocos de resíduos maiores. Em todos os casos, verificamos que a entropia/resíduo é próxima de uma constante, seguida de uma queda acelerada nas regiões próximas da saturação.

Em 1996, Strait and Dewey estimaram, a partir de um banco de dados de 190 proteínas, que a entropia da seqüência de aminoácidos estaria em torno de $2,5$ *bits/aminoácido* [43]. Porém, nesse trabalho, o autor não comenta sobre o limite de saturação do banco de dados. Baseado nos resultados encontrados, é mais razoável, entretanto, considerar essa entropia por volta de 4 *bits/aminoácido*. Desse modo, o valor de 4 *bits/aminoácido* serve como referência para a quantidade de informação possível de ser transmitida para a estrutura a partir da seqüência de aminoácidos.

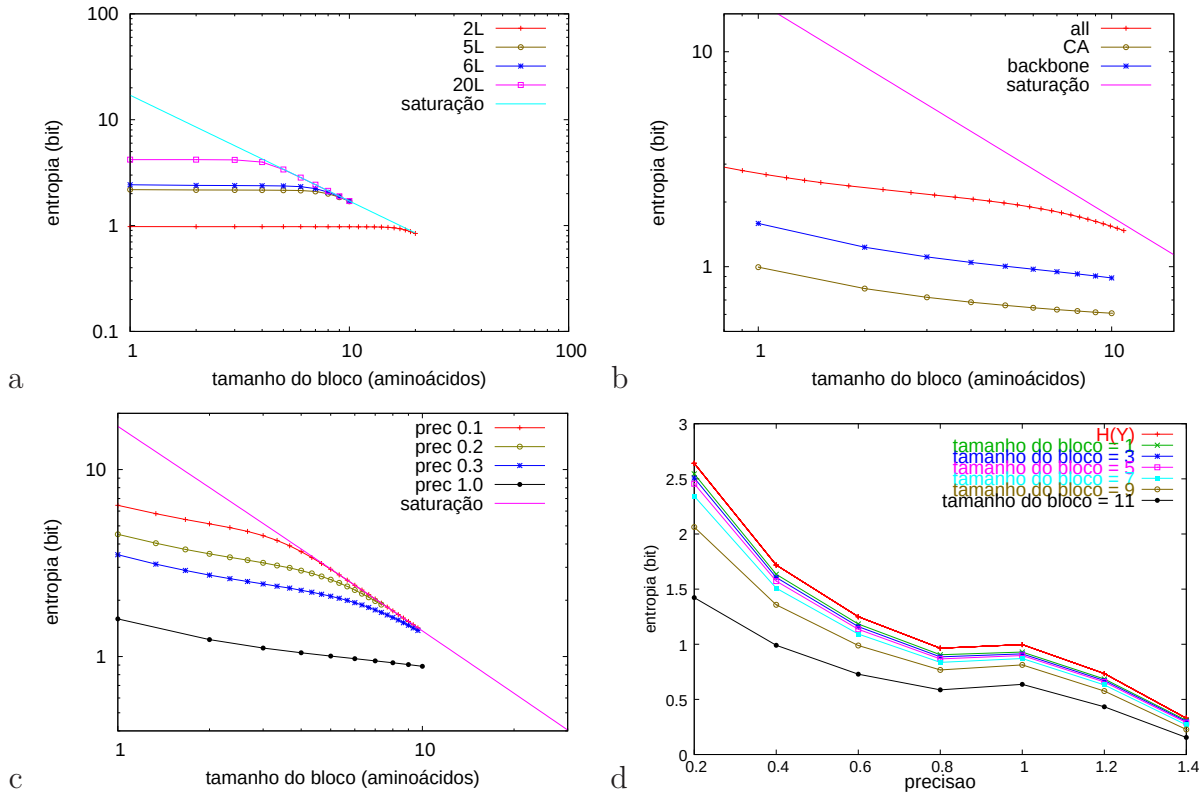


Figura 19: Análise de informação. Em a) constatou-se que a entropia da seqüência de aminoácidos se mantém constante até as proximidades da curva de saturação, ao considerar blocos de aminoácidos. A saturação do banco de dados pode ser prolongada ao agrupar os resíduos por similaridades. Nessas condições, a entropia continua independente da seqüência. Em relação à distância reduzida, a entropia aumenta ao se considerar mais átomos na análise da informação (b) ou ao se aumentar a sua precisão (c). O gráfico em d permite analisar a diminuição da entropia para os enterramentos ao se saber a seqüência de aminoácidos.

Para ser prevista, a entropia das estruturas deve ser inferior àquela pertencente à seqüência de aminoácidos. Neste trabalho, a distância reduzida dos átomos, ou enterramento, refere-se à informação estrutural a ser procurada. Ao discretizar o enterramento dos átomos pela precisão 1 e considerando apenas o C_α dos átomos, encontramos entropia de primeira ordem próxima de 1 $bit/aminoácido$, decrescendo para uma entropia de ordem 10 com valor de $0,6$ $bits/aminoácido$ (figura 19b), valor 3 vezes abaixo da saturação ($\approx 1,7 bits/aminoácido$).

Porém, a entropia dos enterramentos não dependente apenas da seqüência de aminoácidos e, quanto mais átomos foram considerados, mais precisa se torna a entropia. Ao considerar os átomos C, C_α e N, o cálculo do entropia do enterramento partir da equação 17 fornecerá um resultado em $bits/átomo$. Pelo fato de cada aminoácido contribuir com a entropia de 3 átomos, a transformação para $bits/aminoácido$ ocorre ao multiplicar a entropia em $bits/átomo$ por 3. Essa idéia se aplica se mais átomos forem considerados. Se cada resíduo contiver em média $5,4$ átomos, essa transformação ocorrerá por um fator de $5,4$.

Observando a figura 19b, verifica-se que ao se considerar também os átomos C e N, a entropia aumenta em cerca de 50% em relação a quando era considerado apenas o C_α. Esse aumento é muito menor que 200%, o valor esperado caso o enterramento desses átomos fossem independente de C_α. Quando consideramos todos os átomos, incluindo aqueles da cadeia lateral, o aumento da entropia foi um pouco maior que o dobro daquela encontrada apenas para o C_α. À medida em que a precisão da discretização do enterramento foi melhorada, verificou-se um aumento na entropia estrutural. Uma precisão de $0,1$ para o enterramento possui uma informação de cerca de $4 bits/aminoácido$ e, de acordo com essa análise, se torna um limite superior da precisão do enterramento possível de ser codificada a partir da seqüência.

É interessante notar que, ao contrário do observado para a seqüência de aminoácidos, o enterramento de um átomo mostrou bastante dependência do enterramento de átomo vizinho. Esse fato é ilustrado pela queda da entropia ao se considerar "blocos" de aminoácidos e é justificado pelo fato dos átomos estarem ligados covalentemente ao longo da seqüência,

de forma que a "liberdade" de enterramento se torna limitada pela distância entre as ligações.

Existe uma dependência entre os enterramentos preferenciais dos átomos e a seqüência de aminoácidos. Como vimos na sessão 4.3, o enterramento preferencial dos átomos está relacionado com a hidrofobicidade de seus resíduos. Para o caso de considerarmos o enterramento dos átomos discreto, podemos estimar a quantidade de informação $I(x;y)$ a ser extraída da seqüência de aminoácidos para a predição do enterramento, conforme a equação 18.

$$I(X;Y) = \sum_{x}\sum_{y} P(x,y)\log_2 \frac{P(y|x)}{P(y)} \tag{18}$$

onde I(X;Y) é chamado de transinformação entre as variáveis X, Y e P(x) representa a probabilidade de $X = x$. Vale a pena ressaltar que matematicamente, I(X;Y) pode ser representado pela diferença $I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$, em que $H(X)$ é o mesmo definido a partir das equações 16 ou 17.

A figura 19d mostra a "dúvida" para a predição do enterramento, de primeira ordem, em relação à informação, de ordem $2k+1$, referente à seqüência de aminoácidos. Quando não se possui nenhuma informação sobre a seqüência de aminoácidos, a dúvida é máxima e está representada na figura por $H(Y)$. À medida que aumenta a informação sobre a seqüência, essa dúvida decresce. A transinformação I(X;Y) é representada pela diferença entre $H(Y)$ por $H(Y|X)$. $H(Y|X)$ representa a entropia do enterramento ao saber a seqüência X. Quando $bloco = 1$ consideramos apenas a influência individual de cada átomo para o enterramento. Em $bloco = 2k+1$, consideramos uma janela com a informação de $2k+1$ aminoácidos, onde o enterramento considerado pertence ao átomo central. Os resíduos vizinhos foram agrupados como polares ou apolares. Um exemplo de seqüência é $HHP(ALA - C_\alpha)HHP$. Essa seqüência representa uma janela de 7 resíduos, onde um C_α de uma Alanina está cercado por resíduos polares (representados por P) e apolares (representados por H).

Pelo gráfico 19d, percebe-se a queda da dúvida (entropia) do enterramento ao aumen-

tar a informação referente à seqüência de aminoácidos para diferentes precisão de corte. Quanto mais precisa for a busca pelo enterramento, maior será essa dúvida. É interessante ressaltar que a dúvida sempre diminuiu ao piorar a precisão, exceto entre a precisão de $0,8$ para $1,0$. Essa exceção ocorreu devido ao fato do pico de probabilidade de distribuição de enterramentos ocorrer perto de $1,0$. Ao se agrupar em precisão de $0,8$, cria-se um super-grupo compreendendo os enterramentos percentencentes ao intervalo $[0,8 : 1,6]$ que conseqüentemente diminui a entropia. Quando o tamanho do bloco é igual a 11, o banco de dados está saturado e a queda da entropia deixa de ser significativa.

Na seção 3.3, obtivemos uma função contínua para a densidade de probabilidade de um átomo em relação ao enterramento. Para o caso contínuo, uma adaptação da equação 18 permite a quantificação da transinformação. O cálculo para a transinformação a partir de distribuições contínuas é mostrado abaixo, na equação 19.

$$I = \sum_{\tau} P_{\tau} I_{\tau} \tag{19}$$

onde I representa a transinformação, P_τ a probabilidade de encontrar um átomo τ e I_τ a transinformação relativa de um átomo τ. I_τ é mostrado abaixo, onde $p(r)$ e $p_\tau(r)$ seguem as definições de acordo com as equações 7 e 13.

$$I_{\tau} = \int_0^{\infty} \log_2 \frac{p_{\tau}(r)}{p(r)} dr \tag{20}$$

Considerando o caso em que a energia efetiva não necessariamente varia linearmente com r, a transinformação para um grupo atômico τ depende apenas dos parâmetros h_τ e α_τ. À medida que esses dois parâmetros se afastam do padrão ($h_\tau = \alpha_\tau = 1$), ou seja, quanto maior $|h_\tau - 1|$ e $|\alpha_\tau - 1|$, mais informativo será sua função de energia. A figura 20 representa curvas de nível da transinformação em função de h_τ e α_τ. Nessa figura está ilustrada regiões de isohidrofobicidades. A transinformação varia pouco em regiões de mesma hidrofobicidade.

A partir da definição mostrada na equação 19, encontramos $I = 0,096;\ 0,109;\ 0,135;\ 0,161$ *bits/átomo* ao diferenciar os tipos atômicos de acordo com os grupos G_{HP}, G_{res},

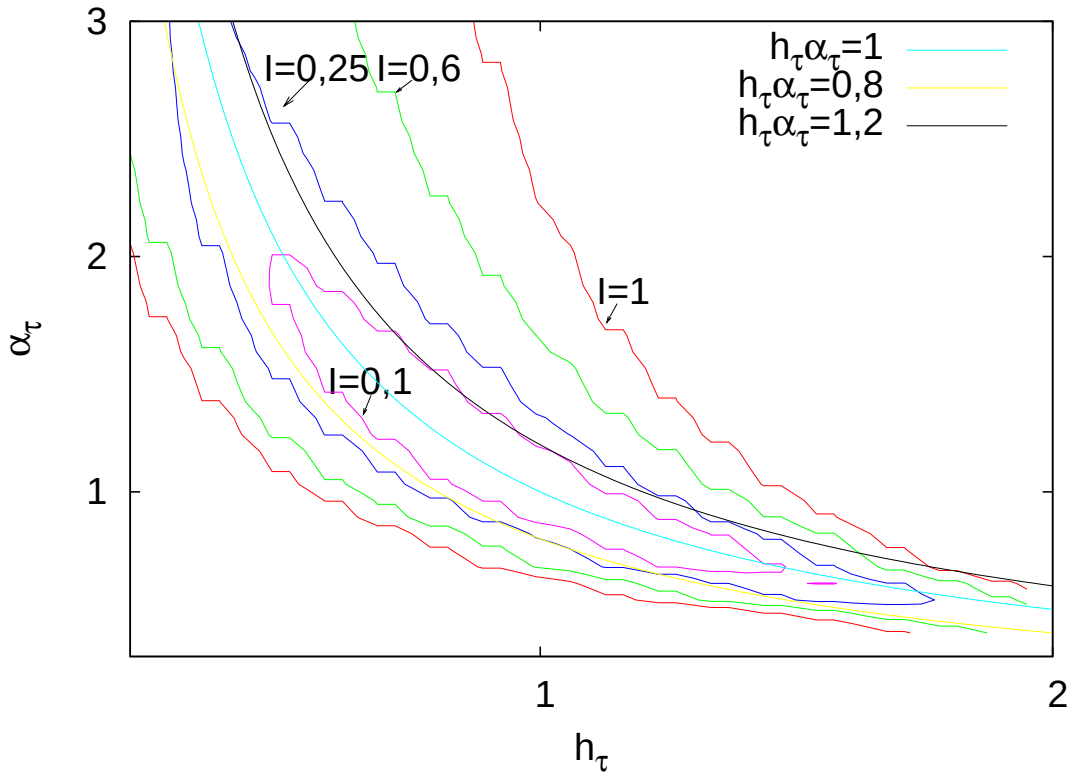


Figura 20: Curvas de nível para transinformação de primeira ordem em função de h_τ e α_τ são mostradas por linhas quebradiças para $0,1$; $0,25$; $0,6$ e $1 bits/átomo$. As linhas suaves representam curvas de isohidrofobicidade onde o produto $h_\tau\alpha_\tau$ é constante. Percebe-se que a taxa de variação da transinformação é maior quanto maior for o distanciamento em relação à curva neutra ($h_\tau\alpha_\tau = 1$).

G_{at}, G_{ph} respectivamente. Ao multiplicar esse valor por $7,8$; a média de átomos por resíduo desse banco de dados, encontramos a transinformação, de primeira ordem, em *bits/aminoácido*. Comparando os resultados, verificamos que a transinformação aumentou de $0,75$ *bits/aminoácido* para um modelo de duas letras (divisão G_{HP}) para $1,26$ *bits/aminoácido* na divisão G_{ph}. Esse resultado leva à mesma conclusão obtida comparando o valor-Z da energia da estrutura nativa para as diferentes maneiras de distinguir os tipos atômicos (seção 4.2). Dessa forma, concluímos que o aumento da especificidade dos tipos de átomos aumenta a informação que pode ser extraída da seqüência de aminoácidos, aprimorando a capacidade de predição de um enterramento ideal.

5.2 Tentativa inicial de predição dos enterramentos (exposto/enterrado).

Assumindo que o enterramento de um átomo deve ser influenciado pelos resíduos próximos ao longo da seqüência que o cerca, utilizamos um esquema análogo ao proposto por Garnier [44] para predição de estruturas secundárias a partir da contribuição de resíduos vizinhos. Um exemplo de janela de aminoácidos é mostrada na figura 21. As setas representam o fato da estrutura de um átomo central ser influenciada por seus resíduos vizinhos. A influência de um resíduo vizinho deve diminuir à medida que este se distância do átomo central. Esse fato é representado pela largura das linhas que compõem as setas e é corroborado pela figura 22, que representa a transinformação de cada tipo de resíduo para um C_α central estar exposto ou enterrado. Nessa análise, definimos duas possibilidades de estruturas, enterrada, em que $r_i < 1$ e exposta, onde $r_i \geq 1$.

Intuitivamente, a transinformação representa o quanto de informação pode ser extraída a partir de uma mensagem para a predição de outra. Uma maneira prática para utilizar a transinformação é relacioná-la como uma função de energia, capaz de distinguir estruturas preferenciais para cada letra. Em 1978, Garnier et al. baseou-se na transinformação para desenvolver um método, chamado de informação direcional, de predição de estrutura secundária [44]. De acordo com esse método, a propensão de um resíduo a uma determinada estrutura secundária pode ser calculada a partir de uma contribuição inde-

pendente de cada resíduo vizinho. Para este trabalho, adaptamos a informação direcional para tentar predizer enterramentos preferenciais.

Na subseção anterior, verificamos que a entropia da seqüência de aminoácidos é quase independente do tamanho de bloco considerado (figura 19a), pelo menos para blocos de poucos aminoácidos e, por isso, é razoável considerar que a probabilidade de se encontrar um par de resíduo x,y seja simplesmente o produto de suas probabilidades independentes, ou seja $p(x, y) = p(x)p(y)$. A inexistência de uma forte tendência dos resíduos em se agrupar em sub-seqüências preferenciais nos induz a considerar razoável que cada resíduo influencie de modo independente o enterramento de um átomo central.

A influência de cada resíduo vizinho para um átomo central pode ser calculada de acordo com a equação que representa a transinformação (equação 18). Nesse caso, Y representa o enterramento de um átomo central e X representa tanto o tipo de resíduo quanto sua posição em uma janela de aminoácidos vizinhos. Seguindo o raciocínio da informação direcional, cada resíduo vizinho contribui independentemente para o enterramento de um átomo central. Dessa forma, a partir do somatório da influência de cada resíduo, somos capazes de predizer o enterramento preferencial de cada átomo.

Na equação 21, representamos formalmente a função de energia para a estrutura de uma átomo central.

$$E(b_0|s) = -\sum_{i=-j}^{j} \log(\frac{P(b_0|s_i)}{P(b_0)}) = -\sum_{i=-j}^{j} I(b_0, s_i) \tag{21}$$

De acordo com essa equação, a energia da estrutura de um átomo é calculada para uma janela de $2j + 1$ resíduos, onde a contribuição de cada vizinho é somada de acordo com o tipo de resíduo e a posição i na janela, com i variando de $-j$ a $+j$. Nessa equação b_0 é o enterramento do átomo na posição central da janela e s_i representa o resíduo ocupando a posição i da janela. $E(b_0|s)$ é a energia que esse átomo, pertencente à uma seqüência s, contribuirá para o sistema ao estar no enterramento b_0. $P(b_0|s_i)$ é a probabilidade do enterramento b_0 tal que s_i, e $P(b_0)$ é a probabilidade de ocorrer o enterramento b_0, independente da seqüência, e $I(b_0, s_0)$ representa a mesma transinformação definida na equação 18.

a

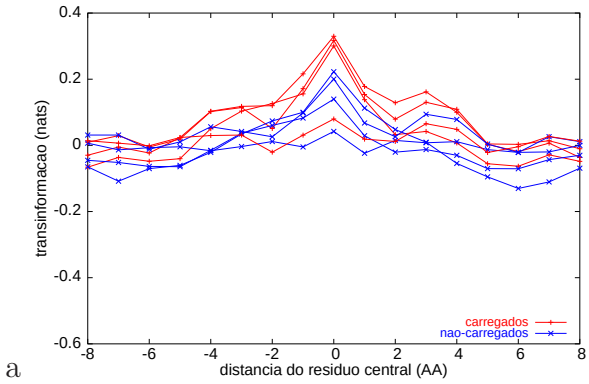
Figura 21: A hidrofobicidade dos resíduos vizinhos influenciam no enterramento de um átomo central. Quanto mais próximo o resíduo, maior será essa influência

Os valores usados nessa equação são referentes a traninsformação da estrutura ao saber o resíduo vizinho. Na figura 22, cada linha representa um tipo de resíduo. O eixo x indica a posição na janela e a abscissa refere-se à contribuição, dependente da posição da janela, que um tipo de resíduo acrescenta na propensão de um C_α central em estar exposto.

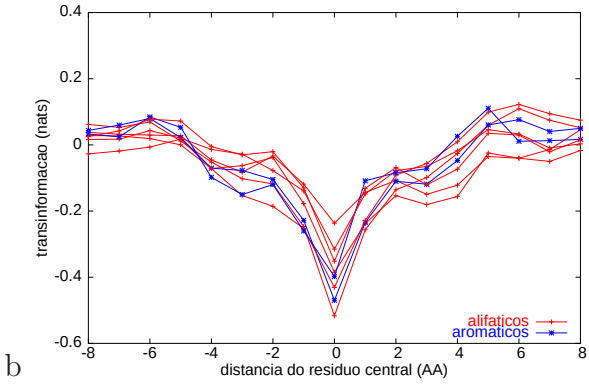
Na figura 22c, é mostrado os resíduos que possuem um comportamento diferente do padrão e eles são a Histidina, Alanina, Prolina e Glicina. Apesar da Prolina ser um iminoácido com cadeia lateral apolar, ela influencia um átomo central a estar exposto. Justifica-se essa situação pelo fato do nitrogênio da cadeia principal estar ligado à cadeia lateral, impossibilitando-o de fazer pontes de hidrogênio. Essa característica restringe a Prolina apenas como resíduo inicial de $\alpha - hélice$, tornando-a rara em regiões enterradas. A Prolina é mais comum em regiões de "loop" e ao final de $folhas - \beta$ antiparalelas [45].

A Alanina possui como cadeia lateral um grupo metil. O grupo metil é um grupo apolar, que combinado com o fato da cadeia principal ser polar, torna a Alanina um resíduo de comportamento neutro. Esse fato justifica a transinformação da Alanina ser próxima de zero, independente de sua posição na janela e é semelhante ao resultado obtido para h_τ. Outros trabalhos também descrevem a hidrofobicidade da alanina como neutra, como o trabalho de Meirovitch et al. [31, 32].

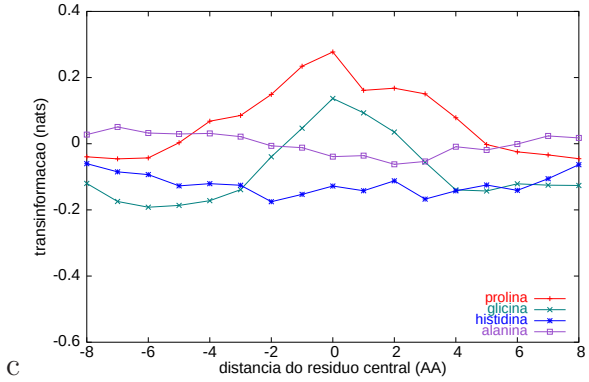
Com exceção das curvas para Histidina e Alanina, todas as curvas mostram um pico na posição do átomo central e o "peso"de cada resíduo vai diminuindo quanto mais afastado



a



b



c

Figura 22: Informação direcional para diferentes resíduos de aminoácidos. A informação direcional é calculada como a transinformação para o enterramento de um átomo central ao saber o tipo de resíduo e sua posição em uma janela de aminoácidos. Esses gráficos estão relacionados à conformação exposta do átomo ($r > 1$). Nota-se um pico na posição central (0), que representa o resíduo ao qual o átomo pertence.

este estiver na janela. Por esse fato, conclui-se que o resíduo central é o fator de maior influência para seu próprio enterramento. E a influência vai diminuindo para resíduos distantes do átomo central.

Além do C_α, também foi calculado a influência de resíduos vizinhos para o enterramento de C_β. Seus valores foram utilizados na equação 21 para previsão de estruturas preferenciais. A figura 23 mostra a percentagem desses átomos que estão em seu enterramento de menor energia, como uma função do tamanho da janela a ser considerado. Para esse resultado, essa função de energia de dois tipos (enterrado/exposto) foi aplicada ao próprio banco de dados em que foi extraída, indicando que 66% dos átomos estão com a energia mínima e outros 34% estão no enterramento mais energético. Não houve considerável diferença no sucesso da predição para C_α e C_β. A principal diferença encontrada foi que para o C_β era necessário uma janela menor para atingir a mesma informação. A causa da saturação precoce na informação para a predição de C_β pode estar relacionada ao fato de C_β ser um átomo pertencente à cadeia lateral. Seguindo essa linha de raciocínio, C_β de resíduos hidrofílicos estariam mais expostos que C_α e o inverso para os resíduos hidrofóbicos.

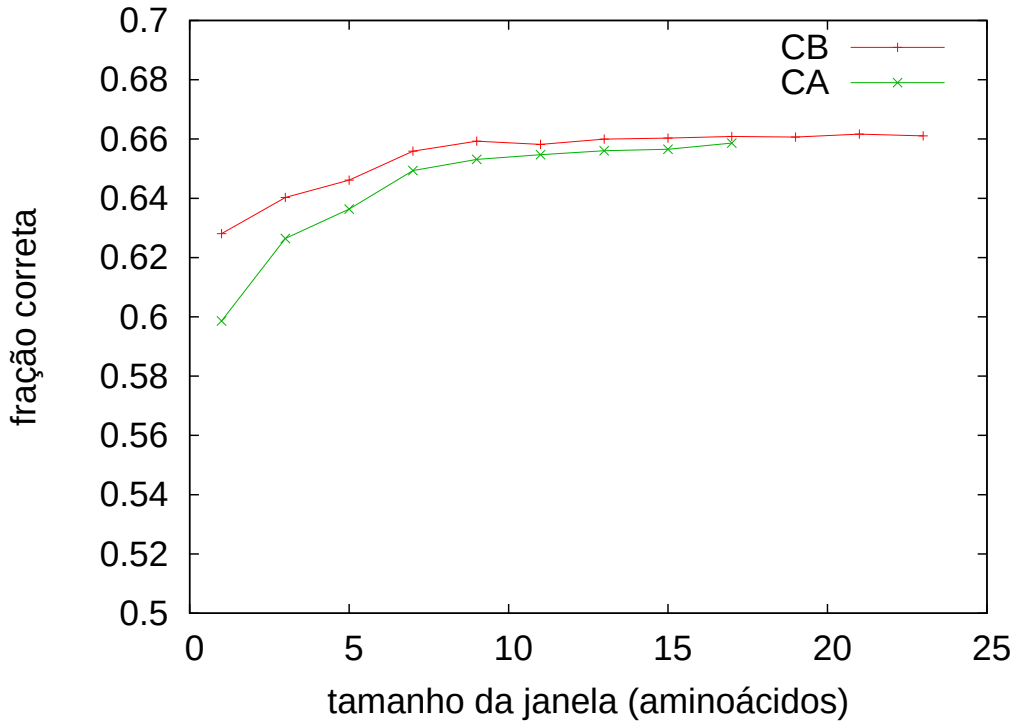


Figura 23: Percentagem de átomos que estão na posição de menor energia, de acordo com a função de energia descrita na equação 21.

6 Discussão

A variável distância reduzida foi pouco utilizada na literatura, mas pode vir a ser importante na busca de estruturas de uma proteína a partir da seqüência de aminoácidos. A definição de enterramento utilizada mostrou correlação com a hidrofobicidade dos resíduos obtida por outros trabalhos. A correlação para esse fato ocorre por causa da escolha apenas de proteínas globulares para o banco de dados. O ponto de corte para definir proteínas globulares foi um tanto arbitrário e um relaxamento em seus parâmetros manteve as mesmas características. Para o caso de selecionar apenas proteínas não globulares, a correlação de Pearson ao comparar h_τ (equação 13) dos resíduos com a hidrofobicade obtida por outros trabalhos diminui. Quando selecionamos apenas aquelas proteínas com $k > 4$ e $asf > 0.2$, o coeficiente de Pearson cai para $C = 0.73$. Ao Aumentar o rigor para definir proteínas não-globulares diminui-se ainda mais a correlação (quando consideramos $asf > 0.25$, o coeficiente de Pearson cai para 0.65).

Independente da relação com a Física, a distribuição de Fermi-Dirac pode ser uma poderosa ferramenta para a predição de estruturas de proteínas. Em primeiro lugar, poucos parâmetros foram necessários para o seu ajuste e esses parâmetros mostraram ser relacionados com propriedades químicas. Em segundo lugar, o cálculo para a variável distância reduzida é relativamente simples e serve como alternativa para outros métodos tradicionalmente utilizados na literatura.

O ajuste da distribuição dos átomos fornece uma função de energia específica para cada grupo atômico. Essa função de energia é possível de ser utilizada como potencial estatístico. Ela mostrou ser capaz de distinguir a estrutura nativa de estruturas compactadas ao acaso. Essa distinção melhorou ao se aumentar a especificidade dos grupos atômicos. Apesar dessa função de energia não ser suficiente para distinguir estruturas próximas da nativa, ela permite distinguir enovelamentos absurdos.

O potencial de energia ED, definido na seção 4.1 mostrou conter informação suficiente para dobrar proteínas globulares, pelo menos para aquelas de cadeia pequena. Comparado com o GO, esse potencial necessita de muito menos informação para ser extraído.

Utilizando ED, encontramos o valor-Z da energia da estrutura nativa, comparado com a energia de estrutura compactadas ao acaso em torno de 10, enquanto para GO o valor-Z se encontrou por volta de 50. Por esse motivo, consideramos $Z_n \approx 10$ uma referência de segregação energética suficiente para atingir a estrutura nativa.

O potencial mais informativo que obtivemos diferencia cada átomo e discrimina o envolvimento da principal em ponte de hidrogênio. Nesse caso, obtivemos $Z_n \approx 4,1$. Tendo como referência ED, precisamos extrair mais informação da seqüência de aminoácidos para se conseguir uma função onde a energia da estrutura nativa seja suficientemente segregada. Uma primeira tentativa para melhorar esse potencial energético é verificar a influência de resíduos vizinhos para a função de probabilidade do enterramento de um átomo.

Ao trabalhar com potenciais estatísticos, estamos sujeitos à saturação da informação e é preciso procurar uma método para conseguir maximizar a extração da informação. Para uma seqüência de m resíduos, existem 20^m possibilidades de combinação. Neste trabalho, o banco de dados continha aproximadamente 500000 resíduos de aminoácidos. Em média, cada combinação poderia ser amostrada $500000/20^m$. Para ilustração, consideramos o caso de $m = 3$. Nesse caso, seriam possíveis $62,5$ amostras em média por conformação. O valor de $62,5$ é pouco para ter dados suficientes e ajustar uma função $p_\tau(r)$, de acordo com a equação 13. Para contornar essa situação, uma maneira comumente usada para maximizar a informação é agrupar resíduos com propriedades semelhantes. Em um trabalho de 2000, Solis & Rackovsky [46] propuseram um método, baseado em Monte Carlo, para o agrupamento de resíduos no qual a informação a ser extraída de um banco de dados é maximizada.

Outra possibilidade é buscar algum padrão de influência de resíduos vizinhos. Para exemplicar o fato, suponha que a Lisina sempre desloque a distribuição dos átomos de seus resíduos vizinhos em $0,1$ à direita. Seria importante se fosse obtida uma função influência para cada resíduo, na qual fosse possível prever como cada resíduo afeta um resíduo vizinho. A função influência capacitaria a busca por informações que não são explicitamente presentes no banco de dados.

Ainda na busca por uma função de energia ideal, precisamos levar em conta que a definição de enterramento adotada nesse trabalho (distância reduzida), é uma aproximação e teria significado físico apenas em proteínas globulares. Como nem todas as proteínas são globulares, uma outra definição de enterramento, por exemplo a distância do átomo à superfície, pode fornecer uma função de energia mais informativa.

Referências

[1] DILL, K. A. Dominant forces in protein folding. *Biochemistry*, v. 29, n. 31, p. 7133–7155,

[2] ANSON, M. L.; MIRSJY, A. E. the reversibility of protein coagulation. *Journal of Physical Chemistry*, v. 35, n. 1, p. 185–193,

[3] EISENBERG, M. A.; W., S. G. The reversible heat denaturation of chimotrypsinogen. *The Journal of General Physiology*, v. 34, p. 583–605,

[4] TANFORD, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of the American Chemical Society*, v. 84, p. 4240–4247,

[5] 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 of the United States of America*, v. 47, n. 9, p. 1309–1314,

[6] LEVINTHAL, C. Are there pathways for protein folding? *Journal de Chimie Physique*, v. 65, n. 1, p. 44–45,

[7] ANFINSEN, C. B.; B., C. Principles that govern the folding of protein chains. *Science*, v. 181, n. 4096, p. 223–230,

[8] KENDREW, G. B. J. C. et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, v. 181, p. 662–668,

[9] GREEN, N. M.; FLANAGAN, M. T. The prediction of the conformation of membrane proteins from the sequence of amino acids. *Biochemistry journal*, v. 153, p. 729–732,

[10] SCHULZ, G. E. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature*, v. 250,

[11] ZWANZIG, R.; SZABO, A.; BAGCHI, B. Levinthal's paradox. *Proceedings of National Academic of Sciences of the United States of America*, v. 89, p. 20–22,

[12] DILL, K. A.; CHAN, H. S. From levinthal to pathways to funnels. *Nature Structural Biology*, v. 4, n. 1, p. 10–19,

[13] ONUCHIC, J. N.; WOLYNES, P. G. Theory of protein folding. *Current Opinion in Structural Biology*, v. 14, p. 70–75,

[14] FINKELSTEIN, A. V. Protein structure: what is it possible to predict now? *Folding and Binding*, v. 7, p. 60–71,

[15] SHIMADA, J.; SHAKHNOVICH, E. I. Evolution-like selecton of fast-folding model proteins. *Proceeds of the National Academic of Sciences of the United States of America*, v. 92, p. 1282–1286,

[16] PRIVALOV, P. L. *Protein Folding*. [S.l.]: T.E. Creighton, 1992.

[17] DILL, K. A. et al. Modelling water, the hydrophobic effect, and ion solvation. *Annual Review of Biophysics and Biomolecular Structure*, v. 34, p. 173–199,

[18] FAUCHÈRE, J.; PLISKA, V. Hydrophobic parameters of pi amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. *European Journal of Medicinal Chemistry*, v. 18, p. 369–375,

[19] MILLER, S. et al. Interior and surface of monomeric proteins. *Journal of Molecular Biology*, v. 196, p. 644–656,

[20] ARAÚJO, A. F. Pereira de. Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation. *Proceeds of the National Academic of Sciences of the United States of America*, v. 96, n. 22, p. 12482–12487,

[21] SCHINDLER, T. et al. Extremely rapid protein folding in the absence of intermediates. *Nature Structural Biology*, v. 2, n. 8, p. 663–673,

[22] HOANG, T. X. et al. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proceedings of the National Academy of Science of the United States of America*, v. 101, n. 21, p. 7960–7964,

[23] LEVY, Y.; ONUCHIC, J. N. Mechanisms of protein assembly: Lessons from minimalist models. *Accounts of Chemical Research*, v. 39, n. 2, p. 135–142,

[24] GARCIA, L. G.; ARAÚJO, A. F. Pereira de. Folding pathway dependence on energetic frustration and interaction heterogeneity for a three-dimensional hydrophobic protein model. *Proteins: Structure, Function and Bioinformatics*, v. 62, p. 46–63,

[25] LAU, K. F.; DILL, K. A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, v. 22, n. 10, p. 3986–3997,

[26] SHAKNOVICH, E. I. Proteins with selected sequences fold into unique native conformation. *Physical Review Letters*, v. 72, n. 24, p. 3907–3910,

[27] GARCIA, L. G.; TREPTOW, W. L.; Pereira de Araujo, A. F. Folding simulations of a three-dimesional protein model with an non-specific hydrophobic energy function. *Physical review E*, v. 64, p. 011912–1–011912–5,

[28] HOBOHM, U.; SANDER, C. Enlarged representative set of protein structures. *Protein Science*, v. 3, p. 522,

[29] BAUMÄRTNER, A. Shapes of flexible vesicles at constant volume. *Journal of chemical physics.*, v. 98, n. 9, p. 7496–7501,

[30] TSAI, J. et al. The packing density in proteins: Standard radii and volumes. *Journal of Molecular Biology*, v. 290, p. 253–266,

[31] MEIROVITCH, H.; RACKOVSKY, S.; SCHERAGA, H. A. Empirical studies of hydrophobicity. 1.effect of protein size on the hydrophobic behavior of amino acids. *Macromolecules*, v. 13, p. 1398–1405,

[32] MEIROVITCH, H.; SCHERAGA, H. A. Empirical studies of hydrophobicity. 2. distribution of the hydrophobic, hydrophilic, neutral and ambivalent amino acids in the interior and exterior layers of native proteins. *Macromolecules*, v. 13, p. 1406–1414,

[33] REIF, f. *Fundamentals of statistical and thermal physics*. New York, USA: Mcgraw-Hill,

[34] GOMES, A. 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,

[35] LEVITT, M.; CHOTHIA, C. Structural patterns in globular proteins. *Nature*, v. 261, n. 5561, p. 552–558,

[36] SHIMADA, J.; KUSSELL, E. L.; SHAKHNOVICH, E. I. The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. *Journal of Molecular Biology*, v. 308, n. 1, p. 79–95,

[37] METROPOLIS, N. et al. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, v. 21, p. 1087–1092,

[38] SHIMADA, J.; SHAKHNOVICH, E. I. The ensemble folding kinetics of protein g from an all-atom monte carlo simulation. *Proceeds of the National Academic of Sciences of the United States of America*, v. 99, n. 17, p. 11175–11180,

[39] BOWIE, J. U.; LÜTHY, R.; EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, v. 253, p. 164–169,

[40] PIERCE, J. R. *An introduction to information theory: Symbols, signals and Noise*. New York, USA: Dover Publications, Inc.,

[41] REZA, F. M. *An introduction to information theory*. [S.l.]: Dover Publications, Inc., 1994.

[42] BETANCOURT, M. R.; SKOLNICK, J. Local propensities and statistical potentials of backbone dihedral angles in proteins. *Journal of Molecular Biology*, v. 342, p. 635–649,

[43] STRAIT, B. J.; DEWEY, T. G. The shannon information entropy of protein sequences. *Biophysical Journal*, v. 71, p. 148–155,

[44] GARNIER, J.; OSGUTHORPE, D. J.; ROBSON, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, v. 120, n. 1, p. 97–120,

[45] FASMAN, G. D. *Prediction of protein structure and the principles of protein conformation*. [S.l.]: Plenum Press, 1989.

[46] SOLIS, A. D.; RACKOVSKY, S. Optimized representations and maximal information in proteins. *Proteins: Structure, Function, and Genetics*, v. 38, p. 149–164,

7 Anexo (artigo publicado)

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo