

André Luiz Martins

Predição de Estruturas de Proteínas Utilizando o
Algoritmo da Rede Elástica



1910001433



André Luiz Martins

**Predição de Estruturas de Proteínas Utilizando o Algoritmo da Rede
Elástica**

Dissertação apresentada como parte das exigências para
obtenção do título de Mestre em Biofísica Molecular, área
de concentração Biofísica Molecular do Departamento de
Física do Instituto de Biociências, Letras e Ciências Exatas
da Universidade Estadual Paulista "Julio de Mesquita Filho"
– UNESP.

T.1433/05



Orientador: *Prof. Dr. Jorge Chahine*

São Jose do Rio Preto – SP
2005

577.112
M386p

07/10/2005

Doação

R\$ 20,00

Autor(a)

UNIVERSIDADE ESTADUAL PAULISTA
"JULIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

unesp

André Luiz Martins

Predição de Estruturas de Proteínas Utilizando o Algoritmo da Rede
Elástica

Martins, André Luiz.

Predição de Estruturas de Proteínas Utilizando o Algoritmo da Rede
Elástica / André Luiz Martins. - São José do Rio Preto : [s.n.], 2005
62 f. : il. ; 30 cm.

Orientador: Jorge Chahine

Dissertação (Mestrado) - Universidade Estadual Paulista. Instituto de
Biociências, Letras e Ciências Exatas

1. Proteínas - Estrutura. 2. Problema do caixeiro viajante. 3. Algoritmo da
rede elástica. 4. Predição de estrutura. I. Chahine, Jorge. II. Universidade
Estadual Paulista. Instituto de Biociências, Letras e Ciências Exatas. III.
Título.

CDU - 577.112

Orientador: Prof. Dr. Jorge Chahine

São José do Rio Preto - SP
2005



DECLARAÇÕES

Eu, **Prof. Dr. Jorge Caramelo**, pelo meu cargo de **Reitor**, e pelo meu cargo de **representante** da **Comissão** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**, **declaro** **que** **o** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná** **está** **em** **conformidade** **com** **o** **regulamento** **de** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**.

A **COMISSÃO** **de** **procuradores** **do** **Departamento** **de** **Defesa** **do** **Estado** **de** **Paraná**, **por** **meio** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**, **declaro** **que** **o** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná** **está** **em** **conformidade** **com** **o** **regulamento** **de** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**.

A **COMISSÃO** **de** **procuradores** **do** **Departamento** **de** **Defesa** **do** **Estado** **de** **Paraná**, **por** **meio** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**, **declaro** **que** **o** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná** **está** **em** **conformidade** **com** **o** **regulamento** **de** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**.

Aos **Procuradores** **do** **Departamento** **de** **Defesa** **do** **Estado** **de** **Paraná**, **por** **meio** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**, **declaro** **que** **o** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná** **está** **em** **conformidade** **com** **o** **regulamento** **de** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**.

Aos **Procuradores** **do** **Departamento** **de** **Defesa** **do** **Estado** **de** **Paraná**, **por** **meio** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**, **declaro** **que** **o** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná** **está** **em** **conformidade** **com** **o** **regulamento** **de** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**.

A **COMISSÃO** **de** **procuradores** **do** **Departamento** **de** **Defesa** **do** **Estado** **de** **Paraná**, **por** **meio** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**, **declaro** **que** **o** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná** **está** **em** **conformidade** **com** **o** **regulamento** **de** **procedimento** **de** **investigação** **de** **conduta** **do** **Procurador** **de** **Defesa** **do** **Estado** **de** **Paraná**.

A minha mãe, Marlene
Pereira Martins, que sempre
me amparou e incentivou
em minhas decisões.

AGRADECIMENTOS

Ao Prof. Dr. Jorge Chahine, pela excelente orientação, e pelo crescimento profissional e científico proporcionado durante a realização desse trabalho.

A TODOS os professores do Departamento de Física, pela atenção, companheirismo e incentivo. Em especial aos professores Dr. José Roberto Ruggiero e Augusto Agostinho Neto pela participação no exame de qualificação deste trabalho. E aos professores que participaram na banca de defesa Prof. Dr. Vitor Barbanti Pereira Leite e Prof. Dr. Marco Antonio Alves da Silva.

A TODOS os amigos do Departamento de Física. Em especial ao Leandro Cristante, Sidney J. de Carvalho e Carlos Emídio Sampaio pela ajuda computacional. E ao amigo de sala Diogo de Moraes.

Aos Meus Pais, Gilberto C. Martins e Marlene P. Martins; Irmãos, Célio Roberto Martins e Gilda M. Assunção; e Sobrinhos, Alex C. Martins, Fabrício Jr. M. Assunção, pelo companheirismo e pelo incentivo para que eu pudesse realizar meus sonhos.

Aos meus grandes amigos Ana Helena Casé, Rafaella Flávio de Lima e Ricardo Barcelos Souza por todos esses anos de amizade, por me ouvirem e me apoiarem todas as vezes que precisei.

A CAPES pelo auxílio financeiro para que eu pudesse desenvolver minhas pesquisas.



Índice Geral

Índice Geral	i
Índice de Figuras	ii
Índice de Tabelas	iii
Resumo	1
Abstract	2
Capítulo 1- Introdução	3
Capítulo 2 – Metodologia	5
2.1.1 - Método da Rede Elástica e aplicação do PCV	5
2.1.2 - Resultados do Método da RE Aplicado ao PCV	9
2.2 - Descrição Mecânico – Estatística	12
2.3.1 - Aplicação do Método da RE à Predição de Estrutura de Proteínas	16
2.3.2 - Descrição da Rede Cubo-Octaédrica	20
2.3.3 - Um Método Iterativo para Extrair Potenciais a Partir de Estruturas de Proteína	
Proposto por Thomas, Dill e Colaboradores	22
Capítulo 3 - Resultados e discussões	29
Método da RE Aplicado à Predição de Estruturas de Proteínas	29
Capítulo 4 – Conclusões	44
Capítulo 5 – Referências Bibliográficas	46



Índice de Figuras

Figura 1 – Distribuição de cidades para o Problema do Caixeiro Viajante	11
Figura 2 – Ilustração do modelo de uma cela da rede Cubo-Octaédrica	22
Figura 3 – Diagrama de fluxo do Método ENERGI	27
Figura 4 – Gráficos do RMS em função da temperatura	36
Figura 5 – Trechos de estruturas preditas pelo algoritmo da RE	40
Figura 6 – Comparação das estruturas preditas para a proteína Ieca	41
Figura 7 – Sobreposição de estruturas preditas com a estrutura nativa	42
Figura 8: Curvas de RMS x T para a predição da estrutura da proteína Ivii	43



Índice de Tabelas

Tabela 1 – Parâmetros para a função energia interna da proteína	19
Tabela 2 – Parâmetros de energia de contato resíduo-resíduo	28
Tabela 3 – Valores de RMS para as estruturas preditas pela Rede Elástica em $T = 1$	37
Tabela 4: Classificação dos resíduos de 3 proteínas que apresentaram curvas semelhantes de RMS	38



RESUMO

Neste trabalho utilizamos o método da rede elástica aplicado à predição de estruturas de proteínas, a partir de suas seqüências de aminoácidos. Inicialmente, reproduzimos possíveis soluções de um problema de otimização global clássico, conhecido como Problema do Caixeiro Viajante (PCV). Ao PCV aplicamos a estratégia de otimização da Rede Elástica (RE) de Durbin-Willshaw (DURBIN e WILLSHAW, 1987). Posteriormente, o método da RE, associado com potenciais estatísticos obtidos de um banco de dados, e algumas informações experimentais, como raio de giro da proteína, foi aplicado à predição de estruturas protéicas.



ABSTRACT

In this work we use the elastic net method applied to prediction of protein structure from their amino acid sequence. Initially we reproduce possible solutions of a classic optimization problem known as Salesman Travelling Problem (STP). We apply the strategy of Durbin-Willshaw (DURBIN e WILLSHAW, 1987) elastic net (EN) method. To the STP the EN method can be associated with data base-derived potentials and some experimental data, for example radius of giration, for predicting protein structure.



CAPÍTULO 1

INTRODUÇÃO

Proteínas são cadeias polipeptídicas responsáveis por grande parte das funções básicas da vida no nível molecular. Estas cadeias lineares enovelam-se em complexas estruturas 3D, cuja forma está relacionada à função da proteína. Uma proteína é constituída por uma sequência (estrutura primária da proteína) de 20 aminoácidos existentes na natureza. No entanto, os aminoácidos de qualquer proteína não ocorrem nas mesmas proporções, e nem todas as proteínas contém todos os 20 aminoácidos comuns. As proteínas são sintetizadas dentro das células (CREIGHTON, 1993). As instruções usadas pela célula para construir uma proteína são escritas no DNA. A dupla hélice do DNA é constituída por duas longas cadeias compostas por quatro diferentes nucleotídeos. O DNA contém genes, sequências de nucleotídeos codificador de proteínas. Cada tripleto de nucleotídeo num gene corresponde a um aminoácido da proteína decodificada (VOET, VOET E PRATT, 2002).

Graças aos vários projetos de sequenciamento genômico, toda sequência de DNA de muitos organismos têm sido determinadas experimentalmente. Em cada genoma as posições dos genes são descobertas usando sinais específicos, sequências particulares de nucleotídeos usados pelas células durante a transcrição. A partir destes genes identificados a estrutura primária da proteína é extraída (CERONI et al, 2003). O conhecimento da sequência de aminoácidos de uma proteína é pré-requisito para determinar sua estrutura



tridimensional (3D) e é essencial para entender seu mecanismo de ação. Infelizmente nosso conhecimento freqüentemente para por aqui. As estruturas 3D da proteína, essencial para o estudo de suas funções, permanecem desconhecidas. Existem métodos físicos para estimar as coordenadas de cada átomo da proteína, mas eles precisam da cristalização de proteína. Se o número de proteínas cuja seqüência primária conhecido fosse da ordem de milhares, somente poucas centenas delas teriam sido cristalizadas e suas estruturas 3D depositadas no *Protein Data Bank*. A estrutura cristalizada pode ser diferente daquela em atividade. Por isso torna-se importante prever estrutura terciária de proteínas através do uso de algoritmo computacional.

O problema clássico de enovelamento de proteína está em entender este processo, a fim de prever a conformação 3D da proteína, dada somente a sua seqüência de aminoácido (MAIOROV E CRIPPEN, 1992). Neste trabalho aplicaremos o método da Rede Elástica (RE) à predição de estruturas de proteínas. Este algoritmo contém uma série de parâmetros que influenciam na formação da estrutura terciária da proteína. Para cada estrutura primária, temos que encontrar a melhor relação entre estes parâmetros que forneça o menor RMS da estrutura 3D predita em relação à nativa. Testamos o algoritmo em um conjunto de 20 proteínas, que variam de 36 a 136 aminoácidos. Nas simulações fizemos o processo de annealing iniciando com temperatura por volta de setenta, e reduzida até um.

Inicialmente, reproduzimos possíveis soluções de um problema de otimização global clássico, conhecido como Problema do Caixeiro Viajante (PCV). Ambos PCV e enovelamento de proteína podem ser descritos em termos do perfil da superfície de energia. No próximo capítulo mostraremos como estruturar a predição de estrutura de proteína em termos do PCV o qual é resolvido usando o método da rede elástica, RE, (DURBIN e WILLSHAW, 1987).



CAPÍTULO 2

METODOLOGIA

2.1.1 - Método da Rede Elástica e aplicação ao PCV

Neste tópico ilustraremos as características essenciais do método da rede elástica fazendo uma revisão de sua formulação original feita por Durbin-Willshaw e aplicado ao PCV: *Dado as posições de M cidades, as quais no caso mais simples estão no plano euclidiano bi-dimensional, qual é o menor trajeto (tour) fechado no qual cada cidade pode ser visitada uma única vez?*

No PCV são dadas as coordenadas de um conjunto fixo de m cidades, $S = S_1, S_2, \dots, S_m$ no plano xy , onde o vetor S_i é a localização da cidade i . O caixeiro viajante deve visitar cada cidade somente uma vez, e assim, seguindo um caminho ou *tour*, passando pelas cidades numa ordem particular. A linha fechada consiste de n *tour points*, $R = R_1, R_2, \dots, R_n$, com R_j sendo o vetor posição da j -ésima parada do caixeiro.

Resolver o PCV, é uma questão de adequar os pontos de parada (*tour stops*) do caixeiro viajante às cidades, tal que:

- (i) Cada cidade seja visitada exatamente uma vez (restrição de ajuste);
- (ii) O *tour* é fechado ($R_{n+1} \equiv R_1$);



(iii) O *tour* é o menor caminho possível que passa pelas cidades obedecendo (i) e (ii).

Isto é, o comprimento total do *tour*,

$$L = \sum_{j=1}^{n-1} |\mathbf{R}_{j+1} - \mathbf{R}_j| \quad (1)$$

está em seu mínimo global.

Para se ater a esses vínculos, o algoritmo faz sucessivas iterações das posições de um número de pontos, no plano em que as cidades ficam. Inicialmente o caminho fechado descrito pelos pontos, é um pequeno círculo, localizado sobre o centro de distribuição das cidades, sendo gradualmente alongado, não uniformemente, para passar eventualmente próximo de todas as cidades, e assim definir um *tour* em torno delas. Cada ponto sobre o caminho é movido sob a influência de dois tipos de forças, ou melhor, por dois conjuntos diferentes de molas harmônicas.

O primeiro conjunto de n molas conecta os *tour points* num *loop* fechado. Chamaremos estas de *tour springs*, porque elas descrevem como os *tour points* estão conectados em sucessão. Nesta representação, o *tour* é análogo a uma cadeia de polímero linear, ou contas conectadas por molas, as quais podem explorar diferentes conformações no plano passando através das cidades. Cada *tour spring* tem energia $E_j = \alpha_t (\mathbf{R}_{j+1} - \mathbf{R}_j)^2$, onde α_t é uma constante de força. Estas molas atrativas servem para manter o *tour*, tão pequeno quanto possível. O segundo conjunto de mn molas conecta cada *tour point* $\mathbf{R}_j$ em cada cidade $\mathbf{S}_i$, com energia proporcional a $(\mathbf{S}_i - \mathbf{R}_j)^2$. Denominadas de *city springs*, essas molas tendem a forçar o *tour* do caixeiro a passar pelas cidades.

Desta forma, cada cidade torna-se associada com uma seção particular do caminho. Inicialmente, todas as cidades têm grosseiramente, a mesma influência sobre cada ponto no



caminho. Posteriormente, maiores distâncias tornam-se menos favorecidas, e cada cidade gradualmente, torna-se mais específica para os pontos mais próximos dela sobre o caminho. Este aumento gradual da especificidade corresponde à redução da ‘temperatura’ em *simulated annealing*. Devido à forma como o caminho inicial é gradualmente modificado para produzir o *tour* final, este algoritmo é conhecido como método da Rede Elástica (RE). No algoritmo RE, existem três escolhas possíveis para o número de cidades m , e o número de *tour points* n :

- (i) $m < n$,
- (ii) $m = n$,
- (iii) $m > n$.

Para que a restrição de ajuste no PCV seja satisfeita, devemos selecionar (i) ou (ii). No caso (i) é possível ter mais *tour points* do que cidades, e somente m destes *points* finalmente tornam-se *tour stops*. Na literatura é comum encontrar $n = 2,5m$. Em temperatura finita, um *tour* elástico tenderia a ‘cruzar’ próximo às cidades para minimizar a penalidade do comprimento. Usando mais *tour points* do que cidades no PCV reduzem esta tendência, pois assim, diminui-se o intervalo médio entre os *tour stops*, relaxando a tensão elástica no *tour*. Este relaxamento permite às *city springs* no caso (i), puxarem os *tour stops* mais próximos para elas, do que no caso (ii).

O método da RE, baseia-se num procedimento iterativo, que calcula sobre cada ciclo, uma função de distribuição de Boltzmann ϕ , para cada *city springs*:

$$\phi(|S_i - R_j|, K) = \exp\left(-(|S_i - R_j|)^2 / 2K^2\right) \quad (2).$$

Onde K é um parâmetro. Escrevemos a energia livre para este sistema, E , é definida como:



$$E = -\alpha_c K \sum_{i=1}^m \log \sum_{j=1}^n \phi(|S_i - R_j|, K) + \alpha_t \sum_{j=1}^n |R_{j+1} - R_j|^2 \quad (3)$$

As constantes α_c e α_t são coeficientes que controlam a intensidade relativa das *city* e *tour springs*, respectivamente. Como estamos descrevendo a energia do sistema como uma energia de mola, podemos calcular a força que atua sobre um tour point $F = -\nabla E$. O caixeiro viajante desloca-se $\Delta R_j = -K \partial E / \partial R_j$ (ver equação (5)), e dessa forma qualquer mudança em R_j resulta na redução no valor da E , e um mínimo local de E pode ser alcançado. No limite de $K \rightarrow 0$, o caminho deve passar através de todas cidades para minimizar a energia pois para todo i existe um j tal que $|S_i - R_j|$ tende para zero, condição essa que maximiza o peso de Boltzmann ϕ para $K \rightarrow 0$ (equação (2)).

Normalizando sobre todas as *city springs*, conectadas para a cidade i , obtém o seguinte peso estatístico para cada uma destas molas:

$$w_{ij} = \frac{\phi(|S_i - R_j|, K)}{\sum_{k=1}^n \phi(|S_i - R_k|, K)} \quad (4)$$

Estes pesos especificam a influência da cidade i sobre um tour point j , atuando como constantes de molas variáveis, as quais para um valor de K fixo tornam-se mais fracas quando a distância $|S_i - R_j|$ aumenta. Para encontrar o *tour* mais favorável, o método da RE, progressivamente, acrescenta mudanças nas coordenadas de cada R_j sobre o *tour*, de acordo com a expressão abaixo, obtida de $\Delta R_j = -K \partial E / \partial R_j$.

$$\Delta R_j = \alpha_c \sum_{i=1}^m w_{ij} (S_i - R_j) + 2 K \alpha_t (R_{j+1} - 2 R_j + R_{j-1}) \quad (5)$$



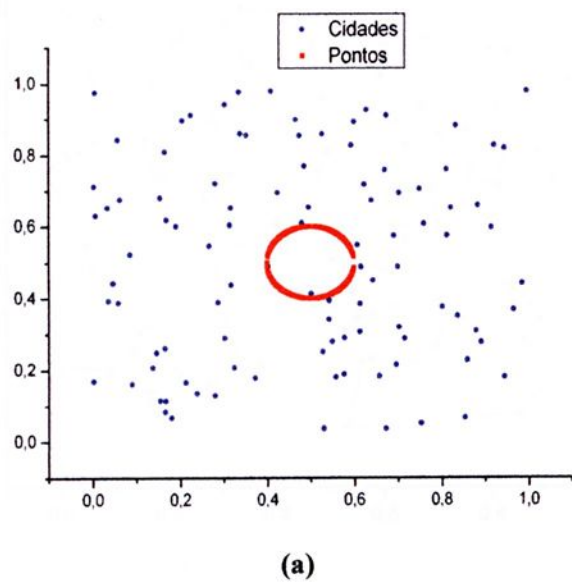
O método da RE começa num valor de K suficientemente alto, tal que todas *city springs* tem essencialmente a mesma intensidade, e fornecem uma força uniforme sobre cada ponto ao longo do *tour* do caixeiro viajante. Devido ao fato dos pontos não sentirem uma força líquida, na direção de uma cidade qualquer, o *tour* encolhe para seu menor comprimento possível, ou seja, $L = 0$. Quando K é reduzido, o peso da mola determinado pela equação (2), aproxima-se da unidade para o alongamento da menor *city spring* ligada a uma dada cidade, enquanto que, a intensidade das outras $n - 1$ molas associadas com esta cidade, diminuem exponencialmente. Conseqüentemente, abaixo de algum valor crítico de K , o *tour point* mais próximo, é puxado preferencialmente na direção desta cidade, enquanto os outros são puxados para longe dela, movendo-se na direção das outras cidades. A escolha de w_{ij} na equação (4), assegura que no limite $K \rightarrow 0$, exatamente um *tour point* será puxado para cada cidade, assim satisfazendo a restrição de ajuste.

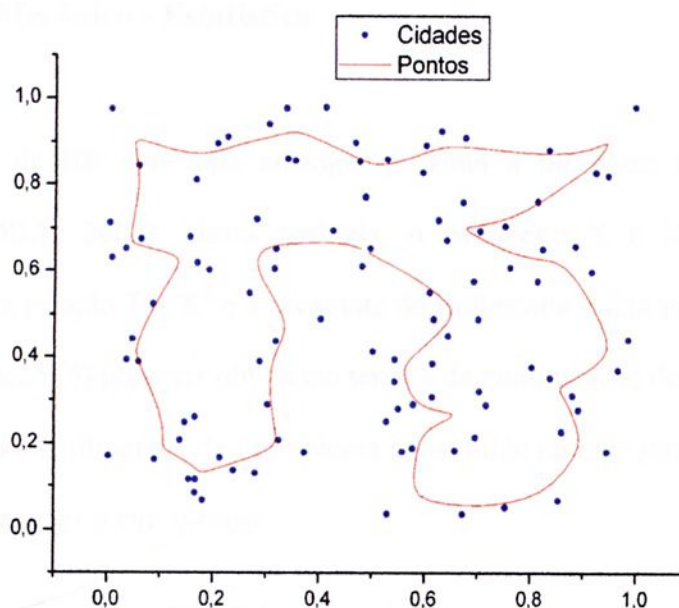
2.1.2 - Resultados Método da RE Aplicado ao Problema do Caixeiro Viajante

Apresentaremos neste tópico os resultados que foram obtidos na resolução do PCV, como descrito na seção anterior. Inicialmente temos uma distribuição aleatória de 100 cidades no plano cartesiano xy . Um conjunto de 250 pontos foi colocado no centro de distribuição destas cidades, num círculo de raio igual a 0,1, conforme mostrado na figura 1(a). Quando o parâmetro de comprimento, K , na equação (2) é suficientemente alto a força

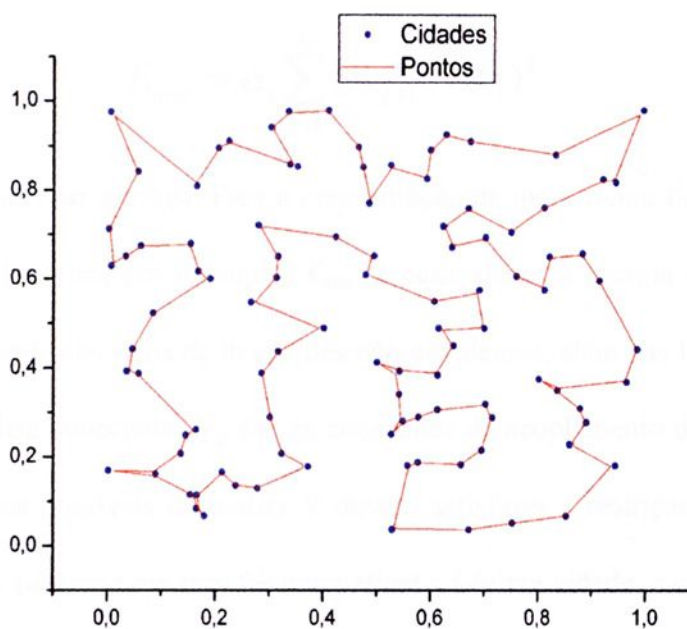


das *city springs* têm aproximadamente a mesma influência sobre um determinado *tour point*. Assim as *tour springs* conectam os pontos para formar um *tour* fechado, minimizando as distâncias entre os pontos. O valor inicial de K para esta distribuição das cidades, pode variar de 0,2 até 0,25. Acima de 0,25 as *tour springs* compactam os pontos no centro da rede (0,5; 0,5). À medida que K é reduzido (análogo à redução da temperatura no *simulated annealing*), ocorre um aumento gradual da especificidade de uma cidade para puxar o *tour point* mais próximo dela, formando assim o *tour* do caixeiro viajante, figura 3(b). Em nossa simulação reduzimos K por 1%, de 0,25 até 0,01, fazendo 300 iterações para cada valor de K fixo. Os valores das constantes α_c e α_t na equação (5) foram de: $\alpha_c = 0,15$ e $\alpha_t = 2,0$. Com este conjunto de parâmetros encontramos um tour final de $L = 7,59$, figura 3(c). Na literatura para este mesmo conjunto de 100 cidades o comprimento final do tour é de $L = 7,7$ com os parâmetros $\alpha_c = 0,2$ e $\alpha_t = 2,0$.





(b)



(c)

Figura 1: Exemplo do progresso do método da rede elástica para 100 cidades distribuídas randomicamente em uma rede quadrada unitária. Configuração inicial (a), em $K = 0,06$ (b) e final $K = 0,01$ (c).

2.2 – Descrição Mecânico - Estatística

O método da RE tem uma analogia próxima à mecânica estatística (BALL, ERMAN AND DILL, 2002). Nessa analogia, o parâmetro K é identificado com a temperatura T via a relação $T = K^2$ e a constante de Boltzmann é fixa na unidade. Pode-se mostrar que a equação (5) pode ser obtida em termos da minimização de uma energia livre. Começamos com a Hamiltoniana de um sistema constituído de *tour points* e cidades, com as interações *tour springs* e *city springs*:

$$H = \sum_{i,j} \frac{V_{ij}}{2} (\mathbf{S}_i - \mathbf{R}_j)^2 + E_{tour} \quad (6)$$

onde

$$E_{tour} = \alpha_t \sum_{j=1}^n (\mathbf{R}_{j+1} - \mathbf{R}_j)^2 \quad (7)$$

é a energia total das *tour springs*. Para a conformação de um sistema físico (uma proteína ou uma cadeia polimérica, por exemplo) E_{tour} é equivalente à energia interna, a qual em geral deve incluir todos os tipos de interações não-covalentes, além das ligações covalentes que mantém a cadeia conectada. V_{ij} são as constantes de acoplamento das cidades com os *tour points*. Valores possíveis da matriz V devem satisfazer a restrição de ajuste, a qual requer $V_{ij} = 1$ se o j -ésimo *tour stop* for compatível a i -ésima cidade, e caso contrário $V_{ij} = 0$. Isto é, a matriz V_{ij} deve conter somente um único 1 em cada coluna e linha. A função de partição Z para a cadeia é dada por



$$Z = \int_{\mathbf{R}} \sum_{V_{ij}} e^{-H(\mathbf{R})/T} d\Gamma \quad (8)$$

onde a soma é realizada sobre todas as conformações possíveis $\{\mathbf{R}\}$ da cadeia e sobre todas matrizes possíveis $V = [V_{ij}]$. Da Hamiltoniana da eq (6)

$$Z = \int_{\mathbf{R}} \sum_V e^{-\sum_{i,j} \frac{V_{ij}}{2T} (\mathbf{S}_i - \mathbf{R}_j)^2} \cdot e^{-E_{tour}/T} d\Gamma \quad (9)$$

Podemos reescrever a primeira exponencial

$$e^{-\sum_{i,j} \frac{V_{ij}}{2T} (\mathbf{S}_i - \mathbf{R}_j)^2} = \prod_{i=1}^m e^{-\sum_{j=1}^n \frac{V_{ij}}{2T} (\mathbf{S}_i - \mathbf{R}_j)^2}$$

Logo obtemos

$$Z = \int d^{3n} \mathbf{R} \sum_V \left\{ \prod_{i=1}^m e^{-\sum_{j=1}^n \frac{V_{ij}}{2T} (\mathbf{S}_i - \mathbf{R}_j)^2} \right\} \cdot e^{-E_{tour}/T} \quad (10)$$

Em geral, encontrar a matriz V que maximiza a função de partição através do esgotamento de todas as possibilidades, é um problema combinatório proibitivo, mesmo com as restrições de ajuste sobre V_{ij} . Ao invés disso, toma-se um *ensemble* de combinações possíveis, e assume que o acoplamento de um *tour point* para cada cidade é governado por uma distribuição de probabilidades de Boltzmann que são independentes para cada cidade. Com essas suposições, a função de partição pode ser escrita numa forma mais simples:

$$Z = \int d^{3n} \mathbf{R} \prod_{i=1}^m \left\{ \sum_{j=1}^n e^{-(\mathbf{S}_i - \mathbf{R}_j)^2 / 2T} \right\} \cdot e^{-E_{tour}/T} \quad (11)$$

A função de partição Z é a integral sobre as contribuições de cada conformação $\mathbf{R}$ e pode ser escrita em termos da energia livre microscópica $F(\mathbf{R}, T)$ na forma:

$$Z = \int d^{3n} \mathbf{R} e^{-F(\mathbf{R}, T)/T} \quad (12)$$

onde $F(\mathbf{R}, T)$ é a energia livre interna definida como:

$$F(\mathbf{R}, T) = U(\mathbf{R}) - TS(\mathbf{R}) \quad (13)$$

$$F(\mathbf{R}, T) = E_{tour} - T \sum_{i=1}^m \log \left\{ \sum_{j=1}^n e^{-(\mathbf{S}_i - \mathbf{R}_j)^2 / 2T} \right\} \quad (14)$$

A energia livre é composta da energia interna conformacional $U \equiv E_{tour}$, e dos efeitos entrópicos das *city springs* atuando através $\mathbf{S}$, a qual tende a dispersar os *tour points* igualmente no espaço em altas temperaturas. Quando T é reduzida, as conformações que estiverem mais próximas de satisfazerem a restrição de ajuste têm maior entropia e, portanto têm menor energia livre do que outras conformações. Quando $T \rightarrow 0$ somente as conformações que satisfazem a restrição têm probabilidade de ocupação significativa (YUILLE and KOSOWSKY, 1994).

Na busca pelo estado nativo, a proteína tenta encontrar a conformação $\mathbf{R}$ que minimiza $U(\mathbf{R})$ e maximiza $S(\mathbf{R})$. No limite de temperatura zero, a conformação de



equilíbrio estável está no mínimo de energia livre global; devido a $TS \rightarrow 0$ neste limite, o estado global mais favorável da E_{tour} é também obtido.

O problema está reduzido a encontrar o mínimo global de $F(\mathbf{R}, T)$ no limite de temperatura zero. O algoritmo da RE faz isto através de *annealing* determinístico sobre a superfície da energia livre $F(\mathbf{R}, T)$. As equações (13) e (14) têm a forma de um problema de otimização com restrições, onde U é a função otimizada, sujeito à restrição S . Tomando T como um multiplicador de Lagrange, ambos, $\mathbf{R}$ e T são simultaneamente otimizados, satisfazendo o critério de otimização com restrição:

$$\vec{\nabla}F(\mathbf{R}, T) = \vec{\nabla}U(\mathbf{R}) - T\vec{\nabla}S(\mathbf{R}, T) = 0 \quad (15)$$

$$\frac{\partial F(\mathbf{R}, T)}{\partial T} = -S(\mathbf{R}, T) - T \frac{\partial S(\mathbf{R}, T)}{\partial T} = 0 \quad (16)$$

Estes critérios podem ser satisfeitos através do método de minimização “*steepest-descent*” e “*steepest-ascent*”, respectivamente:

$$\Delta \mathbf{R}_j = -\nabla_{\mathbf{R}} F(\mathbf{R}, T) \Delta \tau \quad \text{e} \quad \Delta T = \frac{\partial F(\mathbf{R}, T)}{\partial T} \Delta \tau, \quad \text{onde } \Delta \tau \text{ é o intervalo de tempo para}$$

atualizar o valor da temperatura. Em termos das eqs (14), (15) e (16) estes passos são:

$$\Delta \mathbf{R}_j = \alpha_c \sum_{i=1}^m w_{ij} (\mathbf{S}_i - \mathbf{R}_j) - \alpha_t \nabla_{\mathbf{R}_j} E_{tour} \quad (17)$$

$$\Delta T = \gamma \left[-\sum_{i=1}^m \log \sum_{j=1}^n e^{-(\mathbf{S}_i - \mathbf{R}_j)^2 / 2T} - \frac{1}{2T} \sum_{j=1}^n \frac{\sum_{i=1}^m (\mathbf{S}_i - \mathbf{R}_j)^2 e^{-(\mathbf{S}_i - \mathbf{R}_j)^2 / 2T}}{\sum_{k=1}^n e^{-(\mathbf{S}_i - \mathbf{R}_k)^2 / 2T}} \right] \quad (18)$$

onde α_c e α_t são as constantes da equação (5), que determinam a contribuição relativa de cada passo para atualizar as forças das *city* e *tour springs*, respectivamente; e γ controla a taxa de *annealing*. O intervalo de tempo do passo $\Delta\tau$ está implícito na escala das constantes α_c , α_t e γ . O uso do *steepest-ascent* para atualizar T , ao invés do *steepest-descent*, se deve a que $F(\mathbf{R}, T)$ é monotonicamente crescente para $T \rightarrow 0$.

2.3.1 – Aplicação do Método da RE à Predição de Estrutura de Proteínas

Para aplicar o método da RE do PCV à predição de estrutura de proteínas, considera-se a cadeia da proteína como um *tour* do caixeiro. Cada ponto j do *tour*, situado na posição $\mathbf{R}_j$, representa o átomo C_α de um aminoácido. O problema em aplicar a estratégia do PCV ao enovelamento de proteínas, consiste em que, no PCV, as coordenadas das cidades são conhecidas antecipadamente, e qualquer *tour* viável deve passar por elas. Mas para o enovelamento de proteínas, não existe antecipação no conhecimento da conformação nativa. A chave desta estratégia está em reconhecer que, em altas temperaturas, o termo entrópico na equação (14) que é dependente das cidades alisa o *landscape* da energia livre. Em T alto o suficiente, este alisamento cria um *landscape* convexo, o qual evita que o sistema torne-se aprisionado em um mínimo metaestável local. Além do mais, no limite de baixas temperaturas da RE, ou em qualquer solução viável do PCV, um *tour* deve passar por cidades distintas (ou conjunto de cidades, como veremos abaixo). Conseqüentemente o efeito de S na equação (14) é similar ao da entropia de



volume excluído da teoria de polímeros. Nos estudos conformacionais por dinâmica molecular ou *annealing* estocástico, as superfícies de potenciais são extremamente rugosas, devido à inclusão de termos de caroço duro repulsivo. Estes métodos gastam a maior parte do seu tempo tentando superar barreiras de alta energia. No caso da RE, a restrição de ajuste das *city springs* reforçam o volume excluído de forma aproximada e gradativa sem aumentar a rugosidade do *landscape*.

A distribuição ideal de cidades seria aquela compatível com a estrutura da proteína nativa, e que cobre o espaço ocupado por uma proteína enovelada, consistente com o princípio de máxima entropia. Distribuímos as cidades na rede cubo-octaédrica (seção 2.3.2), a qual tem se mostrado adequada para representar os ângulos virtuais de ligações entre carbonos α consecutivos e reproduzir simetrias das distribuições dos aminoácidos em proteínas reais. Esta escolha da rede ajuda a minimizar distorções de modelos de estruturas de proteínas causadas pelo uso de redes para representar proteínas.

A grandeza S_i é a coordenada da i -ésima cidade relativo ao centro da rede que é definido pela relação $\sum_{i=1}^m S_i = 0$. Uma distribuição com baixa formatação conforme a rede terá mais cidades do que *tour points* ($m > n$). Neste caso, uma restrição de ajuste é mais geral, pois, na média, um dado monômero será puxado aproximadamente na direção do centróide das m/n cidades mais próximas (ROSE e GUREWITZ, 1990). A proteína tem os carbonos α representados pelos *tour points* que formam uma cadeia polimérica imitando o *tour* do caixeiro. As interações são devidas à existência de três forças:

- (i) As *tour springs* que representam as ligações covalentes entre os monômeros j e $j+1$;



(ii) As *city springs* que forçam um alisamento do *landscape* em forma de funil, principalmente devido ao volume excluído que é satisfeito de forma gradativa à medida que a temperatura diminui;

(iii) Molas não covalentes intramoleculares, tomadas de potenciais estatísticos (seção 2.3.3) e que ligam cada monômero da cadeia a todos os demais.

O mapeamento do PCV ao problema de busca conformacional da cadeia polimérica é feito mediante a substituição de E_{tour} , presente na equação (14), por E_{prot} , a energia interna da proteína, que, por sua vez, contém os seguintes termos:

$$E_{prot} = E_{stat} + E_{HP} + E_{cov} + E_{Rg} + E_{Hc} \quad (19)$$

$$E_{stat} = \sum_{i < j} a_{ij}^{stat} r_{ij}^2$$

$$E_{HP} = \sum_{j=1}^n a_j^{HP} r_j^2$$

$$E_{cov} = \sum_{j=1}^{n-1} a^{cov} (r_{j,j+1} - \overline{r_{j,j+1}})^2$$

$$E_{Rg} = a^{rg} (r_g - r_g^{nat})^2$$

$$E_{Hc} = \sum_{i < j} \left(\frac{r_0}{r_{ij}} \right)^{10}$$

Os valores dos coeficientes presentes nos diversos termos estão contidos na tabela 1



Tabela 1 - Parâmetros para a função energia.

Parâmetro	Intensidade Relativa
a^{stat}	0,05
a^{cov}	6,0
a^P	0,1
a^H	-0,1
a^{rg}	10,0

O valor de a^{stat} corresponde ao máximo valor do acoplamento entre dois aminoácidos não ligados, ou seja, $a^{stat} = \max |a_{ij}^{stat}|$, a^{cov} é o acoplamento referente à ligação covalente, tal que é, pelo menos, 120 vezes maior que a_{ij}^{stat} , a^P e a^H são os parâmetros para resíduos polares e hidrofóbicos e a^{rg} representa um forte acoplamento para manter fixo o raio de giro, caso seja conhecido por técnicas experimentais.

O conjunto de valores dos parâmetros usados para o modelo, está apresentado na tabela 1. Análogo ao modelo Gaussiano (RAGHUNATHAN and JERNIGAN, 1997), interações entre resíduos, não covalentes, são representadas pelo potencial estatístico E_{stat} , onde r_{ij} é a distância do aminoácido i ao j ; o coeficiente a_{ij}^{stat} é a energia potencial de contato entre os aminoácidos i e j , extraída através de métodos que fazem uso de banco de dados de estruturas conhecidas (THOMAS and DILL, 1996a). Esses coeficientes são normalizados para que seu valor máximo seja $a^{stat} = 0,05$, conforme mostra a tabela 1. Ligações peptídicas covalentes são representadas como potenciais harmônicos duros, E_{cov} cuja constante de força é a^{cov} e comprimento de equilíbrio da ligação igual a $\overline{r_{j,j+1}} = 3,8\text{\AA}$ (esta é a distância média entre carbonos α consecutivos na cadeia protéica). O potencial E_{HP} é responsável pela junção dos resíduos hidrofóbicos para dentro da estrutura, e pela segregação dos resíduos polares para a superfície da proteína. Em casos em que alguma informação experimental é disponível, adiciona-se à função energia mais uma mola. Por exemplo, se por alguma técnica se souber a distância entre dois resíduos, pode-se colocar



uma mola dura cuja distância de equilíbrio corresponde à distância encontrada experimentalmente. A técnica de espalhamento de raios X a baixo ângulo tem sido bastante utilizada nos últimos anos no estudo de proteínas em solução e tem se mostrado capaz de determinar valores de raio de giro, e dimensão máxima de moléculas por gráficos de Guinier e transformada inversa de Fourier (SOUZA, 1990). Se a informação sobre o raio de giro estiver disponível acrescenta-se na função energia o termo de potencial E_{Rg} , sendo r_g o raio de giro de qualquer conformação da cadeia, calculado por

$$r_g^2 = \frac{1}{n^2} \sum_{i < j}^n r_{i,j}^2$$

e r^{nat} o raio de giro da estrutura da proteína nativa, que são acoplados via a^{rg} . Com o intuito de verificar a influência de um termo explícito de volume excluído, incluímos o termo E_{Hc} , cujo efeito é o de excluir uma região de raio r_0 em torno de cada monômero. Fixamos o valor de $r_0 = 4 \text{ \AA}$, a idéia é comparar as simulações com e sem a presença desse termo, que, em teoria torna o landscape mais rugoso, devido às barreiras energéticas oriundas da exclusão de volume.

2.3.2 – Descrição da Rede Cubo-octaédrica

Um modelo de cela unitária desta rede é mostrado na figura 2. Cada ponto tem 12 vizinhos mais próximos. Esta unidade sólida semi-regular é referida como cubo-octaédrica (RAGHUNATHAN AND JERNIGAN, 1997), e tem 14 faces e 12 vértices. Ela é um único poliedro convexo e contém somente dois tipos de polígonos regulares, triângulos e



quadrados. As 14 faces são compostas de oito triângulos e seis quadrados. Cada triângulo está cercado por três quadrados, e cada quadrado por quatro triângulos. Existem seis pontos no plano horizontal, três acima e três abaixo deste plano. Vetores adjacentes no plano horizontal fazem um ângulo de 60° com o centro. Os três vetores do topo fazem ângulos de 120° um com o outro, e da mesma forma os três vetores de baixo. Esta rede é uma versão da rede cúbica de face-centrada, que tem conexão permitida somente entre pontos de face-centrada e 12 extremidades mais próximas. Todas conexões têm o mesmo comprimento.

Os 12 vetores são,

no plano xy: $(1, 1, 0)$, $(1, -1, 0)$, $(-1, 1, 0)$, $(-1, -1, 0)$;

no plano yz: $(0, 1, 1)$, $(0, -1, 1)$, $(0, 1, -1)$, $(0, -1, -1)$ e

no plano zx: $(1, 0, 1)$, $(1, 0, -1)$, $(-1, 0, 1)$, $(-1, 0, -1)$.

Quando conectados aos pares, estes, dão os ângulos de ligação 0° , 60° , 90° , 120° e 180° . E em conexões triplas, ângulos de torção 0° , $54,7^\circ$, $70,5^\circ$, $109,5^\circ$, $125,3^\circ$ e 180° . Quando expressados desta forma, os vetores unitários são mais simples de serem entendidos. O conjunto de pontos na figura 2, quando repetidos, formam um arranjo periódico de pontos da rede no espaço. Um número infinito de camadas podem ser crescidas a partir do centro, na direção radial. O número de pontos sobre a superfície de qualquer camada é dado pela relação:

$$N = 10L^2 + 2 \quad (20)$$

Onde L é o número da camada. Portanto, a primeira camada tem 12 pontos, a segunda 42, a terceira 92, e assim por diante. Este modelo oferece umas das melhores formas para inspecionar visualmente traços na cadeia da proteína, bem como estereoquímica da cadeia



lateral em relação à cadeia principal, a estrutura secundária e outros detalhes geométricos de estruturas de proteínas (RAGHUNATHAN e JERNIGAN, 1997).

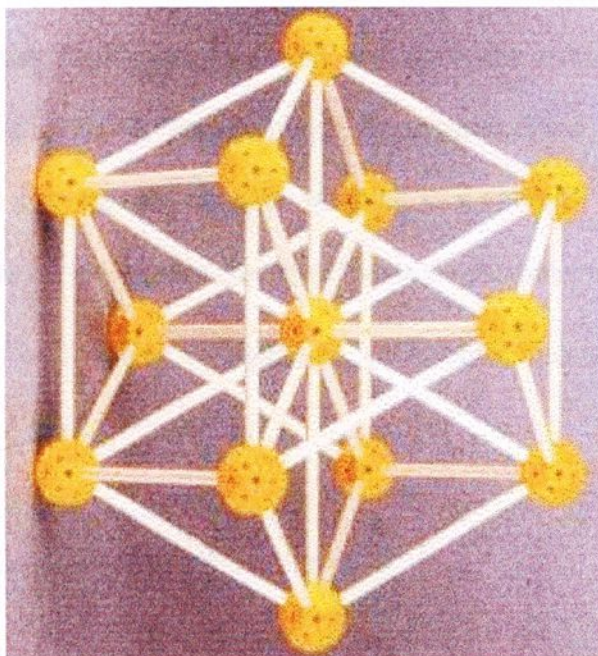


Figura 2 – Ilustração do modelo de uma célula unitária da rede proposta.

2.3.3 – Um Método Iterativo para Extrair Potenciais a partir de Estruturas de Proteínas Proposto por Thomas, Dill e Colaboradores

Descreveremos o método para obter potenciais de interações entre resíduos, a partir de estruturas de proteínas no *Protein Data Bank* (PDB). Potenciais obtidos de banco de dados são quantidades tipo-energia, amplamente usados em algoritmos computacionais, que objetivam de algum modo, prever a estrutura de uma proteína a partir de sua sequência de aminoácidos (THOMAS e DILL, 1996b). Existem vários métodos para obter

potenciais de contato entre pares de resíduos das estruturas conhecidas de proteínas. Muitos deles usam a relação de Boltzmann para a frequência de contato entre aminoácidos observados em estruturas conhecidas, a fim de obterem quantidades tipo energia. No entanto, existem alguns problemas na forma em que estes potenciais estatísticos tipo energia, são obtidos a partir da frequência de pareamento. Para descrever o problema principal, definem-se dois termos: energias verdadeiras e energias extraídas. As energias extraídas são baseadas na observação das frequências de pareamento, ρ_{ij} , do aminoácido tipo i com o tipo j. Toma-se o logaritmo para obter quantidades tipo-energia (adimensionais), e_{ij} ,

$$e_{ij} = \text{energia}/kT = -\ln\left(\frac{\rho_{ij}}{\rho_{ij}^*}\right) \quad (21)$$

ρ_{ij}^* representa a frequência de pareamento no estado de referência, k é a constante de Boltzmann e T é a temperatura absoluta. A energia verdadeira é definida como a energia de interação livre real, E_{ij} , que a natureza utiliza para conduzir a proteína ao enovelamento. Se a suposição e o procedimento dos potenciais estatísticos forem perfeitos, então os potenciais extraídos igualar-se-iam às energias verdadeiras, $e_{ij} = E_{ij}$. Para proteínas reais estas energias não são conhecidas precisamente. Usando um modelo de rede, com energias verdadeiras conhecidas para testar a confiabilidade do procedimento de extração, verifica-se que, energias extraídas por métodos padrões que utilizaram potenciais estatísticos não se igualam às energias verdadeiras (THOMAS e DILL, 1996a).

O problema destes potenciais estatísticos surge de um descuido quando não se leva em conta o volume excluído, conectividade da cadeia ou a sequência dos aminoácidos. Por



exemplo, os métodos de extração padrão assumem que a frequência de pareamento da alanina com o triptofano é independente da frequência de pareamento da glicina com a lisina, ou em geral, que a frequência de pareamento dos resíduos ij é independente da frequência de pareamento dos tipos kl . No entanto isso não é verdadeiro. Em moléculas de cadeia compacta tais como proteínas globulares, a propensão de certo aminoácido para formar um contato com outro, afeta as posições de outros aminoácidos na cadeia. Como resultado, energias estatísticas extraídas podem ter magnitude e até sinais incorretos.

A origem destes problemas está na construção dos estados de referência ρ_{ij}^* na equação (21). Normalmente, estes estados de referência são baseados em glóbulos randômicos de aminoácidos desconectados. Consertar estes problemas requer mais precisão, levando em conta volume excluído, seqüências e conectividade da cadeia no estado de referência. Nesse sentido, esforços têm sido feitos para obter potenciais otimizados de conformações de proteínas no estado nativo e não-nativo (MAIOROV e CRIPPEN, 1992).

No procedimento conhecido como ENERGI (THOMAS e DILL, 1996a), propõe-se um método iterativo para melhorar a obtenção de potenciais estatísticos, no qual um *ensemble* de estruturas no estado de referência da equação (21) é iterativamente ajustado para refletir a melhor aproximação para as interações energéticas. Neste método destacam-se duas características essenciais:

- (i) O estado de referência é uma cadeia totalmente conectada, com a seqüência de aminoácidos correta em diferentes conformações.
- (ii) Como o estado de referência não pode ser descrito por uma expressão analítica, este método envolve iterações.



Os parâmetros ou escores tipo-energia e_{ij} são ajustados em cada iteração, usando a taxa de pareamento entre o *ensemble* de estrutura observada e a predita. Os parâmetros convergem quando eles predizem que a estrutura (nativa) observada é suficientemente favorecida energeticamente sobre as estruturas alternativas. Um fluxograma é apresentado na figura 2. Este método inicia-se com um primeiro valor estimado para o potencial estatístico na matriz, $e_{ij}^{(0)}$, onde o sobrescrito (0) indica que este é o valor da aproximação zero, e e_{ij} é o parâmetro tipo energia para o pareamento dos aminoácidos do tipo i e j. Calcula-se a média de Boltzmann (sobre conformações diferentes) da frequência de pareamento usando esta função peso. O processo iterativo se caracteriza pelo uso da medida Δe_{ij} , para o erro entre as energias verdadeira e a energia estimada:

$$\Delta e_{ij}^{(n)} = -\ln \left(\frac{\rho_{ij}(\text{observada})}{\rho_{ij}^{(n)}(\text{predita})} \right) \quad (22)$$

Δe_{ij} é equivalente a eq (21) usada para obter potenciais estatísticos, onde são usadas médias de Boltzmann para a frequência de pareamento como o estado de referência. Os escores da primeira aproximação são iguais à aproximação zero mais a correção:

$$e_{ij}^{(1)} = e_{ij}^{(0)} + \Delta e_{ij}^{(0)}$$

E em geral:

$$e_{ij}^{(n)} = e_{ij}^{(n-1)} + \Delta e_{ij}^{(n-1)} \quad (23)$$

Calcula-se novamente a média de Boltzmann usando o novo valor e_{ij} e repete-se o ciclo até que ele convirja, ou seja, até que $|\Delta e_{ij}|$ seja menor que uma tolerância pré-estabelecida. Desde que este processo convirja e seja adequadamente parametrizado,



obtem-se um conjunto final de $e_{ij}^{(n)}$ que pode discriminar totalmente, para uma dada sequência, seu enovelamento correto entre um conjunto de conformações. Especificamente, se a conformação nativa é significativamente favorecida por uma função escore sobre todas as conformações alternativas, ela predominará no ensemble. Se isso for simultaneamente verdadeiro para todas as conformações diferentes no banco de dados, Δe_{ij} aproximará de zero e o processo convergirá. O processo de determinar os potenciais estatísticos de contato, é equivalente a começar com uma aproximação zero, na qual todas interações $e_{ij}^{(0)}$ são assumidas sendo iguais e favoráveis, e calcular o termo da primeira correção $\Delta e_{ij}^{(0)}$. A matriz de escore ENERGI obtida dessa forma por Thomas e colaboradores está listada na tabela 2.



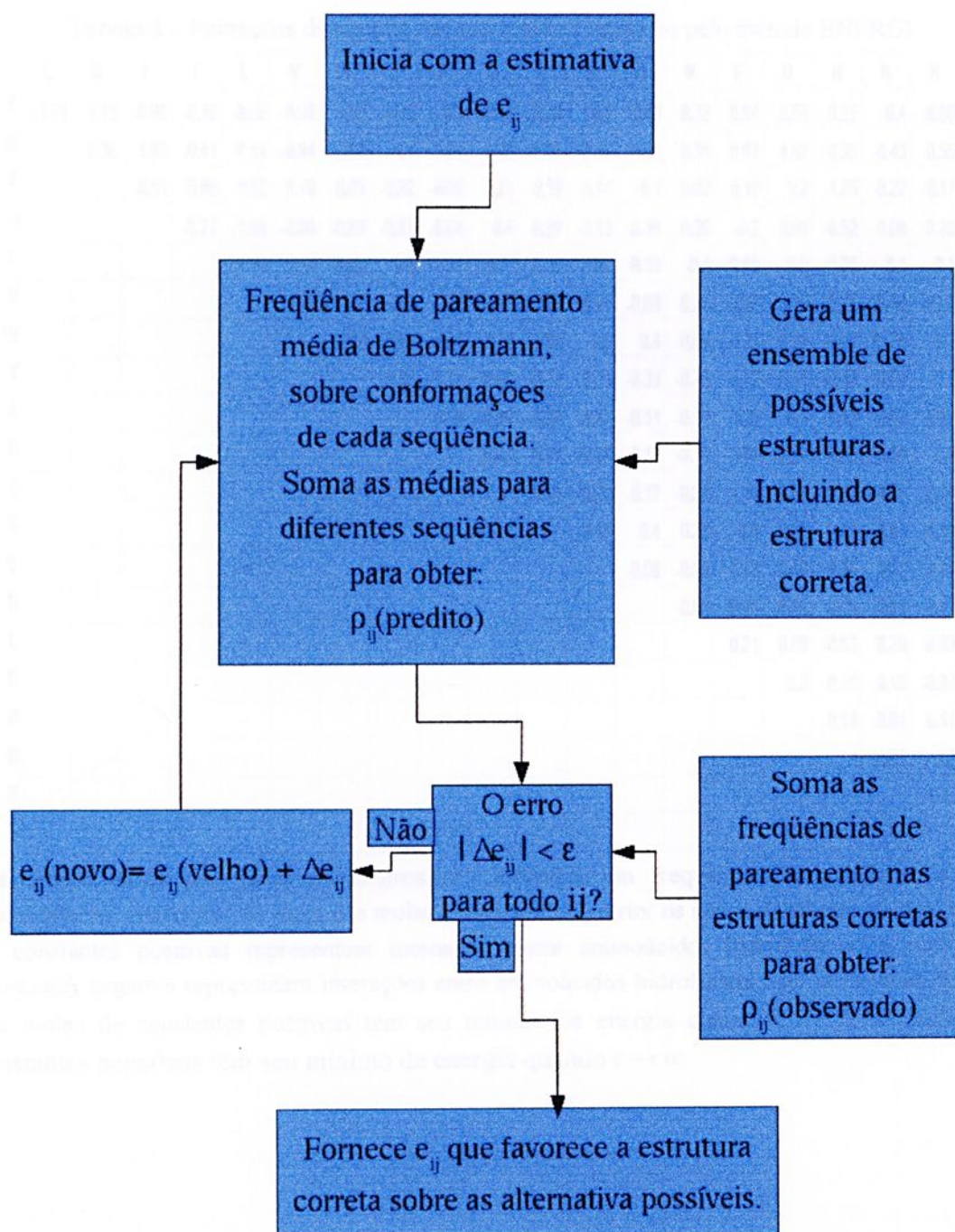


Figura 3 – Diagrama de fluxo do método ENERGI aplicado a interações de pares de aminoácidos em proteínas.

Tabela 2 – Interações de contato resíduo-resíduo extraídos pelo método ENERGI

	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	-1.79	-1.23	-0.98	-0.48	-0.69	-0.94	-0.3	-0.96	-0.3	-0.42	-0.38	-0.2	-0.49	-0.32	0.04	0.55	-0.82	-0.4	0.00	0.07
M		0.36	-1.03	-0.41	-0.31	-0.94	-0.07	-1.1	0.05	0.00	0.06	-0.47	-0.54	0.31	0.02	1.07	-0.35	-0.43	0.55	-0.25
F			-0.61	-0.66	-1.02	-0.78	-0.89	-0.82	-0.05	0.21	-0.19	0.14	0.1	-0.02	0.19	0.2	-1.75	-0.22	-0.17	-0.43
I				-0.71	-1.04	-0.98	-0.89	-0.87	-0.64	0.4	-0.29	-0.13	-0.39	0.39	-0.2	0.04	-0.52	-0.08	-0.26	0.25
L					-1.14	-1.03	-0.97	-0.6	-0.57	-0.08	-0.39	-0.07	-0.13	-0.1	-0.05	0.5	-0.36	-0.1	0.1	0.09
V						-1.15	-0.6	-0.7	-0.6	-0.2	0.06	-0.31	-0.09	-0.24	-0.02	0.25	-0.35	-0.48	-0.08	-0.08
W							0.02	-0.99	-0.08	-0.14	0.07	-0.2	0.4	-0.68	0.32	0.24	-0.41	-0.78	-0.3	-0.44
Y								0.35	-0.37	-0.32	-0.23	0.25	-0.39	-0.74	0.22	0.11	-0.67	0.21	-0.2	-0.45
A									-0.08	-0.09	-0.22	-0.01	-0.11	-0.14	0.03	0.1	-0.15	0.07	0.00	0.41
G										0.04	0.13	-0.04	0.12	-0.18	0.4	-0.06	0.00	0.15	0.1	0.4
T											0.26	0.05	-0.17	-0.27	0.15	-0.03	-0.27	-0.17	0.09	0.36
S												-0.13	0.4	0.37	0.3	-0.09	-0.59	0.61	0.18	0.44
Q													-0.08	-0.05	0.62	0.46	0.05	0.62	0.04	-0.21
N														-0.86	-0.25	-0.12	0.06	0.04	0.18	0.11
E															0.21	0.68	-0.53	-0.26	-0.09	0.33
D																0.6	-0.06	-0.15	-0.09	0.84
H																	0.14	-0.01	0.14	-0.22
R																		0.23	0.3	-0.02
K																			1.45	0.51
P																				0.28

*Importante notar que esses parâmetros são baseados em frequências de contatos, e para representarem constantes de força das molas é necessário inverter os sinais da tabela, ou seja, molas de constantes positivas representam interações entre aminoácidos hidrofóbicos, e molas de constantes negativa representam interações entre aminoácidos hidrofílicos. Isto se deve ao fato de que molas de constantes positivas tem seu mínimo de energia quando $r \rightarrow 0$, e molas de constantes negativas têm seu mínimo de energia quando $r \rightarrow \infty$.

CAPÍTULO 3

RESULTADOS E DISCUSSÕES

Método da RE Aplicado à Predição de Estruturas de Proteínas

A predição de estrutura foi implementada através da rede cubo-octaédrica, em cujos sítios são colocadas as cidades. O número de camadas, L , da rede é escolhido de forma a abranger a estrutura da proteína nativa, cujos raios de giro variam de 8 a 17 Å. Como cada aminoácido está conectado com todas as cidades, quanto maior o número de cidades, maior será o tempo computacional na simulação. Por isso o número de camadas deve ser o menor possível, que ainda abranja as estruturas desnaturadas ou nativas. Análogo à caixa de água usada em dinâmica molecular, que deve ser a menor possível, porém que hidrate toda a cadeia. As distâncias entre os pontos da rede são normalizadas em 3,8 Å, que é a distância entre carbonos α consecutivos em proteínas. Desta forma, obedecendo a restrição de ajuste do PCV, em que cada cidade deve ser visitada somente uma vez, quando um resíduo for atraído por uma cidade, o seu vizinho na sequência será atraído por outra cidade mais próxima que está a 3,8 Å. Por essa razão a estrutura da rede pode polarizar a estrutura da cadeia. Para minimizar esse efeito Ball e colaboradores (Ball, Erman e Dill, 2002), propõem colocar cidades somente sobre uma fração dos sítios da rede, determinada por



uma distribuição Gaussiana $p_{\text{sítio}} \propto \exp\left[-(r/\rho)^2\right]$, onde $\rho = 10\text{\AA}$. Com essa distribuição pretende-se reproduzir, parcialmente, a densidade de aminoácidos em proteínas que é maior para regiões mais internas (menor r) devido às interações de compactação.

Aqui aparece a primeira diferença entre a proposta da literatura e o presente trabalho. Para reduzir a polarização da estrutura da rede na estrutura final da cadeia, optamos por incluir cidades sobre todos os sítios da rede e enfraquecer a constante de interação α_c dos aminoácidos com as cidades. O valor de α_c usado nas presentes simulações foi de 50 a 500 vezes menor do que α_t . Devido a instabilidades numéricas, o maior valor de α_t que o algoritmo suporta é $\alpha_t = 10^{-2}$. Qualquer outro valor maior do que este, provoca grandes deslocamentos, o que aumenta o potencial E_{cov} . Fixamos este valor para α_t em todas simulações, e variamos o acoplamento com as cidades, α_c . O valor de α_c está relacionado com a estrutura primária. Se a proteína for constituída em sua maior parte por aminoácidos hidrofílicos, o valor de α_c deve ser alto o suficiente para manter os resíduos dentro dos limites da rede. Por outro lado, se a maior parte dos aminoácidos for hidrofóbica, o valor de α_c pode ser baixo, uma vez que a restrição de volume excluído, E_{Hc} , esteja presente na equação (19) que representa a energia interna da proteína.

Vale notar que, o primeiro termo da equação (19), E_{stat} , denominado de potencial estatístico, utiliza os parâmetros de contato extraídos, conforme tabela 2. Fizemos também a normalização destes parâmetros para que o maior valor da interação entre dois resíduos, não consecutivos, na proteína seja de 0,05.

A segunda diferença desse trabalho daquele de Ball e colaboradores diz respeito ao termo de potencial E_{HP} , cujo efeito é o de aumentar a presença de resíduos hidrofóbicos no



interior da estrutura e mantendo os hidrofílicos na superfície. Acreditamos que esse termo é bastante inespecífico, o que poderia prejudicar a encontrar estrutura tão específica como é a estrutura nativa. Além disso, é possível que esse termo introduza alguma frustração, pois a ligação covalente mantém juntos os resíduos polares e hidrofóbicos que são consecutivos na sequência protéica torna impossível satisfazer a ida dos hidrofóbicos para o interior da cadeia e manter os hidrofílicos na superfície. Daí a conjectura de que o termo E_{HP} pode se mostrar altamente frustrado. Além disso, o potencial estatístico E_{stat} , que é baseado na frequência de contato entre os diferentes aminoácidos, já deve expressar o fato que resíduos hidrofóbicos possuem alta frequência de contato, quando comparados com a frequência entre resíduos hidrofílicos. Em várias simulações fizemos o teste de incluir e retirar o termo E_{HP} . De fato, verificamos que quando retiramos este termo da equação (19), sempre diminuía o RMS da estrutura predita. Dessa forma, optamos por não incluir este termo nos resultados apresentados na Figura 4. Até o momento não conseguimos compreender as razões que levaram aqueles autores a introduzir o termo E_{HP} e tampouco encontramos simulações na literatura que comparassem os resultados na presença e na ausência do termo hidrofóbico-hidrofílico. Nesse sentido, essa é uma contribuição do presente trabalho.

Aplicamos o algoritmo da RE para um conjunto de 20 proteínas do PDB, com estrutura primária variando de 36 a 136 aminoácidos, listadas na tabela 3. Executamos o processo de annealing decrescendo a temperatura T do sistema de $T = 70$ até $T = 1$. Para cada valor da temperatura, são realizadas em torno de 100 a 200 iterações. O critério para decidir quantas iterações serão realizadas em cada T baseia-se no raio de giro da estrutura na temperatura T . Quando o algoritmo não detecta variações significativas no raio de giro,



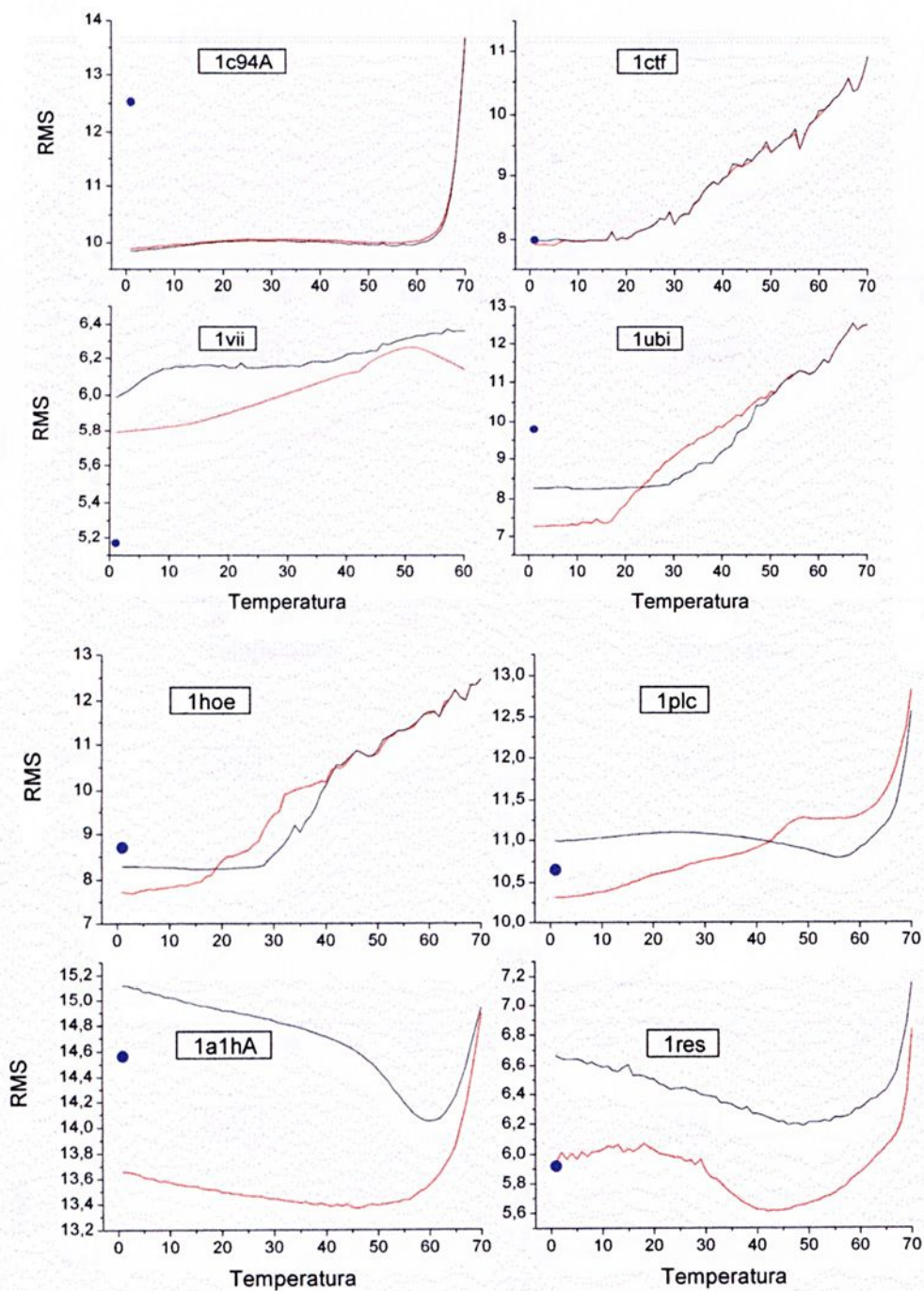
interrompe as iterações e reduz-se a temperatura. Nessa nova temperatura começa-se novamente o ciclo das iterações e assim por diante até $T = 1$.

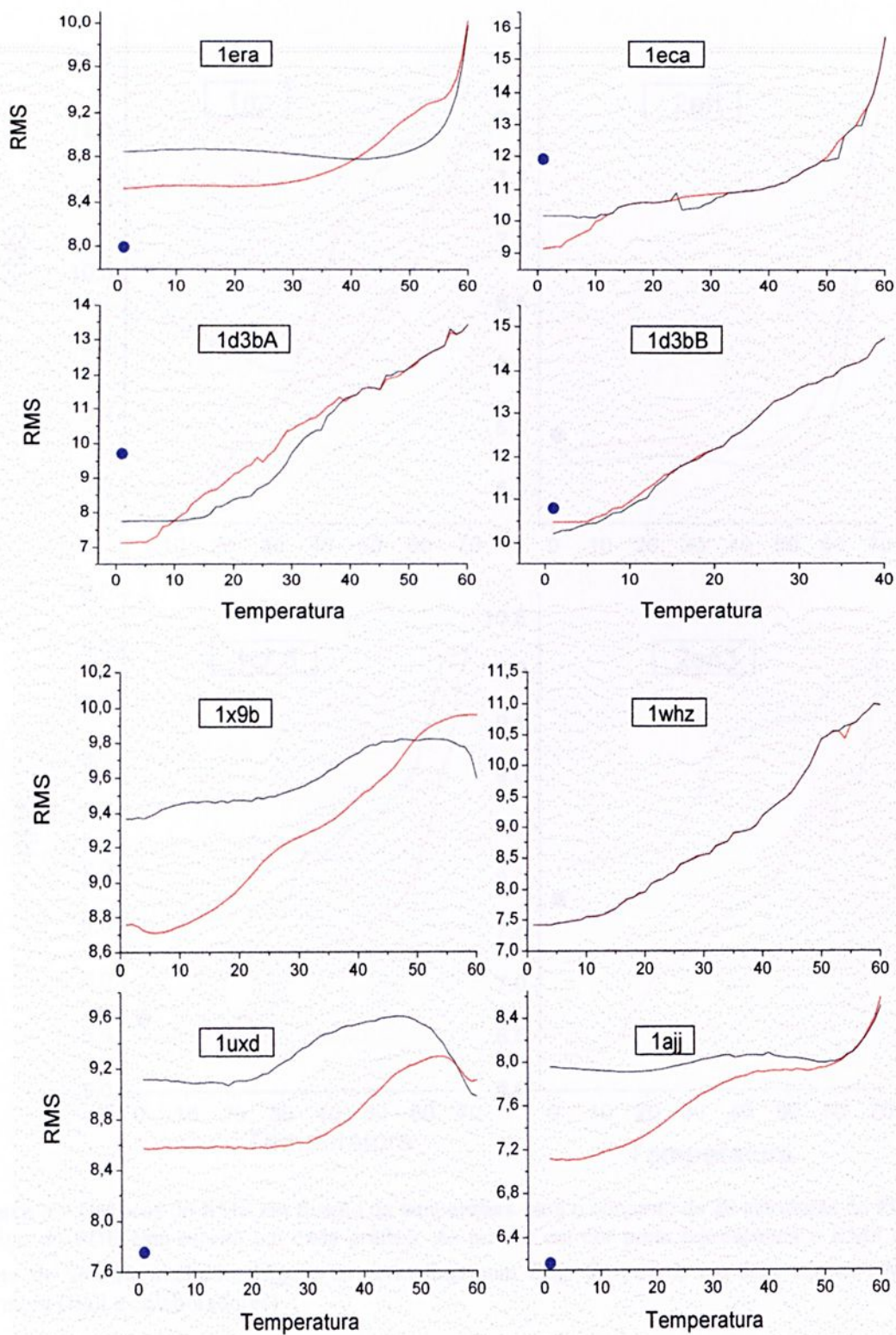
Ao optarmos por enfraquecer o acoplamento α_c dos aminoácidos com as cidades, enfraquecemos o atendimento a dois requisitos fundamentais para a obtenção de boas estruturas: a manutenção da distância entre dois carbonos α consecutivos em $3,8 \text{ \AA}$ (que é a distância entre duas cidades vizinhas), e a condição de volume excluído, pois agora, a tendência de cada aminoácido ocupar uma cidade distinta está, também, enfraquecida. Para corrigir esses dois aspectos fundamentais implementamos duas mudanças: aumentamos de cinco a seis vezes o valor da constante covalente, a^{cov} , presente no termo E_{cov} , que é minimizado quando a distância entre dois carbonos α consecutivos é $3,8 \text{ \AA}$ e introduzimos explicitamente o termo de volume excluído E_{Hc} , presente na equação (19). Fizemos as simulações com e sem a presença do termo de volume excluído e comparamos com os resultados da literatura. Vale lembrar novamente que nos trabalhos de Ball e colaboradores a constante de acoplamento com as cidades é significativamente maior e em baixas temperaturas a entropia é maximizada (energia livre minimizada) mediante a ocupação de cada cidade por um único aminoácido. Estamos, portanto, comparando diferentes estratégias para se implementar o volume excluído e também manter a distância das ligações covalentes. No mencionado trabalho o termo tipo entrópico é maximizado sempre que, em baixas temperaturas, cada monômero (representado pelo *tour point*) ocupar uma cidade distinta, portanto, esse termo entrópico faz o papel do potencial de volume excluído. No presente trabalho, uma vez que a constante de acoplamento com as cidades foi enfraquecida, optamos por satisfazer o volume excluído incluindo o termo de caroço duro.



Apresentamos na figura 4 os gráficos do RMS, calculado pelo programa PROFIT V2.2, em função da temperatura. Para cada proteína fizemos a predição com e sem o potencial E_{Hc} . Verificamos que na maioria das simulações obtivemos menor valor de RMS quando este termo estava incluído na energia interna. Apresentamos também na figura 4 os valores de RMS da literatura em temperatura, $T = 1$. As proteínas 1x9b, 1whz e 1ttz que foram escolhidas do recente CASP 2004, para complementar o conjunto de 17 estruturas da literatura, apresentam somente as curvas de RMS da predição feita pela RE. Em seguida, para melhor comparação destes resultados com a literatura, mostramos os valores de RMS, em $T = 1$, na tabela 3.







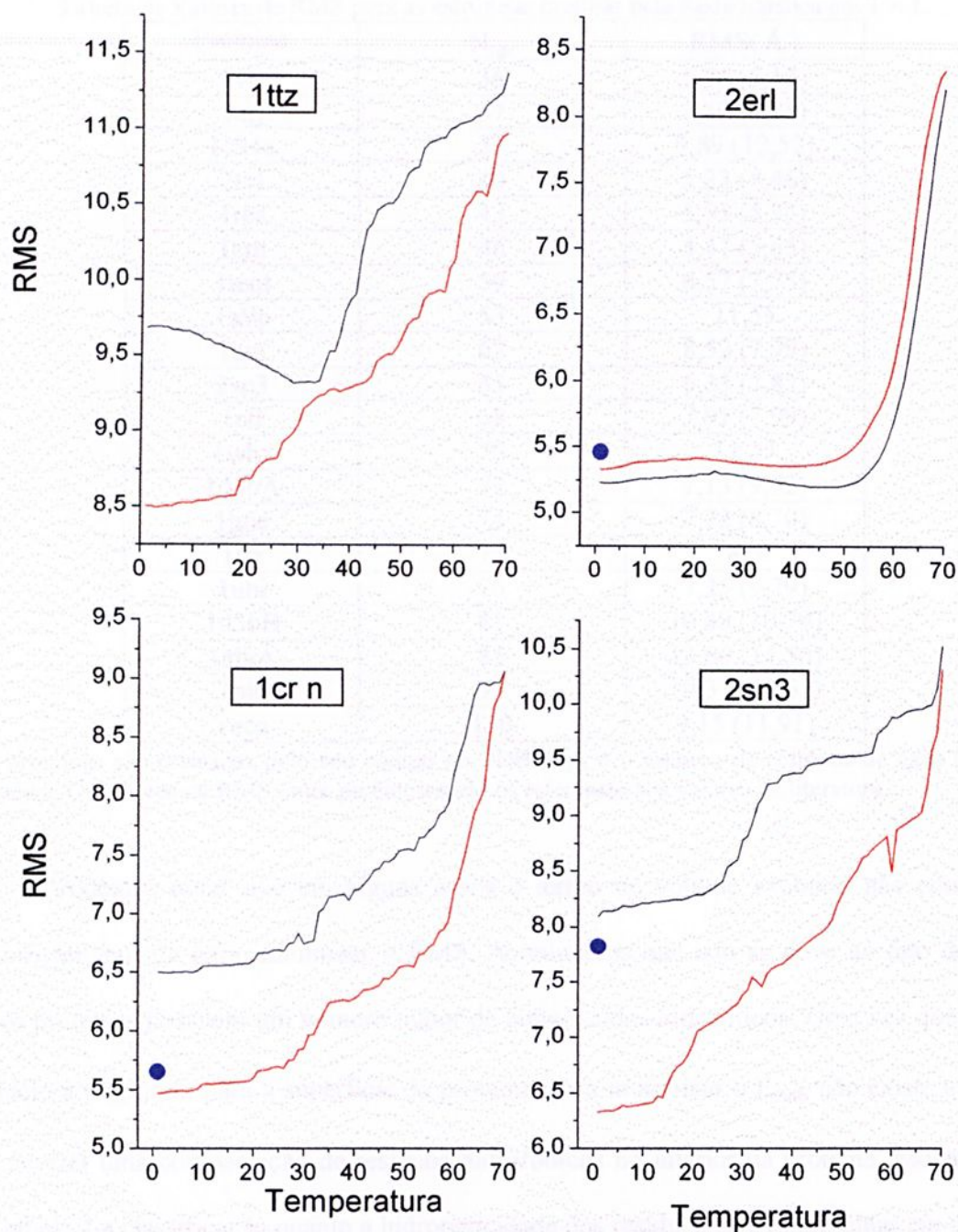


Figura 4 – Gráficos do RMS em função da temperatura para o conjunto de 20 estruturas do PDB. O código do PDB está escrito em cada gráfico. As curvas em cor preta representam o RMS sem o termo de volume excluído, E_{Hc} , e as vermelhas com E_{Hc} . Os pontos em azul são os RMS da literatura (Ball e colaboradores).

Tabela 3: Valores de RMS para as estruturas preditas pela Rede Elástica em $T = 1$.

Proteína	N_{res}	RMS(Å)
lvii	36	5,76 (5,17)
lajj	37	7,12 (6,16)
lc94a	37	9,89 (12,52)
2erl	40	5,33 (5,46)
lres	43	5,95 (5,92)
lcrn	46	5,47 (5,65)
luxd	59	8,57 (7,75)
lx9b	53	11,55
lera	62	8,52 (7,99)
2sn3	65	6,33 (7,82)
lctf	68	7,91 (7,99)
lwhz	70	11,27
ld3bA	72	7,13 (9,72)
lhoe	74	7,72 (8,70)
lttz	75	8,5
lubi	76	7,25 (9,79)
ld3bB	81	10,48 (10,79)
lalhA	85	13,66 (14,56)
lplc	99	10,31 (10,65)
leca	136	9,15 (11,91)

As proteínas são rotuladas pelo seu código no PDB. N_{res} é o número de resíduos de cada cadeia protéica. Os valores de RMS entre parênteses são os referentes aos valores da literatura.

Podemos notar que em alguns casos o termo de volume excluído não contribui significativamente para minimizar o RMS. Acreditamos que isto se deve ao fato de que estas proteínas possuem um número maior de aminoácidos hidrofílicos. Uma vez que estes tendem a migrarem para a superfície da proteína, para minimizar o E_{stat} , não existe o risco de ocorrer uma compactação de resíduos hidrofóbicos no interior da proteína. Mostramos na tabela 4 a classificação quanto a hidropaticidade dos resíduos de três proteínas em que as curvas de RMS com e sem a presença de E_{Hc} se assemelham.



Tabela 4: Classificação dos resíduos de 3 proteínas que apresentaram curvas semelhantes de RMS

Proteína	1c94A	1whz	1d3bB
N _{res} Hidrofóbicos	12	31	33
N _{res} Hidrofilicos	25	39	48
N _{res} Total	37	70	81

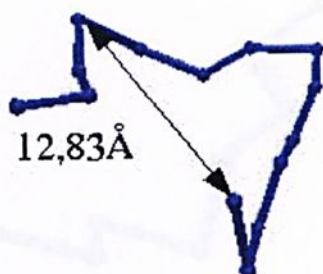
As proteínas são rotuladas pelo seu código no PDB. N_{res} é o número de resíduos.

Podemos verificar na figura 5 que quando incluímos o potencial E_{Hc} as distâncias entre os aminoácidos destacados nos trechos das estruturas apresentadas, estão mais próximas das distâncias apresentadas na estrutura nativa. Na ausência do termo E_{Hc} , o algoritmo, compacta e até mesmo sobrepõe os aminoácidos hidrofóbicos no interior da estrutura desrespeitando o princípio entrópico. Tal fato ocorreu, por exemplo, com a predição da estrutura da proteína 1eca, visualizada na figura 6. A sobreposição destes aminoácidos não ocorre quando incluímos o potencial E_{Hc} . Assim estes fatores contribuíram para minimizar o RMS das estruturas preditas.

Em termos de RMS nossos resultados são comparáveis com os da literatura e, em certos casos até melhores. Apresentamos na figura 7 três sobreposições de estruturas preditas pelo método da RE com suas respectivas estruturas nativas.



1x9b com E_{Hc}



1x9b sem E_{Hc}



1x9b nativa

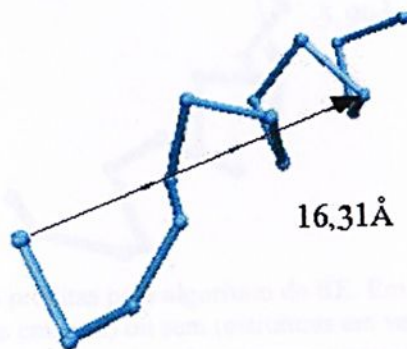
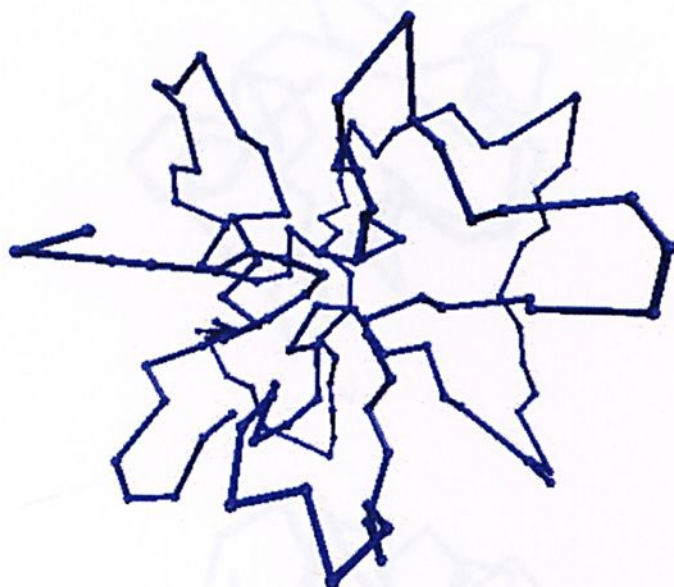




Figura 5: Trechos de estruturas preditas pelo algoritmo da RE. Em cada figura está escrito o código PDB da proteína com (estruturas em azul) ou sem (estruturas em verde) o termo de volume excluído incluído na predição, e também alguns trechos das estruturas nativas (azul claro). Pode-se notar que o termo de volume excluído E_{Hc} aproxima as distâncias ao valor nativo.

1cca com E_{Hc}



1cca sem E_{Hc}



Figura 6 – Estruturas da proteína 1cca previstas pelo algoritmo da RE. Em cada figura está escrito o código PDB da proteína com (estruturas em azul) ou sem (estruturas em verde) o termo de volume excluído incluído na predição. Nota-se o aglomerado central na estrutura em verde, com várias distâncias não nativas.

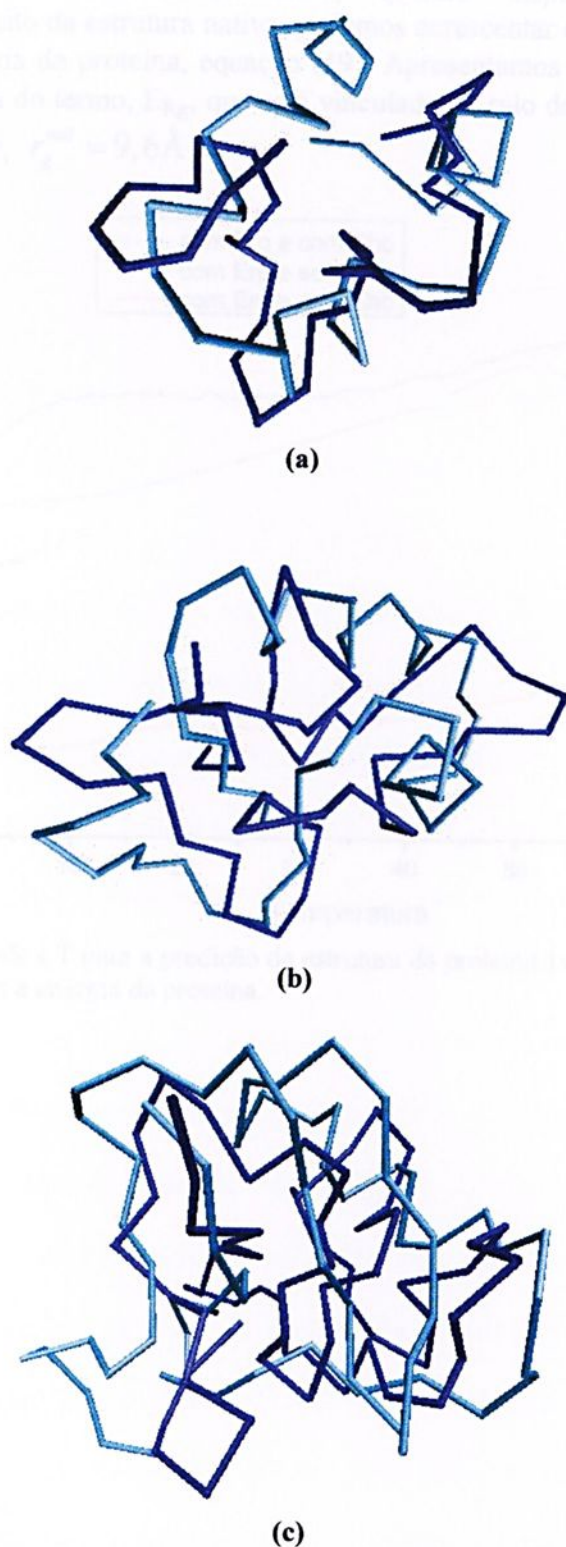


Figura 7 – Sobreposição de estruturas preditas pela RE (azul) com as estruturas nativas (azul claro).
Código do PDB: (a) 1vii, (b) 1crn e (c) 2sn3.

Conforme afirmamos anteriormente, quando dispomos de informações experimentais a respeito da estrutura nativa, podemos acrescentar esta informação no termo que descreve a energia da proteína, equação (19). Apresentamos na figura 8 as curvas de RMS com a presença do termo, E_{Rg} , que está vinculado ao raio de giro da proteína 1vii no estado nativo, a saber, $r_g^{nat} = 9,6\text{\AA}$.

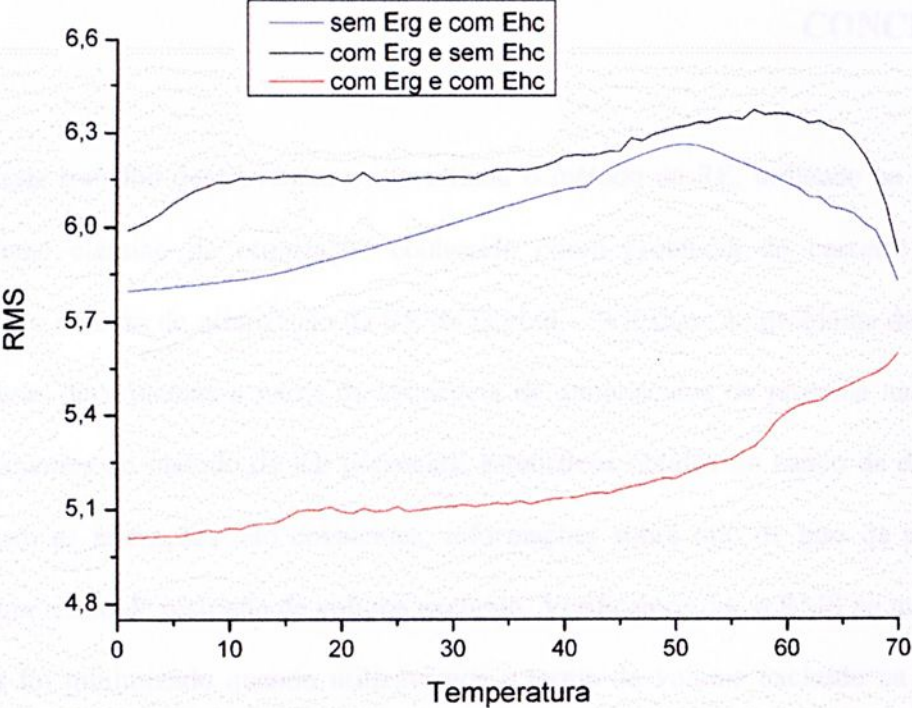


Figura 8: Curvas de RMS x T para a predição da estrutura da proteína 1vii, feita pela RE, variando os termos que descrevem a energia da proteína.

CONCLUSÃO

Neste trabalho descrevemos como utilizar o método da RE, utilizado na resolução do problema clássico de otimização conhecido como problema do caixeiro viajante. Aplicamos o método de otimização da RE de Durbin – Willshaw ao problema de predição de estruturas de proteínas a partir da sequência de aminoácidos da proteína nativa. Para isso, associamos ao método da RE potenciais estatísticos obtidos do banco de dados, que representam as interações não covalentes, informações sobre raio de giro da proteína e também um termo de restrição de volume excluído. Verificamos que o RMS na maioria das predições foi minimizado quando utilizávamos o termo de volume excluído na predição. Uma questão que pode ser levantada é “por que não resolvemos a influência da rede colocando acoplamento zero com as cidades?” Na literatura a influência da estrutura da rede é diminuída através da remoção de algumas cidades. Nós optamos por enfraquecer a constante de acoplamento. A razão de não colocá-la zero se deve a que o potencial hidrofílico tem mínimo na separação infinita entre dois resíduos hidrofílicos. Embora a interação covalente seja bastante intensa em relação ao potencial estatístico, existe $n-1$ molas covalentes, porém são n^2 molas do potencial estatístico. Ao tentar minimizar a interação hidrofílica o algoritmo tenta colocar dois resíduos o mais distante possível. Fizemos o acoplamento zero e notamos que a estrutura é esticada de forma constante. Portanto, algum acoplamento com as cidades é necessário para estruturas aceitáveis.



Finalmente vale ressaltar que vários resultados de RMSD da literatura contêm potenciais de restrição baseados em dados experimentais, como distâncias entre dois resíduos. Nossos resultados não incluem esses dados que, se incluídos poderiam melhorar significativamente a qualidade da predição.

REFERÊNCIAS BIBLIOGRÁFICAS

BALL, K. D.; DEMAR, B.; DELL, K. A. The Elastic Net Algorithm and Protein Structure Prediction. *Journal of Computational Chemistry*, v. 21, n. 1, p. 77-81, 2002.

BEPO, I. VAN DER GIESSEN, J. H. An Analysis of various Elastic Net Algorithms. *Technical Report ESW-FRW-15-25-06, Eindhoven University of Technology, Faculty of Technology, Dept. Of Computer Science, August 1995.*

CERONE, A. et al. Department of Systems and Computer Science, University of Florence. Deposited on <http://www.ist.unifi.it/eng/ai/elastic/elasticStructurePrimitives.html>, 1992.

CREIGHTON, T. H. *Proteins, Structures and Molecular Properties*. Ed. W. H. Freeman and Company, New York, Second Edition, 1992.

CUNNING, R.; WILLISAW, D. The Analogous Approach to the Traveling Salesman Problem Using an Elastic Net Method. *Nature*, v. 326, p. 689-691, 1987.



REFERÊNCIAS BIBLIOGRÁFICAS

BALL, K. D.; ERMAN, B.; DILL, K. A. The Elastic Net Algorithm and Protein Structure Prediction. *Journal of Computational Chemistry*, v. 23, n. 1, p. 77-81, 2002.

BERG, J. VAN DEN; GESELSCHAP, J. H. An Analysis of various Elastic Net Algorithms. *Technical Report EUR-FEW-CS-95-06, Erasmus University Rotterdam, Faculty of Economics, Dept. Of Computer Science, August 1995.*

CERONI, A. et al. Department of Systems and Computer Science, University of Florence. Disponível em <<http://www-dsi.eng.unifi.it/neural/ProteinStructure/proteomics.html>>, 2003.

CREIGHTON, T. E. *Proteins: Structures and Molecular Properties*. Ed. W. H. Freeman and Company, New York, Second Edition, 1993.

DURBIN, R.; WILLSHAW, D. The Analogue Approach to the Travelling Salesman Problem Using an Elastic Net Method. *Nature*, v. 326, p. 689-691, 1987.



MAIOROV, V. N.; CRIPPEN, G. M.; Contact Potential that Recognizes the Correct Folding of Globular Proteins. *Journal Molecular Biological*, v. 227, p. 786-888, 1992.

MIYAZAWA, S.; JERNIGAN, R. L. Residue-Residue Potentials with a Favorable Contact Pair Term and na Unfavorable High Packing Density Term, for Simulation and Threading. *Journal Mol. Biol.*, v. 256, p. 623-644, 1996.

RAGHUNATHAN, G.; JERNIGAN, R. L. Ideal Architecture of Residue Packing and Its Observation in Protein Structures. *Protein Sciense*, v. 6, p. 2072-2083, 1997.

ROSE, K.; GUREWITZ, E.; FOX, G. C. Statistical Mechanics and Phase Transitions in Clustering. *Physical Review Letter*, v. 65, n. 6, p. 945-949, 1990.

SOUZA, C. F. Espalhamento de Raios X a Baixo Ângulo: Estudo de Hemoglobinas Extracelulares em Solução. Tese de doutorado, *Instituto de Física "Gleb Wataghin" UNICAMP*, 1990.

THOMAS, P. D.; DILL, K. A. An Iterative Method for Extracting Energy-like Quantities from Protein Structures. *Proc. Natl. Acad. Sci. USA*, v. 93, p. 11628-11633, Outubro 1996a.

THOMAS, P. D.; DILL, K. A. Statistical Potential Extracted from Protein Structures: How Accurate Are They? *Journal Molecular Biological*, v. 257, p. 457-469, 1996b.



VOET, D.; VOET, D.; PRATT, C.W. Fundamentos de Bioquímica. Ed. Artmed, 1^a reimpressão, 2002.

YUILLE, A. L.; KOSOWSKY, J. J. Statistical Physics Algorithms That Converge. *Neural Computation*, v. 6, p. 351-356, 1994.



