UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM
COMPUTAÇÃO APLICADA

ANTONIO CARLOS STUMPF SOUTO

Uso de Redes Neurais Artificiais na Simulação Monte Carlo Aplicado ao Problema de Dobramento de Proteínas

Monografia apresentada à
Universidade do Vale do Rio dos Sinos
como requisito parcial para a obtenção do título de
Mestre em Computação Aplicada

Prof. Dr. Adelmo Luis Cechin
Orientador

São Leopoldo
julho de 2006

Ficha catalográfica elaborada pela Biblioteca da
Universidade do Vale do Rio dos Sinos

S728u Stumpf Souto, Antonio Carlos

Uso de redes neurais artificiais na simulação Monte Carlo aplicado ao problema de dobramento de proteínas / por Antonio Carlos Stumpf Souto. — 2006.

130 f.: il. ; 30cm

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos. Programa Interdisciplinar de Pós-Graduação em Computação Aplicada. 2006.

"Orientação: Prof. Dr. Adelmo Luis Cechin, Ciências Exatas e Tecnológicas".

1. Rede neural - Computação. 2. Rede neural artificial. 3. Monte Carlo - Método. 4. Proteína - Classificação. 5. Bioinformática. I. Título

CDU 004.855.5

Catalogação na Publicação:
Bibliotecária Vanessa Borges Nunes - CRB 10/1556

Dedicatória

Dedico aos meus pais,
que começaram a construir
o caminho que trilhei,
este trabalho e o meu sucesso nesta etapa:
ambos nada mais que a continuação de sua obra.

Agradecimentos

Agradeço à minha mãe, Medy, por acreditar mesmo quando eu já perdera a fé, por se intrometer na minha vida como um anjo da guarda rebelde e insistente, com o ar decidido de quem parece saber algo que ninguém mais sabe, por não dar ouvidos aos meus protestos e tentativas de errar, e por ansiar clara e constantemente por minha realização e felicidade.

Agradeço ao meu pai, Carlos Ary, pelo apoio incondicional, por me proporcionar sempre, ao alcance da mão, o exemplo de como as coisas podem ser, pelo espírito sagaz e pela insaciedade na busca de saber, por transmitir esta necessidade à nós, seus filhos, nos instigando a sempre buscar mais, e pela obra magnífica de construir a nossa família com o carinho e dedicação de seu grande coração.

Agradeço aos meus pais ainda e sobretudo por criarem a mim e aos meus irmãos em uma casa acolhedora, cheia de amor, carinho e incentivo, que hoje carrego comigo e na qual encontro conforto e força.

Agradeço à minha mulher, Márcia, por suportar e compreender os planos adiados, a indisponibilidade, a solidão que porventura causei, as indisposições de espírito, e continuar ao meu lado com seu amor, cuidando de mim, da casa, da minha vida e, muitas vezes sozinha, do nosso relacionamento.

Agradeço ao meu orientador Adelmo, pela excelência profissional e humana, por me abrir as portas do ambiente científico e acadêmico, por acreditar sempre, pelo incentivo transmitido com o seu fascínio pela pesquisa e a sua alegria a cada resultado promissor.

Agradeço à meu irmão José e minhas irmãs Teresinha e Inês, meus queridos sobrinhos, amigos, a todos que me querem bem, pela compreensão nas minhas ausências em festas, aniversários, confraternizações, almoços em família, pela falta de telefonemas, por não ter estado presente quando talvez precisaram de mim.

Agradeço por fim à minha querida irmã Lígia, por toda a alegria de viver que sempre transmitiu a mim e a todos que com ela privaram, nos poucos e preciosos anos em que nos deu a graça de sua companhia.

Resumo

Neste trabalho é proposto um novo método de otimização do método Monte Carlo (MC) aplicado ao dobramento de proteínas. Este método baseia-se em informações oriundas de Redes Neurais Artificiais (RNAs) treinadas para prever a estrutura secundária de proteínas. Inicialmente, são introduzidos conceitos básicos sobre proteínas e sua estrutura, sobre o método MC, sobre RNAs e sobre os métodos PHD e PROF de treinamento de RNAs para a predição de estruturas secundárias. A seguir, é apresentada uma revisão bibliográfica sobre métodos de previsão de estrutura tridimensional de proteínas e o ganho de informação em sistemas híbridos. Com base nos resultados obtidos em outras abordagens, um novo método é proposto utilizando as predições dos método PROF, disponíveis *on-line* e com índices de acerto para estrutura secundária acima de 76%, para a redução do espaço de busca do método MC aplicado ao dobramento de proteínas. O método MC é apresentado com a previsão da estrutura secundária baseada em RNAs (MC-RNA), e é aplicado a quatro proteínas retiradas da lista de proteínas alvo dos experimentos CASP, para as quais é demonstrado o ganho de acurácia do novo método em relação ao método MC na determinação da estrutura tridimensional das proteínas. Adicionalmente ao método MC e ao novo método MC-RNA, foi desenvolvido o método de controle MC-DSSP utilizando informação real e conhecida a priori sobre a estrutura secundária das proteínas. O método MC-DSSP também foi aplicado às quatro proteínas de teste para demonstrar como a qualidade das predições da estrutura secundária influencia a predição da estrutura terciária. Em todos os testes com os três métodos MC-DSSP, MC-RNA e MC, atingiu-se maior qualidade de predição de estrutura terciária com o método MC-RNA do que com o método MC, utilizando o mesmo esforço computacional. Da mesma forma o método MC-DSSP, que utiliza informação precisa sobre a estrutura secundária, obteve sempre melhores predições sobre a estrutura tridimensional do que os demais métodos, evidenciando a importância da qualidade da informação sobre a estrutura secundária na acurácia da predição da estrutura terciária de proteínas.

Palavras-chave: Bioinformática, Redes Neurais Artificiais, Monte Carlo, Dobramento de Proteínas.

TITLE: "USE OF ARTIFICIAL NEURAL NETWORKS WITH MONTE CARLO SIMULATION APPLIED TO THE PROTEIN FOLDING PROBLEM"

Abstract

This work proposes a new strategy to optimize the Monte Carlo method (MC) applied to the protein folding problem. This strategy is based on the information obtained from Artificial Neural Networks (ANNs), trained to predict the protein secondary structure. The work presents, initially, background knowledge about proteins and their structure. Follows an introduction to the MC method, Neural Networks and to the prediction of secondary structure using PHD/PROF programs. Then, a survey about tridimensional protein structure is presented. Other concepts, such as information gain in the context of hybrid systems, are also presented. Based on state-of-the art results, a new method is proposed using the predictions produced by the PROF program, available *on-line* and with a performance higher than 76% for secundary structure prediction, for the reduction of the MC search space. The MC method is presented with the secondary structure prediction based on ANNs (MC-RNA) and applied to four different proteins obtained from the list of target proteins in the CASP experiments. For these proteins, an improvement in performance is shown in relation to the conventional MC method. Additionaly to the MC method an to the new MC-RNA method, a validation method MC-DSSP was developed using real informations and a priori knowledge about the secondary structure. The method MC-DSSP was also applied to the four test proteins to demonstrate the influence of the quality in the secondary structure prediction on the tertiary structure prediction. In all tests with the three methods MC-DSSP, MC-RNA and MC, a higher score in terms of tertiary structure prediction was obtained with the MC-RNA method than with the MC method, for the same computer power. In the same way, the MC-DSSP method, which uses exact information about the secondary structure, reached better prediction for the tridimensinal prediction than the other methods, showing the importance of a good quality in the secondary structure for the prediction of the tertiary structure.

Keywords: Bioinformatics, Artificial Neural Networks, Monte Carlo, Protein Folding.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ANALYZE	Ferramenta para determinação de superfícies hidrofóbica e total
BLAST	*Basic Local Alignment Search Tool*
CASP	*Critical Assessment of techniques for protein Structure Prediction*
CNTP	Condições Normais de Temperatura e Pressão
CUBIC	*Columbia University Bioinformatics Center*
DP	Distribuição de Probabilidade
DSSP	*Database of Secondary Structure Assignments*
DSTK	*Diedral angles and Secondary strucure TollKit*
EVA	*Evaluation of automatic structure prediction*
MaxHom	Programa de múltiplo alinhamento dinâmico baseado em perfis
MC	Método Monte Carlo
PDB	*Protein Data Bank*
PHDsec	*Profile-based neural network prediction of protein secondary structure*
PROFsec	*Improved version of PHDsec: Profile-based neural network prediction of protein secondary structure*
PSI-BLAST	*Position-specific iterated BLAST*
RASMOL	*Software* de visualização molecular
RNA	Rede Neural Artificial
RMS	*Root Mean Square*
RSCB	*Research Collaboratory for Structural Bioinformatics*
SH	Superfície Hidrofóbica
ST	Superfície Total
SWISSPROT	Banco de dados de seqüências de proteínas
TINKER	Pacote de modelagem molecular para mecânica e dinâmica molecular

Sumário

Capítulo 1

Introdução

O método Monte Carlo (MC) é um método de simulação estocástico que pode ser utilizado para criar uma amostra estatisticamente representativa dos estados de um sistema físico. A simulação MC de um sistema físico consiste basicamente em transições aleatórias entre estados do sistema. Estas transições ocorrem de acordo com as probabilidades de uma cadeia de Markov e resultam, ao atingirem o equilíbrio, em uma amostra de estados visitados correspondente à distribuição de estados possíveis do sistema. Se esta amostra for grande o suficiente e analisarmos os estados em função de determinada variável, os estados que apresentarem valores mínimos para esta variável estarão próximos do mínimo global.

Dado um número de estados visitados, grande o suficiente e que a simulação MC tenha chegado ao equilíbrio, haverá entre eles um ou mais estados próximos ao mínimo global do sistema.

Redes Neurais Artificiais (RNAs) aproximam o comportamento de um sistema através de algoritmos de aquisição automática de conhecimento a partir dos dados do sistema. Além disto RNAs são capazes de aprender o comportamento global do sistema e são capazes de generalizar o comportamento do sistema para dados não vistos previamente. Tão importante quanto as características citadas acima é a capacidade que a RNA treinada tem de, uma vez alimentada com novos dados de entrada, gerar previsões em apenas um passo. Ou seja, enquanto modelos de sistemas dinâmicos dos quais não se conhece a solução analítica tem de ser resolvidos numericamente, com Δt pequeno e alto custo computacional, as RNAs podem aprender e armazenar a resolução analítica destes sistemas, e realizar em um passo o equivalente a N passos da solução analítica.

No entanto há limites teóricos para o que a RNA possa aprender sobre determinados sistemas. Há informações que simplesmente não se encontram codificadas apenas nos dados que descrevem o sistema mas dependem da sua dinâmica no tempo. Para estes casos o aprendizado armazenado na RNA a partir dos dados funciona como informação sobre o comportamento estatístico do sistema. Em outras palavras, baseada nas variáveis do sistema, a RNA pode prever com probabilidade P o estado final deste, e P é dependente da quantidade de informação existente nos dados de treinamento. Métodos Estocásticos por sua vez têm a capacidade de extrair amostras representativas dos estados possíveis de um sistema. É de se esperar portanto que o aumento de quantidade de informação oriunda de RNAs não só acelere a simulação MC como melhore a capacidade de aproximação do estado ótimo em relação aos dois métodos isoladamente.

O método MC-RNA proposto, baseado na premissa acima, é um modelo de simulação estocástico que utiliza o método MC para a minimização de uma função e RNAs para a redução do espaço de busca e conseqüente aceleração da solução do sistema. O modelo consiste em simulação estocástica pelo método MC, com a probabilidade de geração de transição de estado $g(\mu \rightarrow \nu)$ determinada pela RNA.

O problema de dobramento de proteínas foi escolhido como aplicação do método MC-RNA por ser um problema que envolve alto custo computacional tanto pelo uso de métodos de simulação contínua quanto pelo uso de métodos de simulação estocástica como o método MC. O alto grau de complexidade imposto pelas dimensões de uma proteína em termos de número de átomos, a falta de conhecimento sobre os mecanismos de interação internos à proteína e com o meio tais como a hidrofobicidade e o papel da entropia, e a importância da determinação da estrutura nativa de proteínas para a indústria química e farmacêutica, tornam a aceleração dos métodos de determinação do dobramento de proteínas um objetivo de suma importância. A estrutura tridimensional de proteínas é essencial para a determinação da sua função. A descoberta de novas drogas e terapias depende diretamente da nossa capacidade de prever qual a conformação final de uma proteína em seu meio.

O problema da determinação da estrutura tridimensional de proteínas esbarra na dimensionalidade do espaço de busca. A busca em todas as possibilidades de combinações de ângulos de proteínas com centenas de resíduos de aminoácidos (doravante referidos simplesmente como resíduos) resulta em explosão combinatória. Utilizando a mioglobina como exemplo, com ângulos diedrais entre os seus153 resíduos podendo variar livremente de um em um grau entre 0^o e 360^o, teríamos 360^{153} opções de conformações diferentes!

A literatura está repleta de métodos de predição de estrutura secundária de proteínas a partir da seqüência de aminoácidos [CF74, GRG91, Lev97, YL93, KS96, MHA95, RS93, Ros96, SS95, Jon99], e entre estes, o método de predição de estrutura secundária por RNAs têm obtido excelentes resultados. Uma vez treinadas, as RNAs podem realizar a predição diretamente a partir da seqüência de aminoácidos, sem necessidade de comparações com bancos de dados, e são capazes de generalizar o conhecimento para proteínas não vistas durante a fase de treinamento.

Tamanho da Janela	$Q_3(\%)$	C_α	C_β	C_{coil}
1	53,90	0,11	0,14	0,17
3	57,70	0,22	0,20	0,30
5	60,50	0,28	0,26	0,37
7	61,90	0,32	0,28	0,39
9	62,30	0,33	0,28	0,38
11	62,10	0,36	0,29	0,38
13	**62,70**	**0,35**	**0,29**	**0,38**
15	62,20	0,35	0,31	0,38
17	61,50	0,33	0,27	0,37
21	61,60	0,33	0,27	0,32

TABELA 1.1 – Dependência da precisão para os dados de teste (adaptado de [QS88]). Q_3 é á média de acerto na previsão das três estruturas α, β e *coil*. C é o coeficiente de correlação para cada tipo de previsão, como definido por [Mat75] apud [HMK95].

A abordagem do método misto MC-RNA destina-se a acelerar a simulação Monte Carlo com o uso de informação estatística proveniente de RNAs. A RNA é capaz de prever a estrutura secundária diretamente a partir da seqüência local de resíduos. Como a RNA é treinada com informações locais, não é capaz de modelar interações entre resíduos distantes. Qian & Sejnowski [QS88] e outros [HK89, BBB+90] mostraram que utilizar uma janela local de resíduos como entrada da RNA melhora os índices de acerto, como mostrado na Tabela 1.1. Esta Tabela mostra que com janelas pequenas o índice de acerto da RNA cai, evidenciando a importância da informação ao redor da janela para a predição da estrutura secundária. Porém, com janelas maiores do que 6 resíduos de cada lado do resíduo central também ocorre redução na precisão da previsão. Isto mostra que janelas maiores do que6 resíduos não contribuem com mais informação, mas pelo contrário deterioram a performance da predição adicionando ruído [HMK95]. Sabe-se contudo que a estrutura tridimensional da proteína envolve interações entre resíduos distantes, e para levar em conta a influência destes resíduos outros métodos devem ser utilizados que não dependam apenas de informações locais. O método MC leva em conta a iteração de todos os átomos da proteína, pois depende do cálculo da energia potencial das conformações que gera para calcular a probabilidade de transição entre estados.

O novo método MC-RNA utiliza portanto RNAs para acelerar as técnicas tradicionais de simulação estocástica. Esta aceleração acontece porque o método MC é simulado com distribuição de probabilidade baseada em RNAs, o que resulta em diminuição do número de estados passíveis de serem visitados. O método MC-RNA utiliza-se de RNAs treinadas através do método PROF[1], capaz de prever a estrutura secundária de proteínas a partir da seqüência de resíduos com precisão maior do que 76% [Ros01], e desenvolvido por Burkhard Rost a partir do método PHD, que por sua vez era capaz de predições com quase 72% de precisão e foi desenvolvido originalmente pelo próprio Rost e por Chris Sander [RS93, Ros96, Ros01].

O sistema físico responsável pelo dobramento da proteínas na natureza pode ser descrito por uma seqüência de aminoácidos e as diversas interações entre seus átomos. Partindo desta descrição é possível a construção de modelos matemáticos para a simulação da dinâmica das proteínas. O modelo que melhor representa estas interações é o quântico, constituído por equações de Schrödinger e intratável computacionalmente.

A simulação de dinâmica molecular é realizada através do modelo mecânico da interação entre átomos, e é uma das principais ferramentas para o estudo de comportamento de moléculas biológicas. A simulação de dinâmica molecular é utilizada tipicamente em estudos sobre a formação da estrutura tridimensional de proteínas, de sua dinâmica e termodinâmica.

Apesar da aceitação geral e ampla utilização, o Modelo Mecânico para simulação de Dinâmica Molecular ainda é um processo muito caro computacionalmente. Devido à falta de soluções analíticas para o modelo mecânico, a integração numérica de suas equações diferenciais se faz necessária. Inúmeros algoritmos de integração são utilizados pelos pacotes de dinâmica molecular, mas todos apresentam alto grau de complexidade do ponto de vista do custo computacional. De fato, proteínas são comumente compostas por seqüências de 100 a 500 aminoácidos, podendo ultrapassar este valor. Se considerarmos que cada aminoácido possui aproximadamente 10

[1] *B Rost: PROF: predicting one-dimensional protein structure by profile based neural networks. unpublished, 2000.*

átomos, a dinâmica molecular de uma proteína é um problema que envolve centenas ou milhares de graus de liberdade.

Na simulação de dinâmica molecular de uma mioglobina é necessário integrar as equações diferenciais de posição e velocidade para aproximadamente 1530 átomos em cada iteração. Para a simulação de $1\mu s$ de dinâmica molecular da mioglobina são necessários aproximadamente $1,53 \cdot 10^{12}$ integrações das equações diferenciais($\Delta t = 1\text{f}s$). Em um Xeon $2,40GHz$, $1\mu s$ de simulação da mioglobina em solvente explícito, equivale a 200 dias de processamento. De acordo com [LCN00] o tempo de dobramento de uma proteína em ambiente fisiológico é da ordem de ms a segundos, o que dá uma idéia da complexidade do problema da simulação de dinâmica molecular.

Na natureza a dinâmica molecular é responsável por duas etapas do dobramento de proteínas. Na fase inicial ocorre a formação da estrutura secundária, e posteriormente a formação da estrutura terciária. Ou seja, em um primeiro momento ocorre a formação de α-hélices e folhas-β. Uma vez terminado este processo, as estruturas secundárias começam a dobrar umas em direção às outras, formando a estrutura tridimensional final da proteína.

No método proposto, a RNA contribui na aceleração do método MC auxiliando no processo de formação da estrutura secundária. O método MC fica responsável principalmente pelo trabalho de dobrar as α-hélices e segmentos de folhas-β umas sobre as outras. Ou seja, a RNA é responsável pela previsão da estrutura secundária relativa à influência dos resíduos dentro de uma janela local, e o MC é responsável pelas iterações entre resíduos distantes.

Inicialmente são introduzidos no Capítulo 2 os conceitos básicos sobre aminoácidos e proteínas, a descrição de sua estrutura em diversos níveis, os campos de força para proteínas, o dobramento de proteínas e os métodos tradicionais para se atingir este dobramento. Também são introduzidos conceitos sobre o método MC, clusterização, sobre RNAs, sobre particularidades do treinamento de RNAs para predição de estrutura secundária de proteínas, e sobre os métodos PHDsec [Ros96, Ros96] e PROFsec[2] de treinamento de RNAs para a predição de estrutura secundária. O Capítulo 3 traz uma revisão bibliográfica sobre métodos de previsão de estrutura secundária de proteínas, o estado da arte dos métodos de Dinâmica Molecular, Métodos Estocásticos, RNAs, métodos de alinhamento (baseados em homologia) e o ganho de informação em sistemas híbridos. No Capítulo 4 é apresentado o novo método proposto: MC-RNA, que através do ganho de informação proporcionado pelas RNAs treinadas com o método PROF, otimiza o método MC aplicado ao dobramento de proteínas através de redução do espaço de busca. O novo método é dividido em duas fases: a primeira referindo-se à geração de conformações, e a segunda detalhando a clusterização dos dados gerados. O Capítulo 5 traz os resultados dos experimentos divididos em quatro seções, uma para cada proteína utilizada como teste. Finalmente no Capítulo 6 são apresentadas considerações finais e conclusões extraídas dos resultados dos experimentos.

No Capítulo 4, juntamente com o novo método MC-RNA proposto, são apresentados os dois métodos de controle utilizados neste trabalho a título de comparação: o MC e o MC-DSSP. O MC é o método MC aplicado ao problema de dobramento de proteínas, e o MC-DSSP é o mesmo método MC com informação

[2] *B Rost: PROF: predicting one-dimensional protein structure by profile based neural networks. unpublished, 2000.*

conhecida a priori da estrutura secundária das proteínas. O método MC serve como comparativo com o MC-RNA, permitindo mensurar o ganho de performance proporcionando pela informação extra do método MC-RNA. Já o método MC-DSSP utiliza informação conhecida, obtida experimentalmente, da estrutura secundária das proteínas testadas. Como o MC-DSSP utiliza informação já conhecida, ele não tem objetivo de ser uma ferramenta para uso prático em predições, mas foi proposto e aplicado para demonstrar como a maior precisão da informação sobre a estrutura secundária acarreta em aumento da acurácia dos resultados em relação ao MC-RNA.

Todos os três métodos MC, MC-RNA e MC-DSSP foram aplicados a um conjunto de quatro proteínas de domínio público, relativamente pequenas, escolhidas dentre o conjunto de proteínas utilizadas nos experimentos CASP[3] como alvo de simulações: $1j8b$, $1g7d$ domínio C-terminal, $1i74$ (domínio 2) e $1kkg$. A fim de gerar os dados do espaço de busca para os algoritmos, foi criado um banco de dados a partir da lista de proteínas não homólogas do grupo EVA[4], contendo informações sobre a estrutura secundária de 377540 resíduos pertencentes à 2327. A análise das conformações geradas a partir do método MC-RNA e dos métodos de controle é feita no Capítulo 5, analisando-se *clusters* obtidos com o método de clusterização *K-means* aplicado sobre os ângulos formados pelas ligações químicas entre os resíduos. Para otimizar o processo de clusterização, apenas uma fração dos resíduos das proteínas com maior liberdade de movimento é utilizada. Como os resíduos pertencentes à estruturas secundárias formam pontes de hidrogênio entre si, acabam por ter poucos graus de liberdade e podemos desprezá-los para fins de clusterização, adotando como representação da conformação tridimensional da proteína apenas os ângulos dos resíduos pertencentes às alças (*coil*) que interligam as estruturas secundárias, e permitem que elas se dobrem umas sobre as outras. Assim como os ângulos de ligação dos resíduos pertencentes à segmentos *coil* são determinantes para a conformação tridimensional da proteína, pequenas variações nestes ângulos causam grande impacto na estrutura obtida. Soma-se a isto a grande variedade de combinações de ângulos permitida (contra a existência de intervalos de ângulos preferenciais para folhas-β e α-hélices) para explicar o alto grau de complexidade inerente ao problema de dobramento de proteínas.

As proteínas são então representadas pelos ângulos dos resíduos *coil* que determinam como as estruturas se dobrarão umas sobre as outras, e para cada proteína há no Capítulo 5 gráficos e tabelas que comparam os clusters obtidos através destes ângulos. Os dados obtidos após o término dos experimentos e a clusterização mostram que, para todas as quatro proteínas testadas, o método MC-RNA obteve sempre maior acurácia e eficiência na determinação de conformações tridimensionais próximas à conformação nativa (conformação tridimensional da proteína na natureza) do que o método MC. Comparando-se ainda os métodos MC-RNA com o método MC-DSSP, o segundo foi sempre capaz de produzir estruturas mais próximas da estrutura nativa do que o primeiro. Considerando-se que as RNAs treinadas com o método PROF e utilizadas no método MC-RNA conseguem prever a estrutura secundária com acurácia maior que 76% [Ros01], os ótimos resultados do MC-DSSP permitem inferir que a melhora nos métodos de predição de estrutura secundária teriam grande efeito na otimização do MC aplicado à previsão da estrutura terciária, ou seja, ao dobramento de proteínas.

[3] http://predictioncenter.gc.ucdavis.edu/

[4] *EValuation of Automatic protein structure prediction*

Capítulo 2

Conceitos Básicos

Neste Capítulo, são apresentados conceitos necessários ao entendimento deste trabalho. Os assuntos abordados são a composição e estrutura de proteínas, técnica Monte Carlo, Clusterização e Redes Neurais Artificiais.

2.1 Aminoácidos e Proteínas

As proteínas são as mais abundantes macromoléculas biológicas, e estão presentes em todas as células e em todas as partes das células [LCN00]. As proteínas ainda apresentam enorme diversidade: no tamanho elas vão desde pequenos peptídeos a enormes polímeros com peso molecular na casa dos milhões de unidades de massa atômica, e na função onde podem servir para fins tão diversos quanto funções enzimáticas e hormonais, na composição da estrutura de tecidos como o tecido muscular, no transporte de moléculas e muitos outras.

Apesar de toda a diversidade, as proteínas são formadas por combinações de um grupo de apenas 20 aminoácidos padrão (Tabela 2.1), que se unem uns aos outros por ligações covalentes em uma seqüência de resíduos distinta para cada tipo de proteína. Todos os 20 aminoácidos padrão são α-aminoácidos. Eles são compostos de um carbono central denominado carbono-α e, ligados a ele, um grupo carboxila, um grupo amina, um átomo de hidrogênio e uma cadeia lateral. A cadeia lateral é denominada grupo R (de *radical*), e é o que diferencia um aminoácido do outro influenciando na carga elétrica, na estrutura, no tamanho, na polarização e na solubilidade em água do aminoácido (Tabela 2.1 e Figura 2.1).

Aminoácido	Abreviatura	Símbolo	Índice hidropático
Grupos R alifáticos, apolares			
Glicina	GLY	G	$-0,4$
Alanina	ALA	A	$1,8$
Valina	VAL	V	$4,2$
Leucina	LEU	L	$3,8$
Isoleucina	ILE	I	$4,5$
Metionina	MET	M	$1,9$
Grupos R aromáticos			
Fenilalanina	PHE	F	$2,8$
Tirosina	TYR	Y	$-1,3$
Triptofano	TRP	W	$-0,9$
Grupos R neutros, polares			
Serina	SER	S	$-0,8$
Prolina	PRO	P	$1,6$
Treonina	THR	T	$-0,7$
Cisteína	CYS	C	$2,5$
Asparagina	ASN	N	$-3,5$
Glutamina	GLN	Q	$-3,5$
Grupos R carregados positivamente			
Lisina	LYS	K	$-3,9$
Histidina	HIS	H	$-3,2$
Arginina	ARG	R	$-4,5$
Grupos R carregados negativamente			
Aspartato	ASP	D	$-3,5$
Glutamato	GLU	E	$-3,5$

TABELA 2.1 – Nomenclatura dos aminoácidos (adaptada a partir de [LCN00]). Os aminoácidos estão divididos por grupos R. Na última coluna o índice de hidropatia mede a tendência do aminoácido de procurar ambientes aquosos (valores $-$) ou ambientes hidrofóbicos (valores $+$).

Os 20 α-aminoácidos padrão podem portanto ser classificados pelo seus grupos R como [LCN00]: (1) alifáticos e apolares, (2) aromáticos, (3) neutros e polares, (4) carregados positivamente e (5) carregados negativamente.

Os grupos R da primeira classe são apolares e hidrofóbicos. As cadeias laterais da alanina, valina, leucina e da isoleucina tendem a se agrupar no interior das proteínas, estabilizando-as com interações hidrofóbicas.

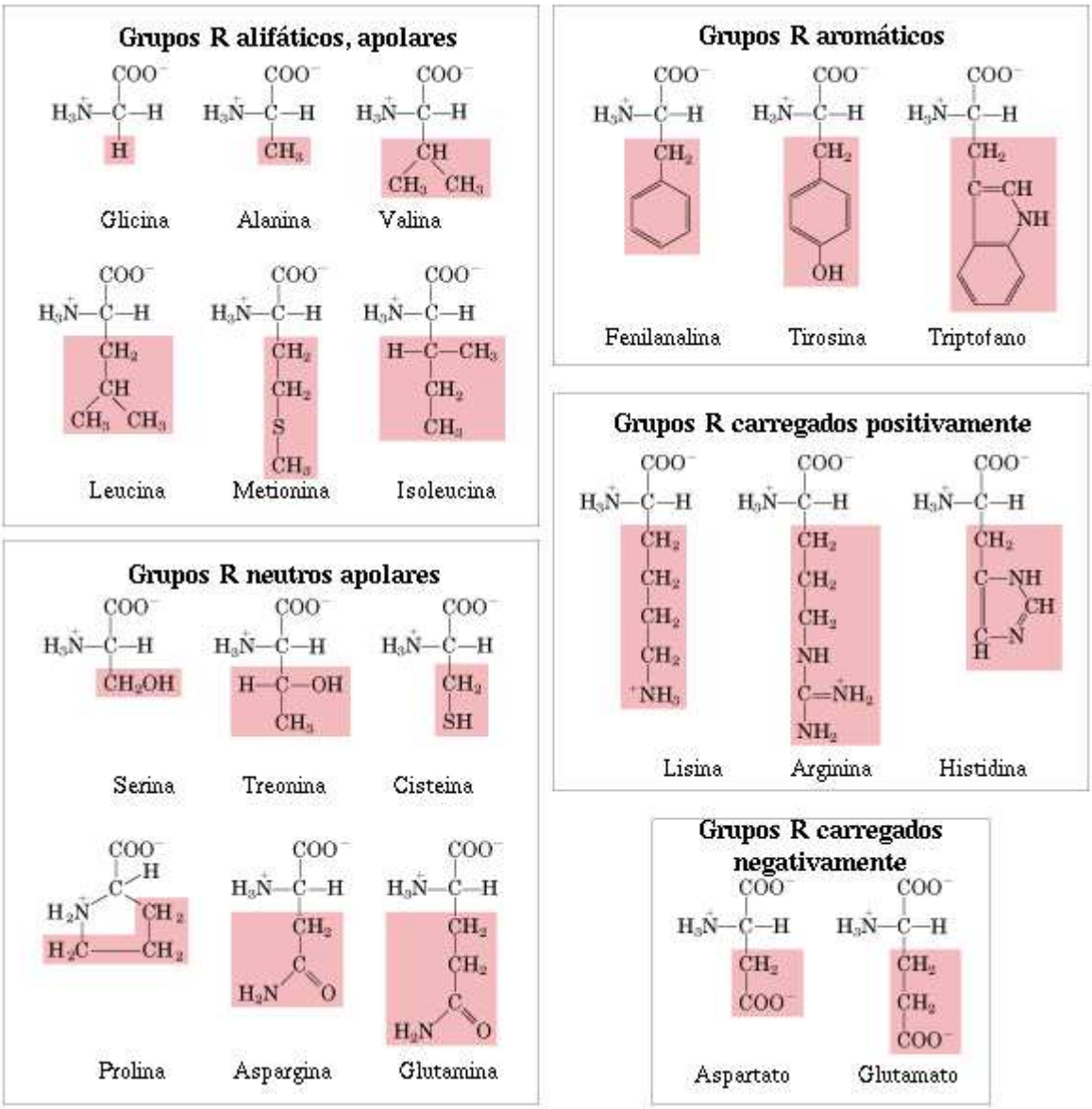


FIGURA 2.1 – Os 20 aminoácidos padrão das proteínas([LCN00]) classificados pelo grupo R. As fórmulas estruturais mostram o estado de ionização predominante em pH fisiológico (7, 0). As partes não sombreadas são comuns à todos os aminoácidos, e as partes sombreadas são os grupos R.

A segunda classe corresponde aos grupos R aromáticos. Estes são relativamente apolares (hidrofóbicos). Os aminoácidos desta classe têm por característica a absorção de luz ultravioleta, aspecto aproveitado por cientistas para caracterizar proteínas.

A terceira classe é composta por grupos R neutros e polares, mais solúveis em água do que os grupos R apolares porque contém grupos funcionais que formam pontes de hidrogênio com a água. A cisteína se oxida na presença de outra cisteína e forma um aminoácido dimérico chamado de cistina. A cistina é formada por duas cisteínas ligadas através de ligação covalente dissulfídica, e é altamente hi-

drofóbica (apolar). As ligações dissulfídicas têm influência importante na formação da estrutura tridimensional de proteínas, pois formam ligações covalentes entre dois segmentos da proteína ou entre dois polipeptídeos.

Os grupos R das classes (4) e (5) apresentam carga elétrica positiva e negativa respectivamente. Estes grupos são os mais hidrofílicos. A histidina, por ter uma cadeia lateral ionizável em ambiente próximos ao pH neutro, tem a função de facilitar inúmeras reações catalizadas por enzimas servindo como doador e receptor de prótons.

A proteína é portanto um polímero de resíduos de aminoácidos, assim denominados devido à perda de água pelo aminoácido ao se ligar através de um ligação covalente à outro aminoácido. As diferentes combinações deste grupo de 20 aminoácidos formam as seqüências de resíduos, ou polipeptídeos, que formam todas as proteínas. A análise da porcentagem de cada tipo de aminoácido presente em uma proteína pode ser obtida por hidrólise, por meio da qual a seqüência de resíduos é desnaturada e resulta em uma mistura de aminoácidos livres. É interessante notar que as porcentagens e mesmo a presença de tipos de aminoácidos varia de proteína para proteína, sendo difícil encontrar duas proteínas diferentes com a mesma proporção de tipos de aminoácidos.

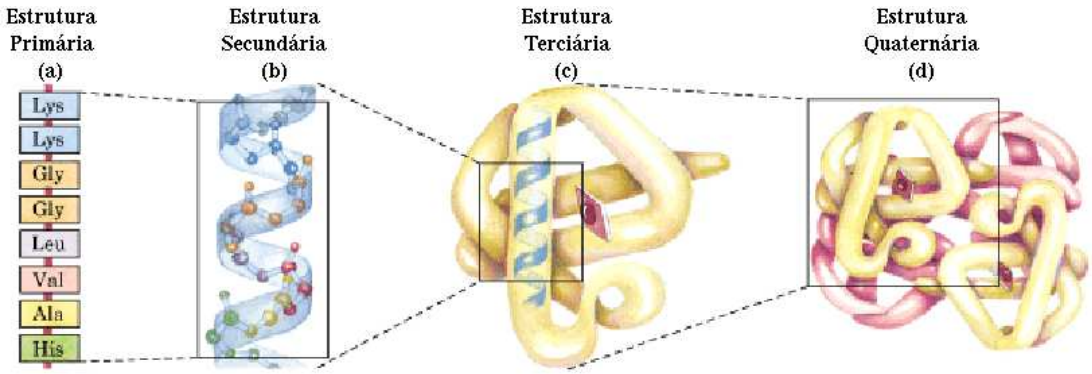


FIGURA 2.2 – Representação hierárquica dos níveis de estrutura em proteínas [LCN00].

O conhecimento sobre a estrutura tridimensional de uma proteína em ambiente natural (estrutura nativa) é essencial para o entendimento de seu funcionamento, pois a função da proteína no organismo é determinada por sua conformação nativa. A conformação de uma proteína é o arranjo espacial de todos os seus resíduos. Mesmo para uma proteína pequena, composta por poucas dezenas de resíduos, as possíveis conformações teóricas seriam tantas quantas as combinações de variações de ângulos possíveis para cada uma das ligações covalentes entre os resíduos. No entanto, apenas algumas conformações tendem a predominar sobre as outras sob condições biológicas (temperatura, íons, nível de pH, temperatura, etc, similares ao ambiente fisiológico). A Hipótese Termodinâmica [Anf93] estabelece que a estrutura tridimensional de uma proteína em seu ambiente fisiológico natural é tal que a energia livre de Gibbs [LCN00] (G) de todo o sistema é mínima. Em [BK00] os autores sugerem que o dobramento da proteína em direção à conformação nativa percorre um túnel de energia decrescente sem barreiras de energia importantes, e que "a taxa de dobramento da proteína é limitada por uma região do túnel onde o ganho

de energia não compensa a perda de entropia conformacional". Os autores sugerem que a cinética da proteína é determinada majoritariamente por barreiras entrópicas. Como a conformação nativa de proteínas (e portanto de energia livre mínima) é determinada por interações fracas [LCN00], simulações de dinâmica molecular que levem em consideração apenas a entalpia do sistema têm dificuldade em encontrar os mínimos globais.

A descrição da estrutura de uma proteína é dividida em três etapas principais (Figura 2.2): (1) a estrutura primária descreve as ligações covalentes entre os resíduos, e tem como elemento mais importante a seqüência de resíduos de aminoácidos, (2) a estrutura secundária se refere à arranjos locais estáveis de resíduos na forma de estruturas recorrentes, e (3) a estrutura terciária é a estrutura tridimensional global da proteína. Para proteínas grandes como a hemoglobina, composta por mais de uma cadeia polipeptídica, o arranjo espacial destas cadeias é denominado estrutura quaternária.

2.1.1 Estrutura Primária

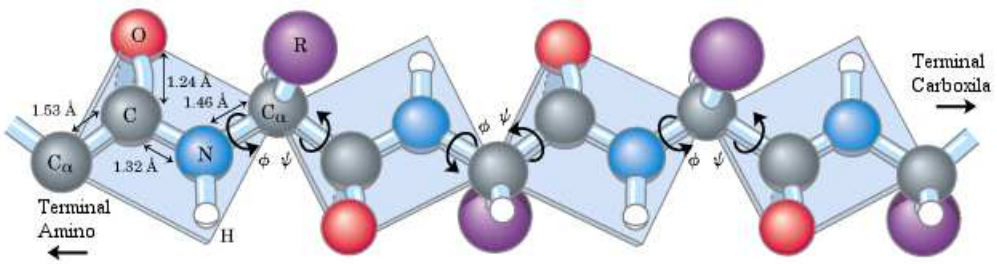


FIGURA 2.3 – Cadeia polipeptídica. Por convenção os ângulos de rotação das ligações covalentes no carbono alfa (C_α) são denominados Φ para a ligação $N - C_\alpha$ e Ψ para a ligação $C_\alpha - C$. Os planos indicam que os átomos das ligações covalentes $C_\alpha - C - N - C_\alpha$ são coplanares e portanto as únicas ligações covalentes com liberdade para rotacionar são as do C_α. ([LCN00]).

A estrutura primária de uma proteína é a descrição de todas as ligações covalentes entre a sua seqüência de resíduos. Linus Pauling e Robert Corey, em meados de 1930 determinaram que entre cada carbono alfa (C_α) há três ligações covalentes do tipo $C_\alpha - C - N - C_\alpha$, e que os 4 átomos participantes são coplanares [LCN00]. As únicas ligações covalentes com liberdade para rotacionar são as do C_α. Por convenção os ângulos de rotação das ligações covalentes no C_α são denominados Φ para a ligação $N - C_\alpha$ e Ψ para a ligação $C_\alpha - C$. Estes ângulos são chamados ângulos diedrais e, devido à rigidez imposta pela coplanaridade dos grupos peptídicos, tem a sua liberdade de rotacionar limitada pela colisão dos outros átomos do grupo (Figura 2.3).

O mapa de Ramachandran [GRB96] é um mapa dos ângulos diedrais permitidos para resíduos pertences à uma proteína, e foi proposto por G. N. Ramachandran em 1963. No mapa de Ramachandran os valores permitidos para os ângulos diedrais são os que se encontram em regiões de pares de ângulos permitidos (Figura 2.4). Todos os demais que se encontram fora destas regiões são considerados ângulos não

permitidos ou proibidos. As regiões permitidas do mapa são construídas através da determinação dos pares de ângulos Φ e Ψ que que respeitam as distâncias mínimas permitidas entre átomos em uma cadeia polipeptídica. Para calcular tais distâncias os átomos são tratados como esferas sólidas de raio igual aos seus raios de Van der Waals. As regiões parcialmente permitidas são obtidas com raio de Van der Waals ligeiramente menor e correspondem às conformações de menor estabilidade [LCN00].

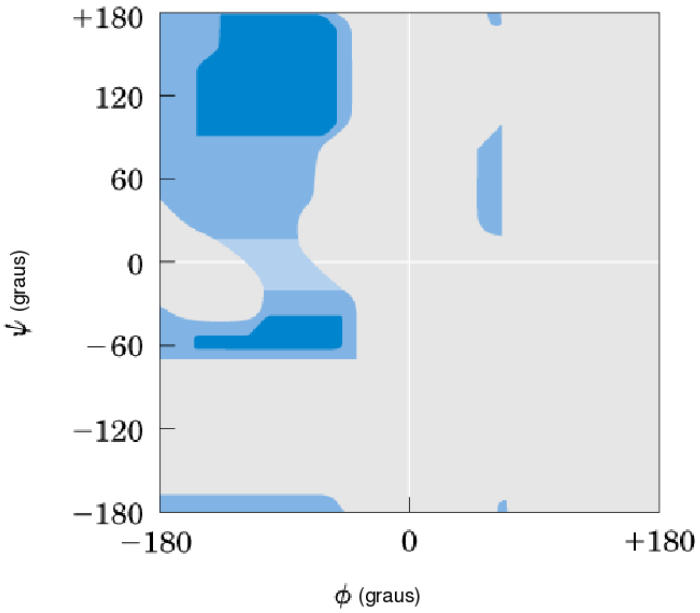


FIGURA 2.4 – Mapa de Ramachandran: os valores permitidos para os ângulos diedrais Φ e Ψ são limitados pela proximidade dos átomos dados os seu raios de Van der Walls [LCN00]. Na área cinza do mapa encontram-se as combinações *proibidas* de ângulos diedrais. Nas regiões azuis encontram-se as regiões permitidas.

Se a estrutura primária da proteína determina a formação da estrutura secundária, a interação entre os segmentos da estrutura secundária determinam a estrutura tridimensional, e a função da proteína depende de sua forma tridimensional nativa, então podemos dizer que a seqüência de aminoácidos define a função da proteína. Porém, de 20 a 30% das proteínas em humanos são polimórficas [LCN00], ou seja, apresentam variações de resíduos na seqüência, porém com pouca ou mesmo nenhuma alteração na função final da proteína. Aparentemente apenas segmentos críticos da estrutura primária tem de se manter inalterados entre proteínas polimórficas para que mantenham a sua função.

Não se sabe exatamente como a seqüência de aminoácidos determina a estrutura terciária de uma proteína, e nem sempre é possível prever a estrutura terciária a partir da primária. Porém, através de comparação com proteínas homólogas com conformações nativas conhecidas, se pode inferir a conformação espacial preferencial da proteína [LCN00]. Ou seja, métodos de alinhamento de estruturas homólogas assumem que proteínas com estruturas primárias e portanto seqüências de resíduos semelhantes tendem a assumir conformações nativas preferenciais semelhantes.

2.1.2 Estrutura Secundária

A estrutura secundária refere-se a conformações recorrentes locais em segmentos da seqüência de aminoácidos. Estas conformações são divididas em dois grupos: α-hélices e folhas-β.

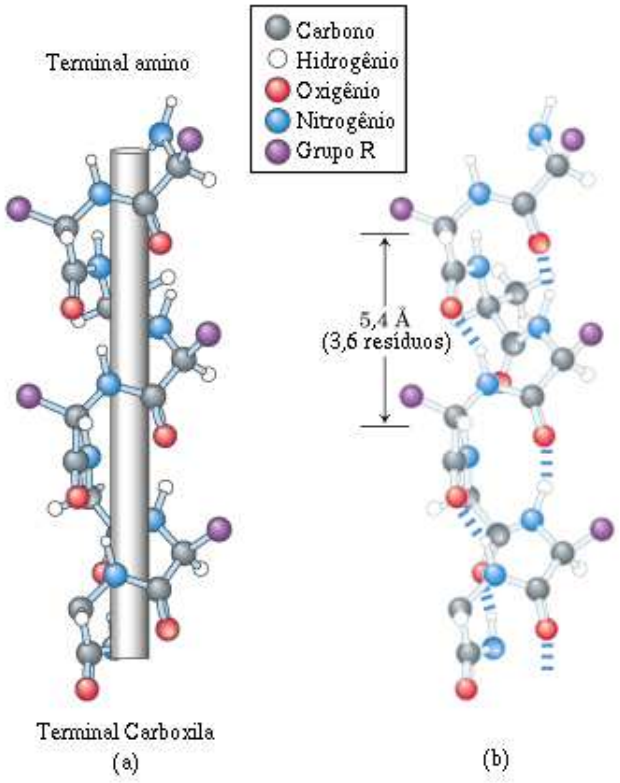


FIGURA 2.5 – Dois modelos da α-hélice de orientação anti-horária (mão direita) [LCN00]. (a) Os planos das ligações peptídicas são paralelos ao eixo da α-hélice representado pelo bastão. (b) Modelo bola e bastão da α-hélice mostrando as pontes de hidrogênio.

A estrutura α-hélice é formada por um segmento de proteína onde os resíduos formam um espiral estreita ao redor de um eixo imaginário, atraindo-se mutuamente por meio de pontes de hidrogênio (Figura 2.5). Para cada volta completa da α-hélice são necessários aproximadamente $3,6$ resíduos, e os grupos R dos resíduos situam-se no lado externo da hélice.

As interações mútuas através de pontes de hidrogênio determinam duas características importantes da α-hélice: ela é a estrutura que se forma mais rapidamente e é a mais estável. Porém a estabilidade desta estrutura depende da identidade dos resíduos que a compõem. Para citar apenas 2 exemplos desta influência (Lehninger cita pelo menos 5), a formação e estabilidade de uma α-hélice é afetada por (1) a atração ou repulsão entre grupos R sucessivos carregados eletricamente, e (2) as dimensões de grupos R adjacentes.

A conformação β, assim como a α-hélice, foi predita por Pauling e Corey. Este tipo de conformação repetitiva resulta em estruturas com formato de zigue-zague. As conformações β com freqüência se alinham através de interações por ponte de

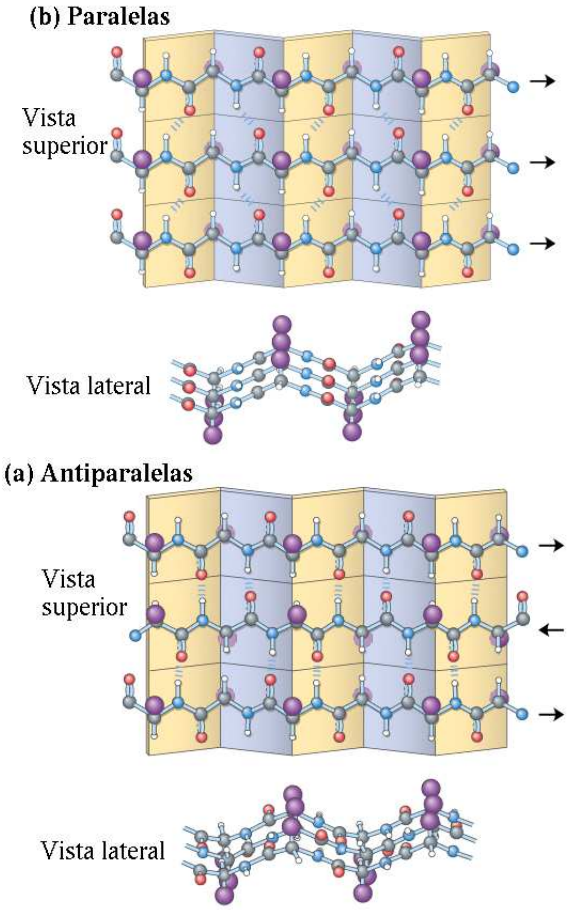


FIGURA 2.6 – Conformação β de cadeias polipeptídicas [LCN00]. As vistas superior e frontal evidenciam os grupos R sobressaindo da forma sanfonada criada pelas ligações peptídicas. As pontes de hidrogênio também são mostradas. Na folha-β antiparalela (a) a orientação terminal-amino para terminal-carboxila é invertida para cada segmento. Na folha-β paralela os segmentos têm a mesma orientação.

hidrogênio, formando estruturas com superfícies em forma de gaita denominadas folhas-β. Neste tipo de conformação, os grupos R são dispostos alternadamente em direções opostas (Figura 2.6), e preferencialmente resíduos com grupos R pequenos são encontrados na seqüência de folhas-β. As folhas-β são formadas, em geral, por segmentos próximos na seqüência de resíduos [LCN00], mas podem ser formadas por segmentos distantes e até por polipeptídeos distintos. Ainda conforme a orientação, as folhas-β podem ser paralelas ou antiparalelas (Figura 2.6 (b) e (a)) conforme a sua orientação.

Os resíduos que não participam das seqüências de estruturas secundárias estão nos segmentos que ligam estas seqüências. Estes segmentos se denominam*coil*, e não possuem estrutura definida, sendo portanto de difícil determinação. Em proteínas globulares de estrutura altamente compacta, mais de 30% dos aminoácidos estão em *coils* em forma de *loops* entre uma e outra estrutura secundária. Alguns tipos de dobramentos são comuns em proteínas como o β-*turn*, composto por 4 resíduos e uma ponte de hidrogênio conectando dois segmentos de folhas-β antiparalelas, em

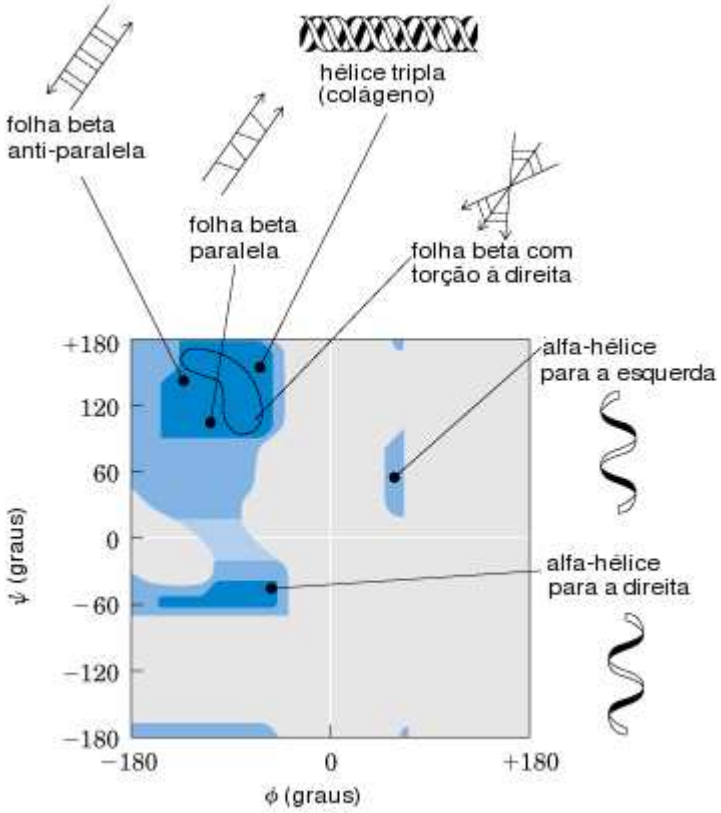


FIGURA 2.7 – Os ângulos diedrais dos resíduos participantes de diferentes estruturas secundárias encontram-se em regiões específicas do Mapa de Ramachandran.

geral localizado na superfície das proteínas [LCN00] globulares.

Os resíduos quando pertencentes a um determinado tipo de estrutura secundária tendem a ter ângulos diedrais característicos, inerentes às restrições impostas pelo tipo de estrutura espacial. Estes conjuntos de ângulos característicos formam regiões específicas no mapa de Ramachandran para cada tipo de estrutura secundária. Apesar de estas regiões variarem conforme o tipo de resíduo, de uma maneira geral as regiões típicas para cada tipo de estrutura secundária são as mostradas no esquema da Figura 2.7. Este mapeamento nos permitirá mais adiante restringir o espaço de busca de ângulos diedrais para estas pequenas regiões sempre que dispusermos de informação sobre a estrutura secundária a que pertence determinado resíduo.

2.1.3 Estrutura Terciária

A estrutura terciária é a estrutura tridimensional dada pela posição espacial de todos os átomos de uma proteína. A estrutura terciária engloba então a descrição da posição relativa de todos os segmentos de estrutura secundária e de todos os resíduos dos segmentos intermediários (coil) (Figura 2.8). A determinação da conformação nativa de uma proteína a partir da seqüência de resíduos é nada mais do que determinar a sua estrutura terciária nativa a partir da estrutura primária.

FIGURA 2.8 – Estrutura terciária da proteína *glutathione peroxidase* do boi. Assinalados em amarelo as folhas-β, em vermelho as hélices-α e em azul os segmentos *coil*.

Como vimos anteriormente, as estruturas secundárias são formadas localmente em segmentos da cadeia polipeptídica em função dos tipos de resíduos presentes na seqüência local. As estruturas secundarias, uma vez formadas no ambiente fisiológico, interagem entre si e com os segmentos de coil, e o resultado da interação das cargas elétricas, das componentes hidrofóbicas e hidrofílicas, das pontes de hidrogênio, das ligações dissulfídicas, e até das restrições de movimento devido ao choque de estruturas é a conformação nativa final da proteína.

Após o processo de formação das estruturas secundárias, estas começam a se dobrar como efeito das interações de suas cargas elétricas, polarização e efeitos hidrofóbicos, em direção à conformação nativa. Portanto, uma vez que tenhamos conhecimento sobre a estrutura secundária de uma proteína, a descrição da estrutura terciária depende dos ângulos diedrais dos resíduos dos segmentos que conectam as estruturas secundárias.

2.1.4 Estruturas Primárias Redundantes

Se o polimorfismo representa um obstáculo no mapeamento da seqüência de resíduos para a conformação nativa e conseqüente função de uma proteína, a determinação da conformação nativa sem o polimorfismo também apresenta limitações. Isto se deve à necessidade de locomoção de algumas proteínas (principalmente enzimas) que acabam por modificar a sua conformação nativa. Disto resultam arquivos PDB[1] ditos redundantes: proteínas 100% homólogas com conformações nativas diferentes. De acordo com [HW02] o limite teórico para a taxa média de acerto na previsão da conformação de proteínas a partir da estrutura primária (utilizando portanto um método com 100% de taxa de acerto) seria de $73,5\%$. Ou seja, o limite teórico de acurácia para um método de previsão da conformação nativa de proteínas baseado apenas na seqüência de resíduos é menor do que 100% porque a estrutura nativa nem sempre é determinada apenas pela seqüência de resíduos.

[1] Arquivos com dados sobre a estrutura tridimensional de proteínas do repositório *on-line* RSCB *Protein Data Bank*, em http://www.rcsb.org/pdb/

A quantidade de conformações nativas de proteínas determinadas por métodos empíricos atualmente é da ordem de milhares. Os dois métodos utilizados são a cristalografia por raio X e ressonância magnética, e ambos se destinam a determinar a posição tridimensional de cada átomo de uma proteína. O alto custo em tempo e recursos para a determinação da conformação nativa por estes métodos e a conveniência de conhecimento a priori sobre as possíveis conformações são os motores da pesquisa em métodos de biologia computacional para o dobramento de proteínas.

2.2 Campos de força em Proteínas

Campo de força é o nome que se dá a um conjunto de informações que permite calcular a energia de um sistema de uma ou mais moléculas em função da distância entre átomos de ligações covalentes, entre 2 ligações covalentes (C-C-C) em função da variação angular, variações de energia de torção, interação de van der Waal entre 2 átomos, forças eletrostáticas, barreiras rotacionais (limites de rotação para ligações entre 4 átomos, responsáveis pelas regiões proibidas do mapa de Ramachandran), e parâmetros como energia de formação de moléculas, constantes ambientais (como a constante dielétrica por exemplo), comprimentos de ligações atômicas e raios de van der Waal, entre outros.

As interações entre átomos são governadas por interações eletromagnéticas, sendo que as interações gravitacionais e nucleares são completamente irrelevantes na escala atômica c. O comportamento de uma molécula pode ser descrito pela equação de Schrödinger (aqui na sua forma independente do tempo)

$$H\phi(r_1, r_2, \ldots) = E\phi(r_1, r_2, \ldots), \tag{2.1}$$

onde r_i são as posições do núcleo e elétrons da molécula, ϕ é a função de onda que contém toda a informação sobre as propriedades dinâmicas do sistema, eE é a energia. O operador hamiltoniano é dado por

$$H = \sum_{i=1}^{n} \frac{-h^2}{2m_i} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{z_i z_j e^2}{4\pi\varepsilon_0 |\vec{r}_i - \vec{r}_j|}, \tag{2.2}$$

onde m_i é a massa da partícula i, z_i é a carga e os outros símbolos têm o significado usual. O primeiro termo é referente à energia cinética clássica e o segundo à energia eletrostática. As propriedades de todas as moléculas, incluindo as proteínas, são governadas por esta equação (excluindo *pequenos* efeitos relativísticos. A solução exata desta equação porém não é possível, nem mesmo se considerarmos moléculas muito simples como H_2.

Born e Oppenheimer em 1927 [BO27] desenvolveram uma boa aproximação para as distribuições eletrônicas e nucleares. Para esta aproximação a energia para uma molécula de N núcleos e n elétrons é dada por

$$E_{total} = E_{eletrons} + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{z_i z_j e^2}{4\pi\varepsilon_0 |\vec{R}_i - \vec{R}_j|}, \tag{2.3}$$

onde R_i são os vetores de posição dos núcleos e z_i as cargas. O primeiro termo representa a contribuição feita pela energia potencial das interações envolvendo os

elétrons, o segundo é o termo coulombiano de repulsão entre os núcleos carregados de uma molécula. A função de onda eletrônica e a energia potencial são dados por

$$H_{eletrons}\phi(r_1, r_2, \ldots) = E_{eletrons}\phi(r_1, r_2, \ldots). \tag{2.4}$$

O operador hamiltoniano para a contribuição dos elétrons para a energia é dado por:

$$H_{eletrons} = \sum_{i=1}^{n} \left\{ \frac{-h^2}{2m_i} \bigtriangledown_i^2 - \sum_{j=1}^{n} \frac{z_j e^2}{4\pi\varepsilon_0 |\vec{r}_i - \vec{R}_j|} \right\} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{e^2}{4\pi\varepsilon_0 |\vec{r}_i - \vec{r}_j|}, \tag{2.5}$$

onde m é a massa do elétron e $\vec{r}_i$ são os vetores posição para os elétrons. O primeiro termo da soma corresponde à energia cinética dos elétrons, o segundo as interações entre os elétrons e os núcleos, e o terceiro as interações entre elétrons.

A solução para as equações acima leva a um grande campo de estudos: a química quântica. A idéia básica dos métodos da química quântica é encontrar a distribuição dos elétrons para um conjunto fixo de núcleos descrevendo a molécula e, com a aplicação de um método de minimização de energia a geometria da molécula pode ser determinada. As equações de Born e Oppenheimer porém, apesar de serem aproximações, ainda tem grau de complexidade computacional apreciável, sendo inviáveis computacionalmente para moléculas compostas por mais do que algumas dezenas de átomos.

Os métodos da química quântica para representar moléculas são impraticáveis quando aplicados a biomoléculas, assim devemos então considerar um modelo mais simples de representação. Os efeitos quânticos são aproximados pela mecânica clássica para facilitar a representação. Muitos campos de força foram desenvolvidos com o uso de dados experimentais para parametrizar um conjunto de funções. Embora os campos de força tenham algumas diferenças, acabam por usar praticamente o mesmo conjunto de funções de energia. A seguir, descrevemos os termos de energia que são utilizadas para quantificar a energia de proteínas.

2.2.1 Interação entre Átomos Ligados

Estas interações aplicam-se aos átomos que estão próximos uns dos outros, ou seja, a não mais do que 2 ligações de distância, e por isso são chamadas também e interações $1-3$.

Ligações Covalentes

Ligações covalentes existem quando dois átomos compartilham elétrons. Caso compartilhem um elétron apenas, temos uma ligação simples, e se compartilham um par de elétrons temos uma ligação dupla.

A lei de Hooke aproxima a energia potencial de uma ligação covalente, é dada por:

$$E_{lig} = k_r(r - r_{eq})^2, \tag{2.6}$$

onde r é a distância entre os núcleos dos átomos em uma ligação covalente, r_{eq} é o comprimento de equilíbrio da ligação, e k_r é a constante de mola.

Os valores de r_{eq} são usualmente obtidos com raios-X de pequenos cristais de moléculas. Já os valores de constante k_r são encontrados por meio de cálculos e

através da comparação dos resultados com dados experimentais das freqüências de microondas. Estes parâmetros também podem ser obtidos por cálculos de química quântica.

Ângulos entre Ligação

Um ângulo θ entre os átomos ABC é definido como o ângulo entre as ligações AB e BC, como pode ser visto na Figura 2.9.

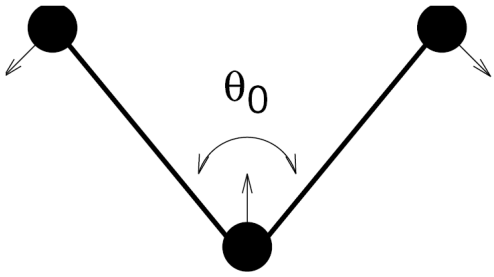


FIGURA 2.9 – Ângulos de ligação.

Assim como a energia associada com a deformação das ligações covalentes, um termo de energia também é associado à deformação dos ângulos de ligação e é dado por:

$$E_\theta = k_\theta(\theta - \theta_{eq})^2, \tag{2.7}$$

onde k_θ é uma constante positiva que depende do tipo de ângulo, θ é o valor do ângulo e θ_0 é o ângulo de equilíbrio.

Os valores de ângulos são encontrados experimentalmente. Um ângulos de ligação em torno de 109^o significa que o átomo central é tetraédrico, possuindo quatro átomos ligados a ele. Já um ângulo em torno de 120^o indica um átomo central com três átomos ligados. Os valores de ângulos de ligação também são obtidos empiricamente através de raios-X de alta resolução de pequenas moléculas. Também podem ser obtidos através de dados de espectrografia ou de cálculos.

Ângulos Diedrais

Estes são os ângulos de torção Φ e Ψ já vistos anteriormente. O ângulo diedral Ψ (ângulo de torção) entre quatro átomos $ABCD$ é definido como o ângulo entre os planos ABC e BCD, como podemos ver na Figura 2.10.

A forma funcional padrão para representar a energia potencial de uma rotação funcional foi introduzida por Pitzer. Esta é uma função periódica representando a interação entre os quatro átomos que formam o ângulo diedral:

$$E_{died} = \sum_{n=1}^{3} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)], \tag{2.8}$$

onde V_n é a barreira de energia para a rotação, n a quantidade de máximos ou mínimos em uma volta e γ determina o *off-set* angular. Nos anos 60, quando funções de

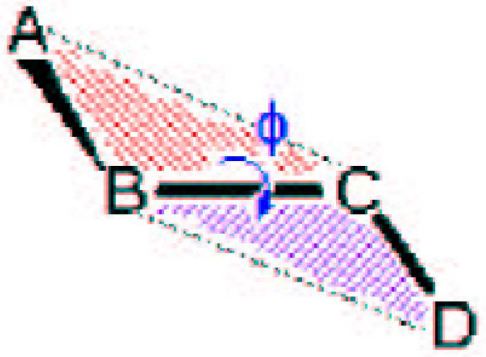


FIGURA 2.10 – Ângulos diedrais.

energia potencial foram desenvolvidas para proteínas, foi verificado que o potencial de Pitzer era insuficiente para fornecer uma representação completa das barreiras de energia nas mudanças de ângulos diedrais. Funções de energia potencial modernas normalmente modelam a dependência da energia em função do ângulo diedral por uma combinação do potencial de Pitzer e interações entre átomos não ligados,

2.2.2 Interações Entre Átomos Não Ligados

Estas interações aplicam-se a átomos que não estão ligados através de ligações covalentes. São chamadas de interações $1-4$ por atuarem em átomos que encontram-se distantes de três ou mais ligações. As interações eletromagnéticas dominam na escala molecular e fornecem a base fundamental para todas as interações, ligadas ou não.

Interações Eletrostáticas

No caso das interações eletrostáticas as cargas do núcleo e elétrons interagem de acordo com a lei de Coulomb:

$$V = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}, \tag{2.9}$$

onde q_i e q_j são as cargas e r_{ij} sua distância, ε_0 a permissividade no vácuo e ε_r a constante dielétrica do meio onde as cargas estão colocadas. A maneira estritamente correta de usar esta lei seria considerar todos os núcleos e elétrons separadamente, colocá-los na equação de Schrödinger e aplicar métodos de química quântica para resolver a equação para a configuração espacial dos núcleos de interesse. Esta solução no entanto é completamente impraticável para sistemas biomoleculares. Então devemos desenvolver um modelo que seja útil para lidar com os núcleos dos átomos sem que seja necessário tratar os elétrons explicitamente.

Pontes de Sal

Como poderia se esperar os resíduos lisina e arginina, que são positivamente carregados, podem formar uma interação forte com os resíduos ASP ou GLU que são

carregados negativamente. Nas proteínas esta interação é conhecida como ponte de sal. Elas são relativamente raras.

Pontes de Hidrogênio

As interações eletrostáticas entre grupos que não possuem carga elétrica são de fundamental importância para a estrutura biomolecular.

FIGURA 2.11 – Molécula de água.

O que ocorre é que grupos sem carga elétrica podem ter uma grande polarização. As órbitas em torno da molécula são distribuídas de uma maneira que partes da molécula tenham menos elétrons, e portanto uma carga positiva, e a outra parte tenha em excesso e portanto carga negativa. Alguns átomos possuem a tendência de atrair elétrons sendo chamados de eletronegativos. Outros átomos têm a tendência de perder elétrons e são chamados de eletropositivos. Em casos extremos esta tendência faz com que um átomo perca totalmente um elétron para outro, levando à formação de compostos carregados conhecidos como íons. Em um caso menos extremo elétrons são compartilhados por dois átomos em uma ligação covalente, mas são puxados para o lado de um dos átomos. Exemplo clássico é a molécula de água esquematizada na Figura 2.11.

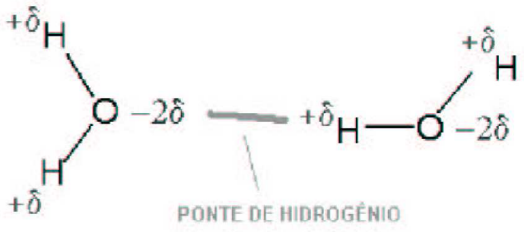


FIGURA 2.12 – Ponte de hidrogênio entre molécula de água.

Como o O_2 é eletronegativo ele atrai os elétrons que está compartilhando na ligação com o H_2, o que distribui as cargas de maneira que os átomos de H fiquem com caráter positivo e o O_2 com caráter negativo. No caso da água o valor da carga efetiva em cada átomo de H é alto (cerca de 1/3 do valor do elétron) e, combinado

com a curta distância entre átomos de hidrogênio e oxigênio, leva a molécula de água a ter um grande momento dipolo. Duas moléculas de água podem formar então uma interação eletrostática muito forte, como esquematizado na Figura 2.12.

Esta interação é conhecida como ponte de hidrogênio. Moléculas de água podem formar uma rede e são muito importantes para a estrutura das proteínas. A capacidade das ligações do carbóxi-oxigênio da cadeia principal em formar pontes de hidrogênio com os aminoácidos da cadeia principal leva à possibilidade de formação das diferentes estruturas secundárias, como as α-hélices e folhas-β.

Dispersão

Um átomo pode ser visualizado como tendo um núcleo com carga positiva envolto em uma nuvem de elétrons com carga negativa. Em um ponto externo do átomo a carga será negativa, uma vez que a carga positiva do núcleo está exatamente no centro balanceado pelas cargas negativas dos elétrons. Entretanto, como os átomos vibram, por alguns instantes a carga positiva não estará exatamente no centro, criando um dipolo instantâneo. Como existem outros átomos na proximidade do primeiro, este será afetado pelo primeiro e terá induzido um dipolo. Os dois dipolos atraem-se mutuamente produzindo uma interação atrativa. Pode-se demonstrar que a interação de dispersão varia de acordo com a potência sexta da distância entre dois átomos: $-B_{ij}/r_{ij}^6$. O fator B_{ij} depende basicamente da natureza do par de átomos interagindo. É normal parametrizar a dispersão empiricamente usando dados energéticos e estruturais de cristais de pequenas moléculas.

Termos de Repulsão

Quando dois átomos são trazidos muito próximos um do outro a ponto de seus orbitais começarem a se sobrepor, existe um custo muito alto de energia. No limite em que os núcleos atômicos são coincidentes os elétrons devem dividir o mesmo sistema orbital. O princípio de exclusão de Pauli diz que dois elétrons não podem dividir o mesmo estado. Desta forma a metade dos elétrons do sistema deverá ir para orbitais com energias superiores às energias de valência. Por esta razão, às vezes o núcleo repulsivo é chamado de "interação de exclusão de Pauli".

A maneira mais simples e antiga de representar o núcleo repulsivo para átomos é usando o modelo de esfera rígida. Neste modelo os átomos têm um raio característico e não podem se sobrepor. Esta porém é uma maneira rude de representação, pois na realidade os sólidos e líquidos são compressíveis, Uma maneira mais realística de representar o custo energético para aproximações muito próximas é utilizar um termo que varie com r^{-12}.

O termo de repulsão cai muito rapidamente quando a distância entre dois átomos aumenta, mas contrariamente fica muito grande em distâncias curtas. Esta aproximação é normalmente usada para função de energia potencial de proteínas. Quando uma precisão maior é necessária, adota-se um modelo de dois parâmetros: $A_{ij}\exp(-B_{ij}r_{ij})$,

Este termo, em conjunto com a representação para a dispersão por um termo R_{ij}^6, é comumente conhecido como "potencial de Buckingham". Ele fornece uma representação mais realística particularmente em distâncias curtas. Contudo, não é normalmente utilizado para simulações macromoleculares porque introduz complexidade computacional com o termo exponencial.

O Potencial de Lennard-Jones e o Raio de Van der Waal

Os termos de dispersão são colocados juntos no potencial de Lennard-Jones:

$$\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}. \tag{2.10}$$

Esta equação pode ser reescrita em uma forma mais instrutiva, escolhendo-se o caso da interação entre dois átomos do mesmo tipo:

$$V = E * \left[\left\{ \frac{2R*}{r} \right\}^{12} - 2 \left\{ \frac{2R*}{r} \right\}^{6} \right]. \tag{2.11}$$

O mínimo da função é em $r = 2R*$ e possui energia $-E*$. A distância $R*$ é conhecida como o raio de Van der Waal e $E*$ é o potencial de Van der Waal. É importante notar que a interação Lennard-Jones entre átomos descarregados é menos atrativa que aquela entre grupos carregados. A diferença é que a contribuição eletrostática dominará a interação Lennard-Jones. Em casos onde grupos descarregados formam estruturas compactas as energias de Van der Waal são freqüentemente citadas como estabilizadoras da conformação.

Efeito do solvente e interações hidrofóbicas

O fato de as proteínas normalmente estarem inseridas em um meio aquoso complica consideravelmente o entendimento das interações dos diferentes grupos. A seguir vamos examinar duas contribuições importantes do solvente sobre as interações com proteínas.

Efeito Dielétrico

Quando duas cargas elétricas interagem no vácuo, a energia da interação entre els é dada pela leio de Coulomb:

$$V = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}. \tag{2.12}$$

Entretanto, se as cargas estão em um meio preenchido por algum material, a energia é reduzida por um fator conhecido como constante dielétrica do meio. Neste caso, a energia é determinada por:

$$V = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}, \tag{2.13}$$

onde ε_r é a constante dielétrica do meio.

A origem do efeito dielétrico deve-se ao fato de o campo elétrico polarizar o material que ele envolve, Podemos imaginar que o meio é composto por um grande número de dipolos microscópicos e estes serão então alinhados com o campo elétrico. Como conseqüência teremos um campo elétrico oposto ao campo original, o que causará redução do potencial elétrico e da energia de interação.

Nas simulações de proteínas, uma aproximação comum é a inclusão de um grande número de moléculas de água em simulações de dinâmica molecular.

Efeito Hidrofóbico e Entropia

O efeito hidrofóbico é a observação de que as moléculas apolares tendem a formar agregados na presença de água, como gotas de óleo por exemplo. Este efeito não pode ser facilmente modelado por interações eletrostáticas e normalmente inclui-se um termo na função de potencial de energia que represente este efeito. Esta interação é de grande importância para o dobramento de proteínas.

Do fato de as moléculas hidrofóbicas tenderem a se agrupar no interior da molécula decorre a constatação de que nem sempre as conformações de menor energia corresponderem à conformação nativa. Ao utilizarmos métodos de minimização de energia para a busca da conformação nativa de determinado polipeptídeo, podemos chegar à conclusão que determinada conformação com moléculas hidrofóbicas na superfície poderia ser preferencial em função de sua baixa energia eletrostática. Porém tal conformação implicaria na formação de agrupamentos de moléculas de água em forma de gaiola (através de pontes de hidrogênio) ao redor das moléculas hidrofóbicas. Tais estruturas implicariam em uma menor entropia, e são portanto menos prováveis. Portanto a forma nativa de um polipeptídeo é determinado pelo balanço entre minimização de energia e maximização de entropia.

2.2.3 Tipos de Campos de Força (Funções Potencial de Energia)

A credibilidade em um cálculo de mecânica molecular é dependente das equações de energia potencial e dos valores numéricos dos parâmetros que são utilizados nestas equações. Também para minimização de energia por mecânica molecular ou para simulações moleculares a qualidade do campo de força e de outros parâmetros que controlam os cálculos definem a qualidade dos resultados computados [BA91, GS91].

Campo de força ou função de energia potencial é o conjunto de termos de energia, cada qual condizente a um tipo de interação, que, reunidos com os parâmetros devidamente escolhidos, descrevem a energia das moléculas.

Os campos de força devem satisfazer dois critérios: (1) reproduzir as estruturas experimentais com uma certa precisão, e (2) as estruturas cristalinas correspondentes ao mínimo de energia encontrado para o potencial devem representar estruturas cristalinas possíveis. Idealmente um campo de força deve ser simples, rápido (computacionalmente barato), transferível e o mais preciso possível.

Existem vários tipos de campo de força, os quais podem ser classificados segundo as características listadas a seguir:

- tipo de componente que será simulado: proteínas, carboidratos ou polinucleotídeos;
- tipo de ambiente do componente de interesse: fase gás, solução aquosa ou não polar;
- alcance dos termos de interação no campo de força;
- forma funcional dos termos de interação;
- tipo de parâmetros de ajuste, isto é, se os parâmetros são teóricos ou experimentais.

As principais diferenças entre os campos de força em uso são: a forma funcional de cada termo de energia, os números de termos cruzados incluídos e o tipo de informação que é utilizada para o ajuste dos parâmetros.

As funções de energia potencial mais comumente utilizadas para peptídeos são:

- ECEPP [MMBS75, SNS84];
- CHARMM [BBO$^+$83];
- GROMOS [HBvGP84];
- AMBER [WK81];
- OPLS [JMTR96];
- MM3 [AYL89a, AL89, AYL89b];

Estas funções de energia potencial são muito parecidas em sua forma funcional, mas diferem em detalhes de suas parametrizações. A maior diferença na forma funcional é que o CHARMM e o AMBER consideram a molécula flexível, e o ECEPP é concebido em termos de uma geometria fixa e não prevê alongamento e contração das ligações ou a curvatura dos ângulos de ligação. CHARMM e AMBER usam coordenadas cartesianas dos átomos como suas variáveis independentes enquanto ECEPP usa como variáveis independentes os ângulos diedrais.

Embora os resultados obtidos com as funções de energia potencial sejam apenas aproximados da realidade, eles possuem uma grande vantagem: são computacionalmente baratos, Isto permite a introdução de representações mais realísticas do ambiente, como por exemplo água envolvendo uma proteína que está sendo simulada.

O método do campo de força ignora o movimento dos elétrons e calcula a energia de um sistema como função das suas posições nucleares somente. Por este motivo é chamada de função de potencial de energia. A maior vantagem dos campos de força é a velocidade dos cálculos, o que permite que cálculos de estruturas grandes como proteínas possam ser realizados em computadores com PCs por exemplo. A capacidade de lidar com grande número de partículas também torna este o único método para realizar simulações onde os efeitos do solvente devem ser estudados. Para sistemas onde são disponibilizados um bom conjunto de parâmetros é possível realizar uma predição da geometria e da energia com uma aproximação muito boa da realidade.

2.2.4 Campo de Força MM3

O campo de força MM3 é muito poderoso para a simulação de moléculas orgânicas. O mesmo foi desenvolvido a partir do campo de forças MM2, do qual herdou muitas características, podendo ser classificado como um campo de força pertencente à classe dos campos de força "complexos". Deve-se tal fato à inclusão de termos cruzados para a representação de ângulos e ligações.

A energia potencial representada pelo campo de força MM3 é, assim como em outros campos de força, composta da soma de vários termos de energia potencial, sendo cada um deles o representante de um tipo de interação. A seguir a relação de cada termo com sua referida equação.

Alongamento da ligação covalente:

$$E_s = 71.94K_s(l - l_0)^2 \left[1 - 2.55(l - l_0) + (\frac{7}{12} \cdot 2.55(l - l_0)^2\right] \tag{2.14}$$

Encurvamento de ângulos

$$\begin{aligned} E_\theta &= 0.021914k_\theta(\theta - \theta_0)^2 \left[1 - 0.014(\theta - \theta_0) + 5.6 \cdot 10^{-5}(\theta - \theta_0)^2 - \right. \\ &\quad \left. 7.0 \cdot 10^{-7}(\theta - \theta_0)^3 + 9.0 \cdot 10^{-10}(\theta - \theta_0)^4\right] \end{aligned} \tag{2.15}$$

Torção entre ângulos:

$$E_\omega = \frac{V_1}{2}(1 + \cos\omega) + \frac{V_2}{2}(1 - \cos 2\omega) + \frac{V_3}{2}(1 + \cos 3\omega) \tag{2.16}$$

Interação Encurvamento-Alongamento:

$$E_{s\theta} = 2.51118K_{s\theta}\left[(l - l_0)(l' - l'_0)\right](\theta - \theta_0) \tag{2.17}$$

Interação Alongamento-Torção:

$$E_{\omega s} = 11.995\frac{K_{\omega s}}{2}(l - l_0)(1 + \cos 3\omega) \tag{2.18}$$

Encurvamento-Encurvamento:

$$E_{\theta\theta'} = -0.021914K_{\theta\theta'}(\theta - \theta_0)(\theta' - \theta'_0) \tag{2.19}$$

Interação Van der Waal:

$$E_{vdw} = \epsilon\left\{-2.25\left(\frac{r_v}{r}\right)^6 + 1.84 \cdot 10^5 \cdot e^{\left[-12.00\frac{r}{r_v}\right]}\right\} \tag{2.20}$$

2.3 Dobramento de Proteínas

As indicações através de medidas termodinâmicas são que as proteínas nativas são fracamente estáveis em condições fisiológicas. A energia livre necessária para sua desnaturação é de aproximadamente $0,4kJ.mol^{-1}$ por resíduo de aminoácido, o que leva a uma energia de $80kJ.mol^{-1}$ para uma proteína com 200 resíduos. Cada um dos efeitos não covalentes, como efeitos hidrofóbicos, interações eletrostáticas e pontes de hidrogênio, pode chegar a milhares de kilojoules por mol em uma molécula inteira de proteína. A conseqüência é que a estrutura de uma proteína é o resultado de um balanço delicado entre poderosas forças concorrentes. A seguir, discutiremos quais são as forças que estabilizam a proteína e como elas atingem um estado final dobrado mais estável.

As estruturas das proteínas são regidas principalmente por efeitos hidrofóbicos e, em menor grau, por interações entre resíduos polares e outros tipos de ligações.

O efeito hidrofóbico que faz com que substâncias apolares minimizem seus contatos com a água é o principal determinante da estrutura de proteínas nativas. A agregação de cadeias laterais apolares no interior de uma proteína é favorecida

pelo aumento de entropia das moléculas de água que, de outra forma, iriam formar gaiolas ordenadas em torno de grupos hidrofóbicos.

Interações Eletrostáticas

Em interiores de agregados de proteínas nativas as forças de Van der Waal, que são relativamente fracas, são uma influência estabilizadora importante. Como estas forças atuam somente a pequenas distâncias, elas desaparecem quando a proteína é desenrolada,

As pontes de hidrogênio possuem apenas uma pequena contribuição na estabilidade das mesmas. Isto ocorre porque os grupos que formam as pontes de hidrogênio em uma proteína desenrolada formam pontes de hidrogênio energeticamente equivalentes com moléculas de água. As pontes de hidrogênio dão a tônica final na estrutura terciária por selecionarem a estrutura nativa singular de uma proteína dentre um número relativamente pequeno de conformações hidrofobicamente estabilizadas.

A associação de dois grupos iônicos de proteínas de carga oposta é designado par iônico ou ponte salina. Cerca de 75% dos resíduos com carga nas proteínas são membros de pares iônicos localizados principalmente na superfície das proteínas. Apesar das fortes atrações eletrostáticas entre membros com cargas opostas de um par iônico essas interações contribuem pouco para a estabilidade das proteínas nativas. Assim ocorre porque a energia livre das interações de carga do par iônico normalmente não é suficiente para compensar a perda de entropia das cadeias laterais e a perda da energia livre de solvatação. quando esses grupos com carga formam um par iônico. Esses fatores explicam a razão pela qual os pares iônicos são pouco conservados entre proteínas homólogas.

Desnaturação e Renaturação de Proteínas

O aquecimento com pouca variação de temperatura causa alterações abruptas de propriedades conformacionais sensíveis.

As variações de pH alteram o estado iônico das cadeias laterais de aminoácidos alterando a distribuição de cargas e a existência de pontes de hidrogênio.

Os detergentes associam-se com os resíduos apolares de uma proteína interferindo com as interações hidrofóbicas responsáveis pela estrutura nativa delas. O íon guanidina e a uréia em concentrações entre $5M$ e $10M$ são desnaturantes protéicos mais comumente usados. Eles atuam aumentando a solubilidade de substâncias apolares na água. Isto deve-se a sua habilidade em romper interações hidrofóbicas.

Rotas de Dobramento de Proteínas

Poderíamos imaginar que a proteína atinge o dobramento de sua conformação nativa procurando cada uma das conformações possíveis aleatoriamente até encontrar aquela que é correta.

No entanto, um cálculo simples realizado por Cyrus Levinthal demonstrou que tal não pode ocorrer. Imagine uma proteína com n resíduos que possui então 2^n ângulos de torção ϕ e ψ, e que cada um possua três conformações estáveis fornecendo $3^{2n} \approx 10^n$ conformações possíveis para a proteína (desconsiderando as cadeias laterais). Como a velocidade em que ligações simples são reorientadas é de $10^{-13}s$

o tempo t em segundos para que a proteína experimente todas as conformações possíveis é

$$t = \frac{10^n}{10^{13}}. \tag{2.21}$$

Para uma proteína pequena de 100 resíduos o tempo seria de $10^{87}s$ o que é muito maior do que a idade estimada do universo $(6 \cdot 10^{17}s)$.

As proteínas dobram-se para atingir sua conformação em poucos segundos, evidenciando que elas se dobram utilizando rotas diretas. Então uma proteína ao dobrar-se apresenta uma quebra abrupta de energia livre.

O estágio inicial do dobramento da proteína é extremamente rápido, sendo que a maior parte da estrutura secundária de proteínas pequenas já está em sua conformação nativa após $5ms$ do início do dobramento.

Nos próximos 5 a $1000ms$ a estrutura secundária torna-se estável e a estrutura terciária começa a ser formada. No estágio final do dobramento, que para as pequenas proteínas de domínio único ocorre nos próximos segundos, a proteína sofre uma série de movimentos complexos por meio dos quais adquire a organização rígida das suas cadeias laterais e pontes de hidrogênio internas, enquanto as moléculas de água remanescentes são expelidas do interior hidrofóbico.

O dobramento, assim como a desnaturação, parece ser um processo cooperativo. Uma proteína que está dobrando-se deve, necessariamente, progredir de um estado de alta energia e entropia, para um estado de baixa energia e entropia. Um polipeptídeo não dobrado apresenta muitas possibilidades de conformação (alta entropia) com o dobramento em número cada vez menor de conformações a sua entropia e sua energia livre diminuem. O diagrama de energia entropia não é um vale suave, mas uma paisagem dentada. Pequenos buracos ou elevações representam conformações que são temporariamente aprisionadas até que, por ativação térmica aleatória, consigam sobrepor a barreira de energia livre e possam progredir para uma conformação de menor energia. Evidentemente as proteínas evoluíram para atingir rotas de dobramento eficientes e conformações nativas estáveis. Contudo, proteínas dobradas erroneamente ocorrem na natureza e acredita-se que seu acúmulo possa ser a causa de várias doenças.

2.4 Métodos Tradicionais de Otimização

Otimização é a procura por um valor ótimo. Quando falamos de otimização estamos falando de procurar um valor máximo ou mínimo para uma função, ou em outras palavras, achar a melhor solução para um problema.

Este problema pode ser unidimensional, quando a função possuir somente uma variável ou multidimensional, quando existem várias variáveis para serem ajustadas. À função para a qual procura-se o valor ótimo, máximo ou mínimo, dá-se o nome de função objetivo

$$f(\vec{x}) = f(x_1, x_2, \ldots, x_n) \tag{2.22}$$

com os limites $l_i \leq x_i \leq u_i$, para $1 \leq i \leq n$, definindo o domínio de cada variável.

Em um problema de otimização estamos interessados em encontrar o máximo ou o mínimo para a referida função, o que pode ser definido como:

Seja S o conjunto das possíveis soluções da função definida por $f : S \rightarrow \Re$; encontre uma solução $(x_1^*, x_2^*, \ldots, x_n^*) \in S$ tal que $f(x_1^*, x_2^*, \ldots, x_n^*)$ satisfaça o critério

$$f(x_1^*, x_2^*, \ldots, x_n^*) \leq f(x_1, x_2, \ldots, x_n), \forall (x_1, x_2, \ldots, x_n) \in S \tag{2.23}$$

para minimização, e

$$f(x_1^*, x_2^*, \ldots, x_n^*) \geq f(x_1, x_2, \ldots, x_n), \forall (x_1, x_2, \ldots, x_n) \in S \tag{2.24}$$

para maximização. Além de encontrarmos um valor máximo ou mínimo para uma função, ao otimizar devemos fazê-lo usando um processo que seja em si otimizado. Podemos até mesmo concluir que encontrar o valor ótimo absoluto com um processo ineficiente, ou seja, custoso em termos computacionais, é pior do que encontrar rapidamente um ponto muito próximo do ótimo.

Existem atualmente vários métodos de otimização que já foram amplamente estudados. A seguir faremos uma breve descrição de alguns destes métodos sem nos preocuparmos com testes de validação para os mesmos.

Métodos Analíticos

Um dos principais métodos de otimização é o método analítico baseado em cálculo. Este pode ser direto ou indireto. No método indireto resolve-se um conjunto de equações não lineares resultantes de se igualar o gradiente da função objetivo a zero, procurando assim um extremo local, mas fazendo a função "saltar"e "mover-se"na direção relacionada com o gradiente local, que é o método conhecido por descida de gradiente.

Este método demonstra falta de robustez porque é local, isto é, inicia a busca a partir de um ponto e encontra a melhor solução na vizinhança deste ponto. Obviamente se este ponto inicial escolhido estiver próximo ao extremo global para o problema, este método o encontrará. Mas se o ponto escolhido estiver próximo de uma solução local, este método encontrará esta solução e ficará preso nela, não mais tendo condições de buscar o extremo global. Como em problemas reais geralmente existem muitas soluções locais e somente uma delas é global, a probabilidade de este método encontrar que não seja global é grande. Outra falha deste método é que ele necessita que existam derivadas definidas, o que não é típico em problemas reais, que geralmente possuem espaços de busca com muitos picos, ruídos e descontinuidades. Devido a estas restrições este método de otimização é aplicado a uma classe restrita de problemas.

Esquemas Enumerativos

Outro método também amplamente estudado são os esquemas enumerativos, como por exemplo a busca exaustiva, que consistem simplesmente em enumerar e observar todos os pontos do espaço de busca, para então encontrar a melhor solução.Evidentemente este método é impraticável para espaços de busca grandes.

Outro método que também já teve seu momento de grande popularidade são os métodos aleatórios. Estes métodos, como por exemplo a busca cega,*beam search* e *Hill-Climbing*, criam uma árvore de busca e procuram pela solução percorrendo esta árvore. Testes com estes métodos mostraram que eles também não possuem um

bom desempenho, ou seja, se o espaço de busca for muito grande estes algoritmos falham no requisito da eficiência e eventualmente encontram extremos locais.

Busca Tabu

É um método utilizado principalmente em problemas de otimização combinatória que procura evitar mínimos locais na busca. Este método inclui algumas técnicas heurísticas nas suas próprias regras de operação e por isso pode ser caracterizado como um procedimento meta-heurístico [HTdW95].

A busca começa com uma única solução inicial, e a partir dela através de um procedimento chamado de *movimento* é gerado um conjunto de soluções. Este *movimento* para a criação de soluções é controlado por restrições tabu e critérios de aspirações. As restrições tabu e os critérios de aspirações são armazenados em listas circulares o que significa que quando um novo elemento é adicionado na lista o ítem mais antigo é eliminado.

As restrições tabu são as soluções que já foram visitadas na busca e que devem ser evitadas nas próximas iterações do algoritmo.

Os critérios de aspirações são as soluções que estão na lista tabu e apresentam uma possibilidade de serem uma solução aproximada para o problema, se após uma longa busca nenhuma solução melhor for encontrada a lista de aspirações contém critérios que eliminam as restrições tabu, Isto significa que um movimento proibido por uma restrição tabu pode passar a ser permitido caso satisfaça a condição de aspiração, permitindo que uma solução aproximada para o problema seja aceita.

Uma característica adicional encontrada na busca tabu é que a função objetivo pode ser substituída por uma nova função objetivo que permite a introdução de intensificação e diversificação da busca.

Recozimento Simulado

Uma maneira de se evitar o aprisionamento do sistema em mínimos locais, baseada em idéias de mecânica estatística e ciência dos materiais, foi proposta por Kirkpatrick, Gellat e Vecchi [KGV83], e batizada de *Simulated Annealing* (recozimento simulado) devido às semelhanças com o processo físico de têmpera.

Da ciência dos materiais sabe-se que para construir um sólido de estrutura cristalina perfeita deve-se fundir o material e depois diminuir lentamente a temperatura do sistema, demorando um longo tempo. Se o sistema é resfriado rapidamente, a amostra resultante apresentará vários defeitos estruturais sem nenhuma ordem de longo alcance.

Para entender como são feitas as mudanças na configuração do sistema em questão, considere um sistema com muitos átomos, a uma temperatura finita, em contato com um banho térmico. Se executamos um pequeno deslocamento ΔX em um átomo, resultando em uma variação na energia do sistema de ΔE, aceitamos esta mudança de a energia diminui ($\Delta E < 0$), e este arranjo atômico passa a ser a nova configuração do sistema. Se a energia aumentar ($\Delta E > 0$), aceitamos a nova configuração com a probabilidade

$$P(\Delta E) = e^{\frac{-\Delta E}{k_b T}} \quad . \tag{2.25}$$

Nesta relação, T é a pseudo-temperatura (doravante denominada simplesmente de temperatura) e k_b é um parâmetro que determina o cronograma de têmpera, isto é,

como a probabilidade varia com a temperatura T. A escolha de $P(\Delta E)$ tem como conseqüência a evolução do sistema com a distribuição de Boltzmann.

Baseado nestes dois conceitos, Kirkpatrick e Vecchi propuseram um algoritmo de otimização iterativa, onde o sistema é iniciado a uma temperatura T_0 bastante alta comparada com as escalas de energia envolvidas. A temperatura permite movimentos que aumentam a energia de uma quantidade ΔE com probabilidade $P(\Delta E)$ dada em 2.25. O algoritmo para a implementação do método é o seguinte:

1. Iniciamos o sistema com temperatura T_0
2. Com a temperatura fixa, um ou mais graus de liberdade do sistema são modificados por um valor ΔX com probabilidade gaussiana $P(\Delta X)$.
3. Cada mudança ΔX leva o sistema de uma energia E para outra com energia $E'i = E + \Delta E$. Aceitamos esta mudança com probabilidade $P(\Delta E) = e^{\frac{-\Delta E}{k_b T}}$.
4. Após executarmos as N mudanças (aceitas ou não) no sistema, diminuímos a temperatura de acordo com uma regra pré estabelecida.
5. Retornamos ao passo 2 até atingirmos $T \approx 0$, onde se espera que o sistema tenha alcançado o estado fundamental.

No caso do Recozimento Simulado, no inicio do processo, podem ser feitos alguns movimentos decrescentes, ou seja, o método aceita possíveis soluções de maior custo (no caso do problema de minimização), na expectativa de escapar de um mínimo local. A idéia é explorar suficientemente todo o espaço do problema logo no início, para que a solução final seja relativamente independente do estado inicial.

Quatro ingredientes são necessários para a implementação do algoritmo de Recozimento Simulado:

- descrição concisa da configuração do sistema;
- gerador aleatório de "movimentos"ou rearranjos dos elementos em uma configuração;
- função objetivo quantitativa contendo os compromissos a serem assumidos;
- cronograma de têmpera das temperaturas e intervalos de tempo para os quais o sistema evoluirá.

A vantagem deste método em relação à descida de gradiente é que ele não necessita usar a derivada para ser executado o que permite que sejam feitas otimizações de funções descontínuas.

2.5 Técnica Monte Carlo

Métodos MC são utilizados para integração numérica, otimização global, teoria de filas, e solução de grandes sistemas de equações diferencias lineares parciais ou equações integrais. Métodos MC são largamente empregados em física estatística e química, em problemas que envolvem o estudo do comportamento de milhares de átomos no tempo e no espaço [Sch02].

A maior parte dos algoritmos MC manipula variáveis aleatórias uniformemente distribuídas e independentes [Sch02]. Ou seja, assume-se que as variáveis independentes obedecem a uma distribuição de probabilidade (DP) ρ que satisfaz $\rho_u(x) = 1$ para x pertencente ao intervalo $[0, 1]$ e $\rho_u(x) = 0$ caso contrário. Destas variáveis aleatórias uniformes é possível obter outras distribuições não uniformes teóricas (como as distribuições normal, Gamma ou Poisson) ou empíricas.

A técnica MC foi originalmente desenvolvida para estimar integrais sem solução analítica. Dada uma função contínua $f(x)$, a integral

$$I(x) = \int_0^x f(x)d(x) \tag{2.26}$$

pode ser estimada a partir da seguinte forma. Se escolhermos um número randômico real h uniformemente distribuído entre 0 e x, e outro número ν entre 0 e o valor máximo de $f(x)$ para o intervalo de 0 a x, e plotarmos o ponto (h, ν) no gráfico de $f(x) \times x$, a probabilidade deste ponto estar abaixo da linha do gráfico da função é dada por $I(x)/x$. Portanto se gerarmos randomicamente uma grande quantidade N de pontos e contarmos a quantidade M de pontos que ficaram abaixo da linha do gráfico de $I(x)$, podemos estimar $I(x)$ por:

$$I(x) = \lim_{N\to\infty} \frac{Mx}{N} \ . \tag{2.27}$$

Em modelos de sistemas mecânicos como um gás dentro de um recipiente ou uma macromolécula em uma "caixa"de água, a simulação MC de equilíbrio térmico calcula o valor esperado (expectância) $\langle Q \rangle$ de uma variável observável Q, como a energia interna U do sistema por exemplo. A maneira ideal de calcular tal valor esperado é calcular a média dos valores da variável para cada estado μ do sistema ponderada pela probabilidade de Boltzmann de cada estado. Então o valor esperado é dado por

$$\langle Q \rangle = \frac{\sum_\mu Q_\mu e^{-\beta E_\mu}}{\sum_\mu e^{-\beta E_\mu}} \ , \tag{2.28}$$

mas só é viável para sistemas muito pequenos. As técnicas MC utilizam apenas um sub-conjunto dos estados escolhidos de maneira randômica de alguma DP p_μ. Dado o subconjunto de estados $M = \mu_1 \ldots \mu_M$, uma estimativa da quantidade $\langle Q \rangle$ é

$$Q_M = \frac{\sum_{i=1}^M Q_{\mu_i} p_{\mu_i}^{-1} e^{-\beta E_{\mu_i}}}{\sum_{j=1}^M p_{\mu_j}^{-1} e^{-\beta E_{\mu_j}}} \ , \tag{2.29}$$

onde Q_M é denominado o estimador de $\langle Q \rangle$, e quando $M \to \infty$ temos $Q_M = \langle Q \rangle$.

Se assumirmos a DP uniforme, ou seja, todos os estados com a mesma probabilidade p_μ, a equação (2.29) se reduz a:

$$Q_M = \frac{\sum_{i=1}^M Q_{\mu_i} e^{-\beta E_{\mu_i}}}{\sum_{j=1}^M e^{-\beta E_{\mu_j}}} \ . \tag{2.30}$$

Porém, em sistemas onde podemos somente nos valer de pequenas amostras do conjunto de estados (o que é usual em se tratando de simulação de sistemas físicos moleculares), se utilizarmos a DP uniforme o resultado será uma representação muito pobre dos estados importantes do sistema como mínimos de energia por exemplo. A

técnica para a escolha dos estados relevantes dentre um grande número de estados é denominada amostragem por importância.

De acordo com [NB99] um sistema físico real não passa por todos os estados possíveis durante o tempo em que são observados[2], mas realiza uma amostra muito pequena dos estados possíveis. Os sistemas físicos reais realizam uma espécie de MC de suas próprias propriedades, o que reforça a corretude da utilização de técnicas MC para obter cálculos razoáveis das propriedades de um sistema a partir de amostras pequenas mas representativas do sistema.

A estratégia para a amostragem por importância é ao invés de se escolher os estados M de acordo com uma DP uniforme, utilizar a probabilidade de Boltzmann correta de cada estado. Agora a probabilidade de que um estado μ seja escolhido é $p_\mu = Z^{-1}e^{-\beta E_\mu}$ e o estimador para $\langle Q \rangle$, equação (2.30), se reduz a

$$Q_M = \frac{1}{M}\sum_{i=1}^{M} Q_{\mu_i} \; . \tag{2.31}$$

Esta expressão é bem mais simples e funciona melhor do que (2.30), pois esta definição de Q_M leva em consideração o tempo real em que o sistema está nos estados mais prováveis.

Porém, para que a expressão (2.31) tenha utilidade é necessário gerar uma amostra randômica de estados que esteja de acordo com a DP de Boltzmann. Para tanto as técnicas MC utilizam processos Markov como gerador do grupo de estados a ser utilizado.

Neste contexto, um processo de Markov é um mecanismo que, dado um sistema no estado μ, gera um novo estado ν do sistema, com a probabilidade de transição $P(\mu \to \nu)$. Para um processo Markov, duas restrições devem ser satisfeitas para todas as probabilidades de transição: (1) não variar no tempo, e (2) depender somente das propriedades dos estados μ e ν, e não de quaisquer outros estados pelos quais o sistema tenha passado anteriormente. Ou seja, a probabilidade do processo Markov estar no estado μ e produzir o estado ν é a mesma sempre que o estado atual for μ. As transições de probabilidade $P(\mu \to \nu)$ devem satisfazer também à seguinte restrição:

$$\sum_{\nu} P(\mu \to \nu) = 1 \; , \tag{2.32}$$

ou seja, o processo Markov gera um estado ν quando recebe um estado μ. É importante observar que a probabilidade de haver transição de estado para o mesmo estado ($P(\mu \to \mu)$) pode ser diferente de zero.

Em uma simulação MC o processo Markov é utilizado repetidamente gerando uma cadeia de Markov de estados. O processo Markov deve ser tal que após suficiente número de iterações a partir de qualquer estado do sistema produza uma sucessão

[2]Se tomarmos como exemplo um litro de gás nas CNTP, o sistema conterá aproximadamente 10^{22} moléculas. As velocidades típicas para estas moléculas são da ordem de $100m/s$, dado o comprimento de onda da ordem de $10^{-10}m$. Cada molécula terá então ao redor de 10^{27} estados quânticos diferentes possíveis dentro do recipiente de 1 litro, e o gás terá ao redor de $(10^{27})^{10^{22}}$ estados possíveis. As moléculas mudam de estado a cada colisão entre si e com as paredes do recipiente a uma taxa de 10^9 colisões por segundo, ou 10^{31} mudanças de estado por segundo para todo o gás. A esta taxa, levaria $10^{10^{23}}$ vezes a idade do universo para que o gás passasse por todos os estados possíveis [NB99].

de estados com probabilidade dada pela DP de Boltzmann. Quando isto acontece o sistema atingiu o ponto de equilíbrio, e para alcançar este tipo de resultado, precisamos de mais duas restrições ao processo Markov: as condições de ergodicidade e de balanço detalhado.

A condição de ergodicidade é a capacidade do processo Markov atingir qualquer estado a partir de outro estado qualquer. Na DP de Boltzmann todos os estados ν tem uma probabilidade p_ν, e se um destes estados é inatingível a partir de um estado μ, não é possível alcançar o objetivo de gerar uma sucessão de estados com probabilidade de acordo com a DP de Boltzmann.

Na prática o método MC pode colocar a maioria das probabilidades de transição em zero, desde que haja um caminho de transições com probabilidade diferente de zero entre cada dois estados da amostra escolhida.

A condição de balanço detalhado se destina a garantir que o sistema seja simetricamente reversível no tempo, ou seja, que não hajam ciclos limitantes e que a taxa total de vezes que uma transição ocorre é igual a taxa de vezes que a transição reversa ocorre. A condição de balanço detalhado

$$p_\mu \cdot (P(\mu \to \nu)) = p_\nu \cdot (P(\nu \to \mu)) \ , \tag{2.33}$$

uma vez satisfeita faz com que o sistema sempre tenda para a probabilidade p_μ conforme $t \to \infty$. Esta condição está de acordo com os sistemas físicos reais, pois estes quase sempre obedecem à condição de balanço detalhado.

Podemos portanto fazer com que a DP dos estados gerados pelo processo Markov tenda a qualquer DP escolhendo um conjunto de probabilidades de transição que satisfaça a equação (2.33). Se desejamos a DP de Boltzmann, simplesmente utilizamos esta DP para determinar as probabilidades da condição de equilíbrio detalhado:

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = e^{-\beta(E_\nu - E_\mu)} \ . \tag{2.34}$$

Por fim, a proporção de aceitação é baseada no fato de que podemos definir transições de probabilidade do tipo $P(\mu \to \mu)$ com valores diferentes de zero. Isto permite quebrar a probabilidade de transição em duas partes:

$$P(\mu \to \nu) = g(\mu \to \nu) A(\mu \to \nu) \ . \tag{2.35}$$

A quantidade $g(\mu \to \nu)$ é a probabilidade de seleção, dado um estado inicial μ, de que um novo estado ν, será gerado, e a quantidade $A(\mu \to \nu)$ é a probabilidade de aceitação do novo estado gerado.

A proporção de aceitação confere liberdade para a escolha do algoritmo de geração de novos estados sem prejuízo das restrições levantadas anteriormente, e obter o conjunto desejado de probabilidades de transições. Ou seja, ajustando os valores das probabilidades $P(\mu \to \mu)$ podemos fazer com que o conjunto de probabilidades de transição tenha o conjunto de valores que mais nos convém, uma vez que a equação (2.34) determina tão somente a proporção:

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{g(\mu \to \nu) A(\mu \to \nu)}{g(\nu \to \mu) A(\nu \to \mu)} \ . \tag{2.36}$$

A outra restrição, a equação 2.32, também é satisfeita pela equação anterior.

Portanto, para construirmos um algoritmo MC, primeiro criamos um algoritmo que gera estados novos ν randomicamente a partir dos estados μ com um conjunto de probabilidades $g(\mu \to \nu)$, e depois aceitamos ou rejeitamos estes estados com grau de aceitação $A(\mu \to \nu)$. Isto satisfará todos os requisitos para as probabilidades de transição, e portanto produzir uma lista de estados que, quando o algoritmo atingir o equilíbrio, serão similares à DP de Boltzmann.

O algoritmo Metropolis é o mais famoso e largamente utilizado algoritmo MC [NB99], e foi introduzido por Nicolas Metropolis em um artigo de 1953 sobre simulação de um gás modelado por esferas sólidas. No algoritmo Metropolis os N estados possíveis de serem atingidos tem todos a mesma probabilidade de geração

$$g(\mu \to \nu) = \frac{1}{N} \, , \tag{2.37}$$

e a probabilidade de geração dos demais estados é zero. Com estas probabilidades de geração, a condição de balanço detalhado da equação 2.34 passa a ser

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{(g(\mu \to \nu))(A(\mu \to \nu))}{(g(\nu \to \mu))(A(\nu \to \mu))} = \frac{A(\mu \to \nu)}{A(\nu \to \mu)} = e^{-\beta(E_\nu - E_\mu)} \, . \tag{2.38}$$

Para obter a maior taxa de aceitação possível, e levando em consideração que $A(\mu \to \nu)$, sendo uma probabilidade, não pode ser maior que 1, podemos escolher

$$A(\mu \to \nu) = e^{-\frac{1}{2}\beta(E_\nu - E_\mu + \Delta E_{max})} \, . \tag{2.39}$$

Na equação acima, a probabilidade de aceitação de transição entre os estados de máxima e mínima energia do sistema é 1, porém cai rapidamente com a redução da taxa de variação negativa da energia, e é praticamente zero para variações positivas de energia. O algoritmo MC com probabilidades de aceitação dadas por (2.39) seria muito ineficiente (Figura 2.13). A simulação MC teria baixa probabilidade de aceitação de transições de estado e pouca ou nenhuma chance de escapar de mínimos locais através de transições para estados de maior energia.

Se determinarmos que a probabilidade de aceitação da transição para um estado de menor energia seja a maior possível, no caso 1, podemos depois ajustar a probabilidade de aceitação da outra transição para respeitar a condição de balanço detalhado da equação 2.38. Suponhamos por exemplo que a energia do estado μ é menor do que a energia do estado ν: $E_\mu < E_\nu$. Então a maior probabilidade de aceitação entre os dois estados é $A(\nu \to \mu)$, e portanto damos a esta probabilidade o valor 1. Para satisfazermos a equação 2.38 basta agora determinarmos o valor da transição inversa: $A(\mu \to \nu) = e^{-\beta(E_\nu - E_\mu)}$. Para otimizar a busca por estado de menor energia, o algoritmo Metropolis determina que:

$$A(\mu \to \nu) = \begin{cases} e^{-\beta(E_\nu - E_\mu)} & \text{se } E_\nu - E_\mu > 0 \\ 1 & \text{caso contrário .} \end{cases} \tag{2.40}$$

Ou seja, se geramos um novo estado com energia menor do que a do estado atual, sempre aceitamos a transição. Se o novo estado gerado tiver energia maior, aceitamos a transição com a probabilidade acima. Desta forma não perdemos tempo rejeitando transições para estados de menor energia, e mantemos a probabilidade de escapar de mínimos locais aceitando transições para estados de maior energia com probabilidade dada pela DP de Boltzmann.

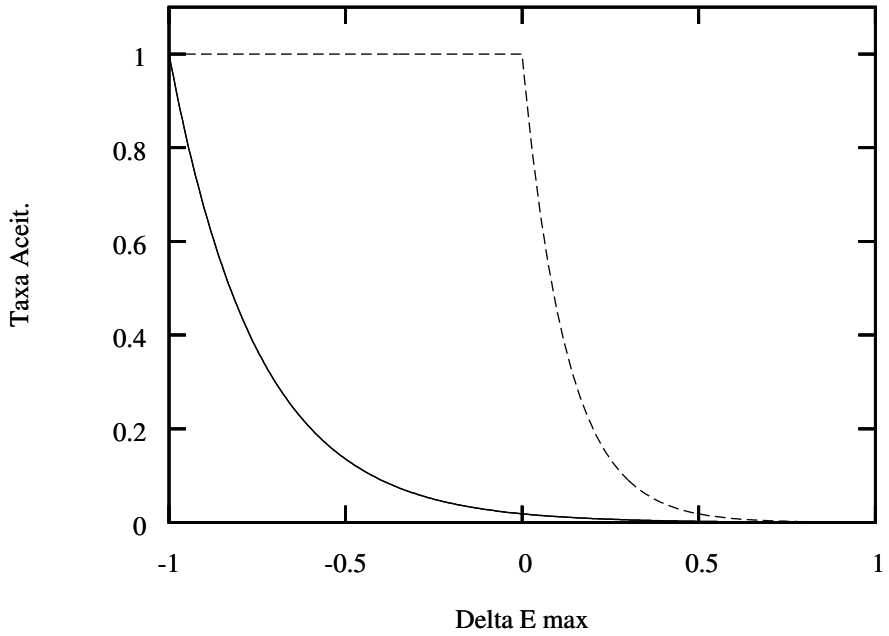


FIGURA 2.13 – A linha contínua é a taxa de aceitação da equação 2.39. A transição para um estado com redução de energia equivalente à $-\frac{1}{2}\Delta E_{max}$ tem probabilidade de ocorrer de apenas $0,13$, e a taxa de aceitação de transições para estados de maior energia é $0,02$ no máximo. No algoritmo Metropolis (equação 2.40) representado pela linha tracejada, as probabilidades de transição são as maiores possíveis para cada ΔE, respeitando-se a condição de balanço detalhado.

2.6 *Clusterização*

Neste trabalho a *clusterização* tem um papel crucial na determinação dos resultados da técnica MC. A simulação MC é inicializada várias vezes com sementes aleatórias diversas. Findas as simulações temos resultados distribuídos probabilisticamente pelos mínimos locais do modelo molecular. A *clusterização* é a ferramenta que pode informar quais conformações de proteínas localizadas em diversos mínimos locais de energia pertencem à mesma classe de solução, permitindo a escolha da conformação que melhor representa cada grupo (*cluster*).

Clusterizar é organizar dados em conjuntos de acordo com algum critério de similaridade. Dados em um conjunto (*cluster*) são mais similares entre si do que com dados pertencentes a outros *clusters*. A *clusterização* é um processo de organização não supervisionado, ou seja, a organização emerge dos dados sem nenhuma pré-classificação.

Os passos típicos de um processo de *clusterização* são:

1. Representação de padrões através de atributos. Estes atributos podem ser medidos ou extraídos dos padrões. Se utilizarmos uma proteína como exemplo, poderíamos representá-la por um ou mais atributos como número de aminoácidos, estrutura secundária, compactação, energia, hidrofobicidade, etc.

2. Definição de medidas de proximidade adequadas para os atributos escolhidos.

3. A *clusterização* propriamente dita.

Antes de entrarmos especificamente nos processos de *clusterização*, é necessário definirmos os termos que serão utilizados nesta seção, sendo eles:

- Padrão: é a instância de um dado. Um padrão $\mathbf{x}$ é representado por um vetor de atributos $\mathbf{x} = \{x_1, x_2, \ \ldots \ , x_d\}$.
- Atributo: os valores escalares dos componentes x_i do vetor $\mathbf{x}$ que representam o padrão.
- Dimensionalidade: é o número de atributos do padrão, e define a dimensão d do hiperespaço que contém os dados.
- Distância entre padrões: é o valor utilizado para aferir a similaridade entre um par de padrões. Quanto maior a distancia, menor a similaridade entre 2 padrões. A função de distância mais intuitiva é a Distância Euclidiana, dada por:

$$d_2(x_i, x_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2 \right)^{\frac{1}{2}} = \|x_i - x_j\|_2 \ . \tag{2.41}$$

  Para evitar que atributos com escalas de valores maiores sejam preponderantes, os atributos devem ser normalizados antes de se utilizar a Distância Euclidiana.

As abordagens para *clusterização* são inicialmente divididas em dois grupos principais: hierárquica e particional. A abordagem hierárquica particiona o espaço de dados recursivamente, enquanto a particional produz apenas um número fixo de partições.

A vantagem dos algoritmos hierárquicos é prescindirem de intervenção na escolha inicial de *clusters*. Porém, como têm alta complexidade computacional, não podem ser utilizados para *clusterização* de grandes conjuntos de dados ou para padrões de grande dimensionalidade.

Os Algoritmos de *Clusterização* Hierárquica produzem dendrogramas a partir de dados e são em sua maioria baseados nos algoritmos *single-link* [SS73] , *complete-link* [Kin67] e *minimum-variance* [War63], sendo os dois primeiros os mais populares [JMF99].

O algoritmo de *Clusterização* Hierárquica consiste nos seguintes passos:

1. Compute a matriz de proximidade contendo a distância entre cada par de padrões. Crie um *cluster* para cada padrão.
2. Encontre o par de *clusters* mais similar de acordo com a matriz de proximidade e una esses *clusters* em um novo *cluster*. Atualize a matriz de proximidade.
3. Se todos os padrões estão em apenas um *cluster*, pare. Caso contrário retorne para o passo 2.

De um algoritmo de *clusterização* particional se obtém simplesmente particionamento dos dados, ao invés de uma estrutura de *clusters* como a do dendrogramas resultante dos algoritmos hierárquicos. Algoritmos particionais são indicados para conjuntos de dados grandes, onde a construção de dendrogramas é computacionalmente proibitiva [JMF99]. A dificuldade dos algoritmos particionais é a necessidade de se escolher o número de *clusters*.

Nesta técnica busca-se otimizar uma função objetivo em relação a algum critério. Como a busca combinatória de todo o conjunto de possíveis *cluster* em busca do ótimo é computacionalmente impraticável, na prática se utiliza rodar varias instâncias do algoritmo com pontos iniciais diferentes. A melhor configuração é então escolhida como resultado do algoritmo de *clusterização*.

O algoritmo k-Means é popular por ser fácil de implementar e devido a sua baixa complexidade ($O(n)$, onde n é o número de padrões). Porém o algoritmo é sensível à escolha das partições iniciais. Uma escolha infeliz de condição inicial pode levar à convergência em um mínimo local.

Na Figura 2.14 as elipses representam os *clusters* formados se escolhermos A, B e C entre os sete padrões bidimensionais como os pontos de partida para a construção de 3 *clusters*. Esta configuração tem Erro Quadrado bem maior do que a *clusterização* representada pelos retângulos. Se escolhemos por exemplo os padrões A, D e F como pontos iniciais, obtemos a *clusterização* mostrada pelos retângulos, que é a configuração de mínimo global.

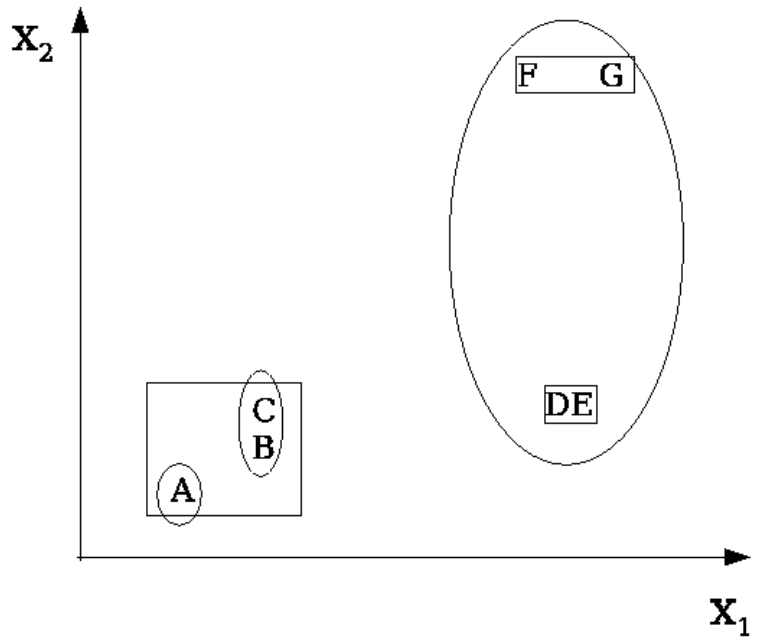


FIGURA 2.14 – O algoritmo k-Means é sensível às condições iniciais

Por isto a necessidade de várias inicializações randômicas com número de partições iniciais diferentes. Os passos para rodar o algoritmo k-Means são os seguintes:

1. Escolha randomicamente k padrões ou k pontos dentro do hiperespaço contendo os dados como sendo os centros inicias de *clusters*.

2. Adicione cada padrão ao centro de *cluster* que estiver mais próximo.

3. Compute o novo centro do *cluster* com o novo conjunto de membros.

4. Se o critério de convergência não é atingido, volte para o passo 2. Critérios típicos de convergência são nenhum ou mínimo numero de padrões trocando de clusters, ou baixa taxa de redução no erro quadrado.

O algoritmo ISODATA [BH65] permite unir e dividir os *clusters* resultantes do particionamento. No exemplo da Figura 2.14 o algoritmo ISODATA, uma vez apresentado à *clusterização* representada pelas elipses, atingiria o mínimo global representado pelos retângulos. Inicialmente ele uniria os *clusters* $\{A\}$ e $\{B, C\}$ dada a proximidade entre seus centróides. O *cluster* $\{D, E, F, G\}$ apresenta alta variância, e seria dividido resultando nos dois clusters $\{D, E\}$ e $\{F, G\}$.

2.7 Redes Neurais Artificiais

Uma RNA é um processador distribuído, maciçamente paralelo, constituído de unidades de processamento simples, denominados neurônios artificiais, que tem propensão natural para armazenar conhecimento e baseia-se no cérebro humano em dois aspectos [Hay01]: (1) o conhecimento é adquirido pela rede a partir do seu ambiente através de um processo de aprendizagem, e (2) o conhecimento adquirido é armazenado nos pesos sinápticos. Os pesos sinápticos são valores associados às sinapses ou conexões entre os neurônios da rede.

O grande sucesso das RNAs se deve à sua capacidade de aprendizagem a partir dos dados, sem auxílio de qualquer tipo de conhecimento prévio sobre o sistema de onde se originam. O treinamento é feito com algoritmos de treinamento que utilizam conjuntos de dados de entrada e saída para modificar os pesos da RNA. Os pesos são modificados de modo a gradualmente fazer com que a RNA apresente na sua saída valores cada vez mais próximos aos dados de saída reais para um conjunto de dados de entrada do sistema. Em outras palavras as RNAs podem ser treinadas através de algoritmos de aprendizagem para modelar o sistema. No entanto, a característica mais atraente das RNAs é a capacidade de generalização, ou seja, a capacidade de prever saídas para dados de entrada não existentes na fase de treinamento. Obviamente há restrições à efetividade das previsões das RNAs ligadas à quantidade de informação existente nos dados de treinamento.

Nas seções seguintes as características das RNAs e da aprendizagem serão abordadas com mais detalhes, desde o modelo de Neurônio Artificial, passando por RNAs multicamadas, aprendizado, algoritmos de treinamento, treinamento supervisionado, até alguns conceitos básicos sobre a RNA treinada para previsão de estrutura secundária: como representar os resíduos, e a técnica de janelamento.

2.7.1 Neurônio Artificial

O neurônio artificial é uma simplificação do neurônio biológico, ou ainda uma unidade de processamento baseada no funcionamento do neurônio biológico. Como no neurônio biológico, o neurônio artificial (doravante denominado simplesmente neurônio):

1. tem conexões de entrada, por onde recebe sinais de ativação de outros neurônios,

2. tem conexões de saída, por onde envia o seu sinal de ativação adiante,

3. tem uma mecanismo para conjugar as ativações de entrada e avaliar a sua intensidade para ativar a sua saída.

As conexões entre os neurônios biológicos são chamadas sinapses e são responsáveis pelo armazenamento de informação. Para tanto os neurônios que são ativados com maior freqüência tendem a aumentar de espessura em relação aos demais. As sinapses dos neurônios biológicos são representadas por conexões ponderadas entre os neurônios, onde o peso w_{yx} representa a espessura da sinapse entre os neurônios biológicos x e y (Figura 2.15, no alto a esquerda). A avaliação dos sinais de entrada para ativação do neurônio é feita aplicando-se o somatório das entradas a uma função sigmoidal. Este tipo de função garante que a saída de um neurônio se estabilize em um valor mínimo ou máximo a medida em que o somatório das entradas tenda a se afastar da origem. Esta capacidade de estabilização permite que o conhecimento adquirido seja estável durante a aprendizagem.

O neurônio de uma RNA é portanto uma unidade de processamento composta por um somador e uma função de ativação não linear (Figura 2.15). Ligado ao neurônio estão as conexões de entrada vindas de outros neurônios ou dos dados de entrada, e a conexão de saída que é ativada por uma função de ativação. Somando-se às conexões de entrada há uma conexão que não provém de outro neurônio ou de dados de entrada, mas da unidade de *bias*. Esta unidade tem a função de adicionar uma constante à função de ativação, permitindo assim que a curva da função se afaste da origem. O neurônio recebe sinais de entrada de outros neurônios e do*bias*, ponderados por pesos sinápticos, e responde na sua saída com o sinal de ativação. O modelo de um neurônio artificial pode ser descrito pelo par de equações:

$$a_k = \sum_{j=1}^{m} w_{kj} x_j \tag{2.42}$$

e

$$y_k = \varphi(a_k + b_k), \tag{2.43}$$

onde $x_1, x_2, \ldots, x_m$ são os sinais de entrada; $w_{k_1}, w_{k_2}, \ldots, wk_m$ são os pesos sinápticos do neurônio k; b_k é o *bias*; $\varphi()$ é a função de ativação e y_k é o sinal de saída ou ativação do neurônio. O *bias* efetua uma transformação afim no combinador linear a_k

2.7.2 RNA multicamada

O *perceptron* é a denominação de uma rede neural muito simples, estudada intensamente nas décadas de 50 e 60 devido à sua capacidade de aprendizagem. O perceptron consiste basicamente de uma camada de unidades de entrada conectadas a um neurônio de saída. As unidades de entrada são neurônios simples, cuja função de ativação consiste em simplesmente repassar o valor da entrada para o neurônio de saída. Em 1969 porém, M. Minsky e S. Papert provaram matematicamente que o perceptron somente conseguia aprender a mapear funções lineares [Hay01]. Com sua

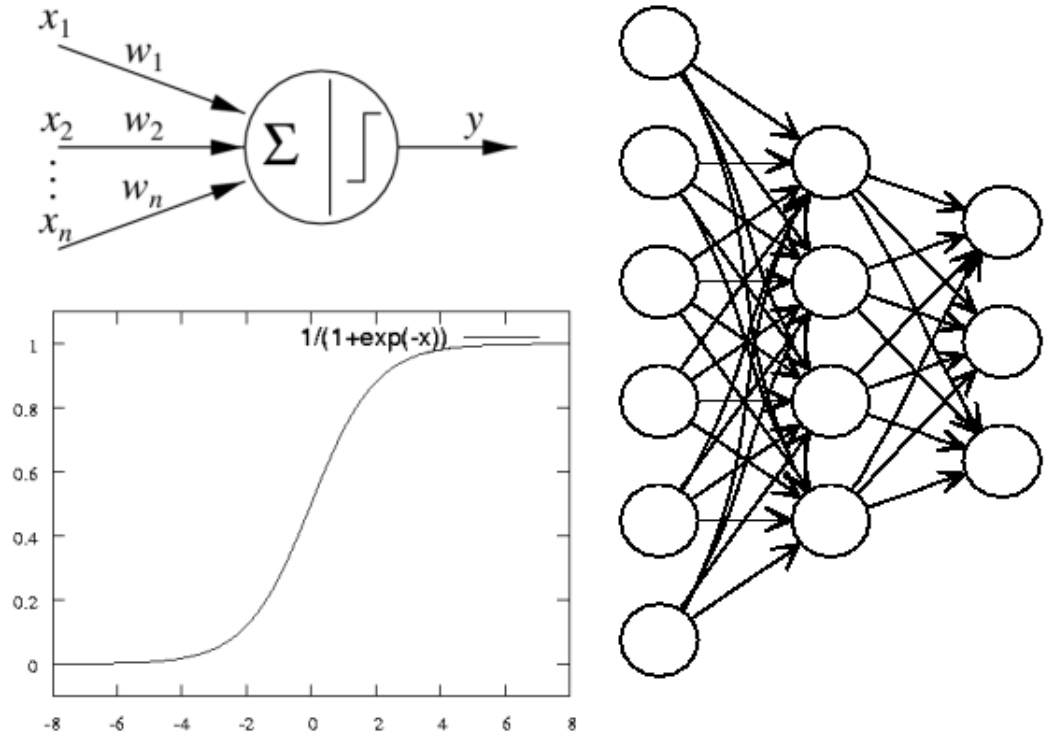


FIGURA 2.15 – Acima à esquerda: representação esquemática de um neurônio artificial. Os valores de entrada $x_1, x_2, \ldots x_n$ são multiplicados pelos respectivos pesos $w_1, w_2, \ldots, w_n$. O somatório das entradas ponderadas pelos pesos aplicado à função de ativação é o valor de ativação y do neurônio. Em baixo na esquerda a função de ativação Sigmóide $y = \frac{1}{1+(e^{-x})}$. Na direita, representação de uma RNA com 6 neurônios na camada de entrada, 4 neurônios na camada escondida e 3 neurônios na camada de saída.

famosa demonstração da incapacidade do *perceptron* de aprender a simples função *XOR*, mostraram que o *perceptron* só era capaz de classificar dados linearmente separáveis, e provocaram um desinteresse da comunidade científica por RNAs que perdurou até meados da década de 80. Foi nesta época que o interesse pelas RNAs ressurgiu, com o advento das RNAs multicamadas.

A RNA multicamada *feed-forward* (Figura 2.15, à direita) é uma RNA com múltiplas camadas de neurônios, sendo que cada neurônio de uma camada está conectado com os neurônios da camada seguinte. A RNA multicamada*feed-forward* é composta de uma camada de entrada, uma ou mais camadas escondidas com neurônios de função de ativação não-linear e uma camada de saída. Devido à camada de neurônios escondida este tipo de RNA é capaz de aprender mapeamentos não lineares entre os dados de entrada e os dados de saída. O nome *feed-forward* ou alimentação à frente se refere à característica das RNAs multicamadas em que as saídas dos neurônios das camadas intermediárias ou *escondidas* é direcionada para

a entrada dos neurônios das camadas seguintes.

Dado que se tenha um conjunto de dados de entrada e de saída representativo do domínio e imagem de uma função não linear e com tamanho suficiente, haverá uma configuração de RNA multicamada que associada a um algoritmo de aprendizado será capaz de aprender esta função em um tempo finito e generalizar os resultados para dados externos a este conjunto com um erro determinado. O modelo mais comumente utilizado de RNA multicamada consiste em RNAs de três camadas: uma camada de entrada, uma camada escondida e uma camada de saída. Descartados os problemas de qualidade dos dados de aprendizado, a limitação de aprendizagem de funções não lineares para este tipo de RNA está ligado principalmente ao número de neurônios pertencentes à camada escondida. Se o número de neurônios for muito pequeno corre-se o risco de a RNA não ser capaz de mapear a função não linear com um erro pequeno. Já se o número de neurônios for demasiado grande em relação à quantidade de dados de treinamento, a RNA pode se especializar em demasia nestes dados e perder a capacidade de generalização, no que se costuma chamar de *decorar*.

Infelizmente não se conhece ainda um método de escolha automático do número de neurônios ideal para cada tipo de conjunto de dados de treinamento para funções contínuas. O algoritmo *Cascade Correlation* proposto por Scott Fahlman permite a construção automática da topologia da rede durante o treinamento, mas é aplicável somente para problemas de classificação. Como a grande maioria dos sistemas na natureza são representados por funções contínuas não lineares, a escolha da topologia ainda depende da experiência do pesquisador e da tentativa e erro.

2.7.3 Aprendizado

O aprendizado de uma RNA é o resultado da aplicação interativa de um algoritmo de treinamento baseado em um conjunto de dados de entrada e saída de um sistema. Em linhas gerais, o método de treinamento consiste em repetir as seguintes operações:

1. atribuir valores iniciais aos pesos da RNA,
2. alimentá-la com dados de entrada,
3. fazer o cálculo da ativação de todos os neurônios na ordem de entrada para a saída (alimentação à frente),
4. comparar os valores de ativação dos neurônios da camada de saída com os valores de saída do conjunto de dados do sistema,
5. e por fim alterar os valores dos pesos para diminuir o erro obtido.

As operações enumeradas acima são repetidas até que seja atingido o critério de parada, tipicamente o erro médio quadrado. Virtualmente todos os algoritmos de treinamento supervisionado, ou seja, em que os pesos são modificados em função do erro da RNA em relação aos dados esperados, se baseiam nas regras e no algoritmo descritos acima. A atribuição inicial é tipicamente mas não obrigatoriamente aleatória. O erro médio quadrado é calculado sobre o erro para cada padrão, que é uma instância de valores de entrada e saída, e é calculado sobre dados não disponíveis

no treinamento. O controle do erro em dados externos ao conjunto de treinamentos permite verificar quão bem a RNA aprendeu a generalizar o aprendizado.

A escolha dos dados que serão separados do conjunto de testes e a maneira que será feita esta separação é essencial para evitar comportamentos tendenciosos ou o super-treinamento da RNA. Por aprendizado entende-se fazer a RNA aprender a simular o comportamento do sistema. Se entregamos todos os dados disponíveis para o treinamento da RNA, mesmo que obtenhamos índices de erro muito pequenos só poderemos afirmar que a RNA aprendeu a mapear o conjunto de dados de entrada disponível para o conjunto de dados de saída igualmente disponível. O aprendizado efetivo do comportamento de um sistema só é atingido se separarmos parte dos dados do conjunto de treinamento e criarmos um conjunto de teste.

O conjunto de dados de teste é utilizado então para acompanharmos o treinamento da RNA e a evolução do erro para os dados de teste, que ela não conhece. Isto se faz interrompendo o treinamento periodicamente, apresentando os dados de entrada do conjunto de teste para a RNA, e comparando a sua saída com os dados de saída do mesmo conjunto. Enquanto o erro para estes dados que a RNA não conhece estiverem diminuindo significa que a capacidade de a RNA generalizar o aprendizado para dados desconhecidos está aumentando, e devemos continuar o treinamento. Quando este erro parar de descer devemos interromper o treinamento independentemente de o erro médio quadrado para os dados de treinamento continuar a descer. Na verdade, a medida que continuamos com o treinamento, é comum o erro de teste parar de descer e começar a subir enquanto o erro dos dados de treinamento continua a descer. Isto chama-se super-treinamento e reflete a perda de capacidade de generalização da RNA em prol da especialização nos dados de treinamento. Em uma comparação livre com o aprendizado humano, pode-se dizer que a rede está deixando de aprender e está começando a *decorar*. Um terceiro conjunto de dados ainda pode ser separado para validação do aprendizado. O conjunto de dados de validação não participa nem do treinamento, nem do teste, mas é apresentado à rede já treinada para validar o erro de generalização obtido nos testes.

A maneira como se separam os dados em conjuntos de treino, teste e validação também influi na aprendizagem, e para garantir a lisura do método deve ser feita de maneira adequada. Não há um consenso sobre as técnicas mais adequadas, mas as mais aceitas são *holdout* e *k-fold-cross-validation*. O método *holdout* consiste em separar o conjunto de dados em dois conjuntos: o conjunto de treino e o de teste, cada um com 2/3 e 1/3 do total de dados respectivamente. O método *k-fold-cross-validation* por sua vez divide o conjunto de dados em K subconjuntos de mesmo tamanho. Então são formados dois conjuntos, o conjunto de treino com $K-1$ subconjuntos, e o conjunto de teste com o subconjunto restante. O processo de aprendizado é feito K vezes, alternando-se a cada treinamento o subconjunto que forma o conjunto de teste. Usualmente se utiliza o método *10-fold-cross-validation*, mas K pode chegar até o número de padrões existente no conjunto de dados. Neste caso o método passa a se chamar *live-one-out*, pois o treinamento é feito N vezes para um conjunto de dados de tamanho N, cada vez retirando apenas 1 dos dados para o conjunto de teste. Este último método é o ideal pois controla o erro para cada um dos dados de teste individualmente, e permite identificar a eficácia do aprendizado por regiões de dados, mas é computacionalmente pesado e é em geral preterido em prol do bem aceito *10-fold-cross-validation*.

Há ainda a escolha dos dados que irão para cada conjunto. Esta escolha tem de ser feita com critério de modo a evitar aprendizados tendenciosos. Os dois métodos principais para a escolha dos dados são a escolha aleatória e a estratificação. O primeiro consiste na escolha randômica dos dados para cada subconjunto, e a segunda se baseia na manutenção da distribuição estatística dos dados em cada subconjunto. Ambas dependem da qualidade do gerador pseudo-randômico utilizado, e a diferença entre ambas diminui com o aumento do conjunto de dados. Ou seja, o cuidado da manutenção da representação de todos os tipos de classes em cada subconjunto só faz sentido se a quantidade de dados for relativamente pequena.

Por fim há a necessidade de se repetir cada conjunto de treinamentos com sementes aleatórias diferentes, pois as RNAs são sujeitas à mínimos locais como qualquer método de minimização. Como os algoritmos de treinamento são determinísticos, dois treinamentos com a mesma inicialização e os mesmos conjuntos de treino e teste vão chegar exatamente ao mesmo resultado. E inicializações com sementes aleatórias diferentes tendem a resultar em erros mínimos diferentes.

2.7.4 Algoritmos de Treinamento

Os algoritmos de treinamento supervisionado são todos variações de um mesmo método: um conjunto de exemplos de dados de entrada e saída é apresentado à RNA, a diferença entre a resposta da RNA e aos dados de saída é calculado, e correções são aplicadas aos pesos baseados neste erro.

O método mais comum de modificar os pesos de uma RNA é baseado na regra de Hebb, segundo a qual uma sinapse entre dois neurônios é fortalecida se as duas unidades são ativadas ao mesmo tempo. A forma geral da regra de Hebb é:

$$\Delta w_{ji} = g(a_j(t), d_j)h(o_i(t), w_{ji}) \tag{2.44}$$

onde:

w_{ji}	peso da sinapse do neurônio i para o neurônio j
Δw_{ji}	variação do peso w_{ji} durante o processo de aprendizagem
$a_j(t)$	ativação do neurônio j no passo t
d_j	valor de exemplo, em geral a saída desejada do neurônio j
$o_i(t)$	saída do neurônio i no passo t
$g(\dots)$	função, depende da ativação do neurônio e da saída desejada
$h(\dots)$	função, depende da saída do neurônio precedente e do peso atual da sinapse

2.7.5 Treinamento Supervisionado

Treinar uma RNA *feed-forward* com aprendizado supervisionado consiste nos seguinte passos.

1. Um padrão de entrada é apresentado à camada de entrada da RNA. A entrada então é propagada para a frente na rede até que a ativação atinja a camada de saída.

2. A saída da RNA é então comparada com o valor desejado, o erro é propagado recursivamente para trás na RNA e os pesos de todas as sinapses são ajustados de acordo com um delta calculado para cada sinapse.

Os passos acima são repetidos para todos os padrões da base de dados de treinamento, quantas vezes forem necessárias para estabilizar o erro da RNA.

Dentre os algoritmos mais comuns para o treinamento de RNAs *feed-forward* encontram-se os algoritmos *BackPropagation* e suas variantes, o *Rprop* e o *QuickProp*.

O algoritmo mais famoso para treinamento de RNAs *feed-forward* é o *backpropagation*. O nome *backpropagation* se refere à retropropagação do erro para cada camada da RNA. A maior dificuldade da utilização de RNAs multicamadas nas décadas de 50 e 60 era a inexistência de uma regra que permitisse escolher corretamente nas camadas intermediárias os neurônios a serem penalizados. Até que Rumelhart, Hinton e Williams desenvolvessem em 1986 o algoritmo BackPropagation, somente era possível treinar RNAs sem camadas intermediárias, pois a regra de Hebb (eq. 2.44) dependia do erro da saída da rede. O problema de determinar um *erro* para a saída de neurônios de camadas intermediárias de forma a poder alterar os pesos que os ligavam à camadas anteriores foi resolvido pela chamada *generalized delata-rule*. O algoritmo *BackPropagation* consiste então nos seguintes passos:

1. *Fase de propagação para a frente.*

   Um padrão de entrada é apresentado à RNA. A entrada é então propagada à frente através da RNA até atingir a camada de saída.

2. *Fase de propagação para trás - Camada de saída*

   A ativação (saída) da camada de saída é então comparada com a saída desejada da RNA. O erro, ou seja, a diferença (delta) δ_j entre a ativação o_j e o valor da saída desejada d_j de um neurônio de saída j é então utilizado em conjunto com o valor de ativação o_i do neurônio i da camada anterior para computar a alteração necessária do peso w_{ji} que conecta os dois.

   Para computar os deltas dos neurônios das camadas internas (escondidas) para os quais não há valores de saída desejados com os quais comparar, os deltas da camada seguinte, que já foi computado, são utilizados na fórmula da *generalized delta-rule* (eq. 2.46). Desta maneira os erros (deltas) são propagados para trás e todos os pesos da RNA são corrigidos de acordo.

O tipo de atualização de pesos pode ser *online* ou *offline*. No treinamento *online* as atualizações de pesos $Delta_{w_{ji}}$ são feitas para cada padrão apresentado, ou seja, a cada passo de propagação para frente e para trás. Já no treinamento *offline*, os deltas de cada padrão são acumulados até que o ciclo (época) esteja completo. Só então o $Delta_{w_{ji}}$ acumulado é aplicado ao peso. A *generalized delta-rule* utilizada pelo algoritmo de *BackPropagation* é dada por:

$$\begin{aligned} \Delta w_{ji} &= \eta\, \delta_j\, o_i \qquad (2.45) \\ \delta_j &= \begin{cases} \varphi'_j(a_j + b_j)(d_j - o_j) & \text{se o neurônio } j \text{ é um neurônio de saída} \\ \varphi'_j(a_j + b_j) \sum_k \delta_k w_{kj} & \text{se o neurônio } j \text{ é um neurônio} \\ & \text{da camada escondida} \end{cases} \end{aligned}$$

onde:

η = taxa de aprendizado (constante)
δ_j = erro
d_j = saída desejada do neurônio j
o_i = saída do neurônio precedente i
i = índice de um neurônio predecessor do neurônio j corrente com peso w_{ji} de i para j
j = índice do neurônio corrente
k = índice de um sucessor do neurônio corrente j com peso w_{kj} de j para k

A *generalized delta-rule* é utilizada pelo algoritmo *BackPropagation*. Os outros dois algoritmos citados, *RProp* e *QuickProp* utilizam regras de atualização de pesos diversas.

2.7.6 RNA aplicada à previsão de estrutura secundária

Para complementar os conceitos básicos de RNA, é necessário ressaltar algumas características específicas dos métodos de treinamento de RNAs para determinação de estrutura secundária de proteínas. Nesta seção analisaremos os seguintes pontos: a representação dos resíduos, o janelamento e o resultado produzido pela RNA.

A Representação dos resíduos

Uma RNA treinada para reconhecer a estrutura secundária de proteínas recebe com dados de entrada tipos de resíduos, que são variáveis discretas. Para ser processada esta informação deve ser transformada em valores numéricos.

A maneira de se efetuar esta representação é através de uma codificação. A codificação mais usual neste tipo é a chamada representação ortogonal. Nesta representação a entrada discreta é representada por N entradas binárias. onde N é o número de classes as quais a variável discreta pode pertencer. Como as proteínas são compostas por 20 tipos principais de resíduos, uma RNA que recebe como entrada o tipo de um resíduo tem uma camada de entrada composta por 20 neurônios. Para cada tipo de resíduo um dos neurônios recebe o valor 1 como entrada, e os demais 19 neurônios recebem o valor zero. Por exemplo, para representar os resíduos Alanina, Arginina e Valina poderíamos utilizar os códigos 00000000000000000001, 00000000000000000010 e 10000000000000000000 respectivamente. Com é comum as RNAs podem receber como entrada não um mas vários resíduos no formato de uma *janela* da seqüência de resíduos de tamanho N. Neste caso o número de neurônios da camada de entrada é necessariamente de $20N$.

Além desta representação binária do tipo de resíduos, podem ser utilizados como entrada de RNAs os atributos físico-químicos dos resíduos. Para os atributos que têm escala contínua podem ser utilizados os valores normalizados, e para os demais a codificação ortogonal descrita acima. De acordo com [WM00], os atributos físico-químicos mais utilizados são hidrofobicidade, volume, massa, área, propensão a formar determinada estrutura secundária, refratividade. Baldi e Brunak [BB01] ainda adicionam à lista: carga, família, distância da extremidade da proteína, entre outros.

Janelamento

A princípio uma RNA que fosse ser treinada para mapear a estrutura primária de uma proteína em sua estrutura secundária deve basicamente aprender a prever a estrutura secundária desta proteína baseada apenas na seqüência da resíduos que a compõe. Portanto esta RNA poderia receber como entrada por exemplo o tipo de resíduo e a posição relativa dele na seqüência e apresentar como saída a estrutura secundária à qual este resíduo pertence. Esta exemplo de arquitetura de RNA não tem a menor possibilidade de aprender o mapeamento desejado por um motivo muito simples: não há informação suficiente no padrão de entrada composto apenas pelo tipo de resíduo e a sua posição relativa. Hoje se sabe que o tipo e a posição dos resíduos anteriores e posteriores ao resíduo em análise influenciam na determinação do tipo de estrutura secundária a qual este resíduo fará parte, findo o dobramento da proteína na natureza [HMK95].

A técnica de janelamento visa entregar para RNA informações não somente sobre o resíduo do qual se quer descobrir a estrutura secundária, mas de parte da seqüência de resíduos que o circundam. Então, cada padrão de dados constitui não mais o tipo e posição do resíduo como entrada e a estrutura secundária como saída, mas um segmento ou janela da seqüência de resíduos. Esta janela é composta por N resíduos, e a saída da RNA é a previsão da estrutura secundária à qual pertence o resíduo central desta janela.

2.7.7 Métodos PHD/PROF

O método PHDsec de Burkhard Rost [RS93, Ros96] se utiliza de RNAs para predição de estrutura secundária de proteínas a partir da estrutura primária. Mais precisamente, as RNAs aprendem a predizer a estrutura secundária a qual pertence um resíduo baseado em informações locais, da vizinhança do resíduo, e em informações globais da seqüência de resíduos. As informações locais não são extraídas diretamente da seqüência de resíduos, mas do resultado do alinhamento da seqüência com proteínas homólogas. Portanto o método têm em duas etapas: na primeira é gerado um alinhamento com múltiplas proteínas homólogas, e na segunda este alinhamento serve de entrada a um sistema de RNAs.

O alinhamento de proteínas é realizado inicialmente com BLAST [AGM+90], que é um método rápido para alinhamento de proteínas. O programa compara a seqüência de cada proteína com o banco de dados de seqüências SWISSPROT e calcula a significância das similaridades entre seqüências. Finalizado o trabalho do BLAST, o método PHDsec utiliza-se do programa MaxHom [SS91], um programa de alinhamento dinâmico de múltiplas seqüencias mais sensível, baseado em perfis de alinhamento. O programa MaxHom constrói uma família[3] de proteínas em duas etapas: (1) refaz o alinhamento utilizando agora somente as proteínas tidas com homólogas pelo método BLAST, e (2) realiza um corte baseado no número de resíduos alinhados.

Os dados locais do perfil resultante da etapa de alinhamento, juntamente com dados estatísticos globais da seqüência original, servem agora de entrada para as RNAs. O método PHD para predição de estrutura secundária de proteínas processa os dados de entrada em múltiplos níveis. O primeiro nível é uma RNA com alimentação à frente de três camadas (de entrada, escondida, e de saída), que faz o

[3]Uma família é definida por um conjunto de proteínas que tendem a ter estruturas similares.

mapeamento entre dados da seqüência e a correspondente estrutura secundária. Os dados de entrada da RNA de primeiro nível provém de dois tipos de contribuições: uma é local e corresponde a dados do alinhamento retirados de um janela de treze resíduos, e a outra consiste em dados da seqüência global (Figura 2.16).

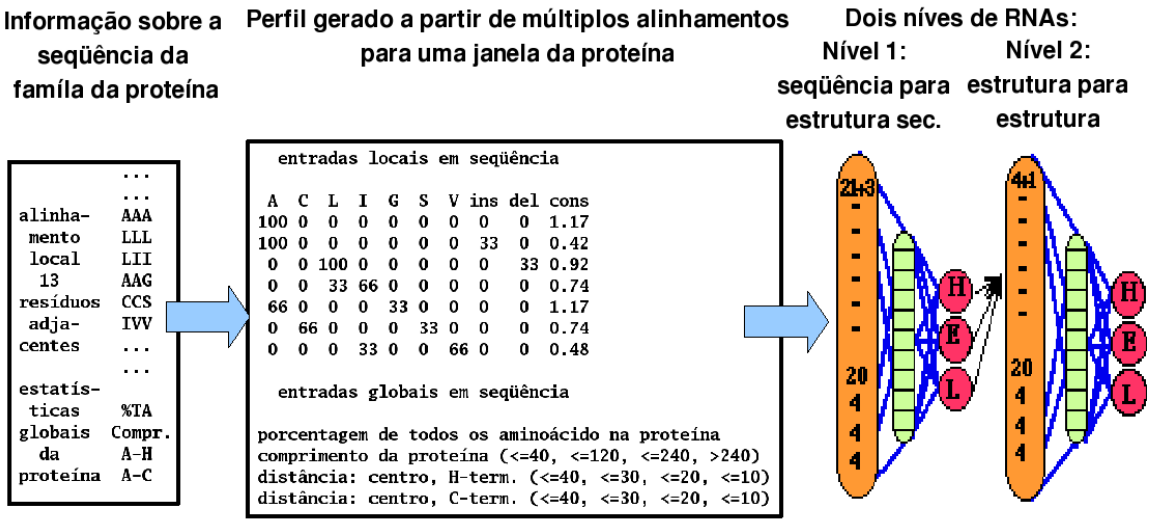


FIGURA 2.16 – Método PHDsec (Figura extraída de [RS93, Ros96]). Primeiro uma janela de 13 resíduos é selecionada do alinhamento da seqüência (Na Figura é mostrada uma janela de apenas 7). Em seguida são computados o perfil e informações globais a partir da seqüência da proteína. Finalmente o sistema de RNAs é alimentado com as informações locais e globais. O sistema de RNAs é composto por RNAs em dois níveis. A RNA do primeiro nível tem 24 neurônios para informação local (20 para os tipos de resíduos, um para um *espaçador* que permite estender a janela além das extremidades da proteína, dois para a quantidade de inserções e deleções, e um para o peso de conservação); e 32 para informação global (20 para a composição de aminoácidos da proteína, 4 para o comprimento da proteína, e 8 para a distância da janela em relação às extremidades da proteína). A camada de saída tem 3 unidades que representam a estrutura secundária do resíduo central da janela. A RNA do segundo nível recebe com entrada a saída do primeiro nível mais as informações globais (espaçador, constante, etc). A saída da RNA de segundo nível á mesma da de primeiro nível: 3 neurônios, uma para α-hélice, outro para segmento de folha-β e o terceiro para o resto.

A saída da RNA do primeiro nível é a previsão da estrutura secundária na qual o resíduo central da janela se encontra. A RNA do segundo nível realiza um mapeamento do tipo estrutura secundária para estrutura secundária, ou seja, recebe uma janela da seqüência de estruturas secundárias gerada pelo sua da de previsão da RNA do primeiro nível e, baseada nesta janela e novamente em dados globais da seqüência, realiza nova previsão de estrutura secundária para o resíduo central da janela. O nível seguinte consiste na média aritmética sobre o resultado de RNAs treinadas independentemente, e o último nível é simplesmente um filtro que corrige previsões drasticamente irreais (ex.: $HEH \rightarrow HHH$; $EHE \rightarrow EEE$; *e* $LHL \rightarrow LLL$) Para

fins de previsão de estrutura secundária, a estrutura prevista é a correspondente ao neurônio com o maior valor, e o grau de confiança na previsão é dado pela diferença entre este valor e o valor dos dois neurônios de saída restantes.

As camadas de entrada das RNAs de primeiro e segundo nível que compõem o sistema de RNAs do método PHDsec estrutura secundária contém respectivamente 344 e 84 neurônios. Para a RNA do primeiro nível, os neurônios da camada de entrada estão divididos da seguinte forma:

- Informação local: para cada um dos 13 resíduos da janela são necessários 24 neurônios. Um grupo de 24 neurônios, representando o resíduo de uma posição específica, é divido da seguinte maneira:
  - Tipos de resíduos: 20 neurônios, um para cada um dos resíduos padrão.
  - Espaçador: 1 neurônio, indica que a janela deve ser estendida para a outra extremidade da proteína. Isto permite que resíduos próximos às extremidades da seqüência possam aparecer no centro da janela de resíduos.
  - Ins/Del: 2 neurônios, contém a quantidade de inserções e deleções no alinhamento para esta posição.
  - Peso de conservação: 1 neurônio para a constante de conservação calculada durante a fase de alinhamento
- Informação global: 32 neurônios são necessários para codificar os dados de entrada sobre a seqüencia de resíduos.
  - Porcentagem de aminoácidos: 20 neurônios recebem o valor da porcentagem de cada tipo de aminoácido presente na seqüência da proteína.
  - Tamanho da proteína: 4 neurônios codificam o tamanho da proteína em valores discretos: ≤ 60, ≤ 120, ≤ 240 ou > 240.
  - Distâncias A-Hterm e A-Cterm: 8 neurônios codificam a distância da janela às extremidades da proteína (≤ 40, ≤ 30, ≤ 20 ou ≤ 10).

A RNA do segundo nível tem a camada de entrada dividida de maneira semelhante à RNA do primeiro nível, trocando apenas o grupo *Tipo de resíduo* pelo grupo *Tipo de estrutura secundária* e sem o grupo de inserção e deleção. A estrutura da RNA de segundo nível é então:

- Informação local: para cada um dos 13 resíduos da janela são necessários 5 neurônios. Um grupo de 5 neurônios, representando o resíduo de uma posição específica, é divido da seguinte maneira:
  - Tipos de estrutura secundária: 3 neurônios, representando α-hélice, segmento de folha-β, e nenhuma das anteriores (*loop*.
  - Espaçador: 1 neurônio, indica que a janela deve ser estendida para a outra extremidade da proteína. Isto permite que resíduos próximos às extremidades da seqüência possam aparecer no centro da janela de resíduos.

  - Peso de conservação: 1 neurônio para a constante de conservação calculada durante a fase de alinhamento

- Informação global: 32 neurônios são necessários para codificar os dados de entrada sobre a seqüencia de resíduos.
  - Porcentagem de aminoácidos: 20 neurônios recebem o valor da porcentagem de cada tipo de aminoácido presente na seqüência da proteína.
  - Tamanho da proteína: 4 neurônios codificam o tamanho da proteína em valores discretos: ≤ 60, ≤ 120, ≤ 240 ou > 240.
  - Distâncias A-Hterm e A-Cterm: 8 neurônios codificam a distância da janela às extremidades da proteína (≤ 40, ≤ 30, ≤ 20 ou ≤ 10).

A estrutura secundária é codificada por três unidades: α-hélice, H (H, G e I do DSSP [KS83a]), segmento de folha-β, E (E e B do DSSP), e nenhuma das anteriores, denotado por L de *loop*. Esta codificação é idêntica nas RNAs dos níveis1 e 2, e os dados de entrada mais significativos para a RNA do segundo nível são as janelas da seqüência de estruturas secundárias geradas pela RNA do nível anterior. O uso da RNA de mapeamento de estrutura secundária pra estrutura secundária se deve à dificuldades da RNA do primeiro nível de aprender certas características específicas sobre formação de estruturas secundárias. Dificuldade esta inerente ao método de treinamento de RNAs, que pressupõe independência entre dados adjacentes e apresenta à RNA exemplos escolhidos de forma randômica. Como resultado a RNA do primeiro nível é capaz de aprender a prever a estrutura secundária provável a que pertence um resíduo em determinada posição da seqüencia, mas é incapaz de aprender por exemplo que α-hélices contém no mínimo três resíduos. A RNA do segundo nível aprende a determinar a estrutura secundária utilizando informação de contexto sobre a *estrutura secundária* adjacente.

Os dados primários para o treinamento das RNAs, ou seja, as seqüências de proteínas propriamente ditas, apresentam proporções diferentes para as três estruturas secundárias. Os resíduos do banco de dados utilizado por Rost estavam distribuídos aproximadamente entre as estruturas da seguinte forma: 32% em α-hélices, 21% em segmentos de folhas-β, e 47% em *loops*. Para evitar que esta distribuição acarretasse em piores resultados para as classes menos representadas (menos dados == menor acurácia), foi utilizado o que se chama de treinamento balanceado: o treinamento foi realizado alternando-se as três estruturas. Por exemplo: se no passo anterior do treinamento o exemplo apresentado à RNA consistia em uma janela com o resíduo central em uma α-hélice, o exemplo atual é uma janela com o resíduo central em uma folha-β, e a próxima janela será escolhida entre as que tem resíduo central em *loop*. 0 treinamento balanceado representa melhora no predição de estruturas menos representadas no dados (ex.: folha-β) mas piora os resultados para as estruturas mais presentes (ex: *loop*), e por conseguinte o desempenho geral da capacidade de predição cai.

Para encontrar o meio termo entre RNAs treinadas com e sem treinamento balanceado, o método PHDsec implementa a estratégia de decisão por júri. Considerando o sistema de RNAs em dois níveis utilizado, foram realizados4 tipos distintos de treinamento cobrindo todas as 4 combinações possíveis de RNAs de primeiro nível com treinamento não balanceado e balanceado, e RNAs de segundo nível com treinamento não balanceado e balanceado. A decisão por júri é a média aritmética

simples sobre o resultado dos 4 sistemas de RNAs, e o resultado final é dado pela unidade de maior valor entre as três unidades de saída. Além da unidade com maior valor (unidade vencedora) determinar a estrutura predita, a diferença entre esta unidade e a de segundo maior valor é usada para calcular o índice de confiabilidade na previsão. Este índice é normalizado entre 0 e 0.9, e é tanto maior quanto maior for a diferença da unidade vencedora para as demais.

Para completar a explanação sobre a metodologia, é importante falarmos sobre como os dados foram escolhidos e divididos para evitar a super-especialização. RNAs com muitos graus de liberdade, treinadas sem dados de validação, ou com dados não representativos da população, tendem a se especializar nos dados de treinamento enquanto perdem a capacidade de generalização. Como a base da utilização de RNAs é a capacidade de generalizar o conhecimento adquirido com os dados de treinamento para dados não conhecidos, deve-se evitar a todo o custo a especialização. Para o treinamento das RNAs do método PHDsec foram formados conjuntos de treino e teste de tal forma que: (1) os grupos fossem diferentes, ou seja, nenhuma proteína de um grupo com mais de 25% de similaridade com proteínas do outro grupo; (2) o treinamento (ajuste de parâmetros livres) fosse feito com *cross-validation*; e (3) a validação fosse feita com novas estruturas de proteínas experimentalmente determinadas após o começo do projeto. Se as RNAs aplicadas aos novos grupos de validação obtivessem menor acurácia do que quando aplicadas aos grupos de teste, isto significa que houve super especialização no treinamento.

O PHDsec (PHD para predição de estruturas secundárias de proteínas) foi o primeiro método a ultrapassar a marca de 70% de acurácia. De fato, à época da publicação, o método atingiu 72% de acerto na predição de estrutura secundária por resíduo. Para atingir esta meta foram utilizadas no treinamento mais de 300 cadeias de proteínas, somando um total de 70.000 resíduos. Originalmente disponível no servidor *Predict Protein*, o método PHDsec foi substituído pelo mais recente PROFsec, também de Burkhard Rost e ainda não publicado. O método PROFsec é similar ao PHDsec, mas utiliza-se do método PSI-BLAST [AMS+97], mais rápido e sensível a similaridades fracas porém importantes do ponto de vista biológico, para realizar o alinhamento das proteínas na primeira fase do processo. Todas as predições utilizadas neste trabalho foram retiradas das RNAs treinadas pelo método PROFsec, através de consultas ao servidor *Predict Protein*.

Capítulo 3

Estado da Arte

Neste Capítulo são abordados temas relevantes a este trabalho, com a intenção de situar a proposta no contexto dos trabalhos científicos relacionados ao trabalho. As áreas abordadas incluem RNAs e suas aplicações na simulação de sistemas dinâmicos e classificação, e métodos de previsão de estrutura secundária de proteínas.

3.1 Predição da Estrutura tridimensional

A Hipótese Termodinâmica [Anf93] estabelece que a estrutura tridimensional de uma proteína em seu ambiente fisiológico natural é tal que a energia livre de todo o sistema é mínima. Ou seja, a conformação nativa de uma proteína é determinada pela totalidade das interações interatômicas e portanto pela seqüência de aminoácidos. Esta idéia enfatiza que a conformação estável tridimensional de uma proteína somente faz sentido em seu ambiente natural, ou seja, na presença de água, íons, nível de pH, temperatura, etc, similares ao ambiente fisiológico.

A maior parte das abordagens utilizadas para determinação da estrutura de proteínas pode ser classificada em dois grandes grupos: métodos *Ab initio* e por homologia. As modelagens *Ab initio* prescindem de conhecimento prévio sobre a estrutura tridimensional da proteína e se baseiam em interações físico-químicas, enquanto as abordagens baseadas em homologia se valem de bancos de dados para analisar a similaridade entre seqüências [DBS04].

De acordo com [HW02] a acurácia dos métodos de predição de estrutura secundária é limitada pela própria natureza flexível das proteínas, que permite diferentes estruturas secundárias para segmentos de resíduos homólogos. Em [VRD+01] vários métodos são utilizados para prever a desordem estrutural a partir da seqüência de aminoácidos, sendo que os melhores resultados são obtidos com conjuntos de RNAs.

3.2 Dinâmica Molecular

Balali-Mood e outros utilizaram dinâmica molecular em [BMHB03] para simular uma camada dupla mista composta por dioleoilfosfatidilcoline (DOPC) e dioleoilfosfatidilglicerol (DOPG) na água. Para tanto utilizaram unidades de água do tipo Carga em Ponto Simples e a estrutura inicial foi previamente construída e dobrada manualmente formando a camada dupla.

A simulação foi ainda realizada com hidrogênio explícito para verificar a interação entre peptídeos e as camadas duplas. O tempo combinado para equilíbrio com minimização de energia e dinâmica molecular para as camadas duplas com 140 moléculas foi de aproximadamente $3,5ns$, o equivalente a aproximadamente 30 dias em um PC Atlhon biprocessado.

Lehninger [LCN00] reporta a simulação de dinâmica molecular de um subdomínio de 36 resíduos da proteína vilina. Para a simulação no tempo teórico de $1\mu s$ foram necessários meio bilhão de passos de integração executados por dois supercomputadores Cray, cada um rodando por dois meses.

3.3 Métodos Estocásticos

Moret *et all* [MBMP02] propuseram em 2002 um método estocástico otimizado de busca no espaço de conformação de polipeptídeos. No trabalho eles utilizaram Descida de Gradiente para encontrar as conformações de baixa energia de polipeptídeos que tipicamente formam α-hélices, em um espaço de busca reduzido através da utilização das regiões permitidas dos Mapas de Ramachandran para cada peptídeo.

3.4 RNAs

N. Qian e T. Sejnowski [QS88] utilizaram em 1988 um modelo neural para predizer a estrutura secundária diretamente a partir da seqüência de aminoácidos. A rede neural treinada era capaz de receber como entrada uma janela de 13 resíduos de proteínas reais e devolver como resposta o tipo de estrutura secundária a que o resíduo central pertence com 63% de precisão.

Em [YY01] um conjunto de RNAs é utilizado para atingir precisão de66%.

O método PHDsec [RS93, Ros96] utiliza duas camadas de RNAs para ultrapassar a barreira dos 70% de precisão na predição da estrutura secundária em 1993.

Em [FA01] RNAs são utilizadas para prever a localização sub-celular de proteínas a partir da composição e ordem dos resíduos.

Em [Wu96] Cathy H. Wu apresenta uma introdução abrangente e detalhada às RNAs e a sua utilização para análise de seqüências moleculares.

Em [CLZ+01] RNAs são utilizadas para predizer segmentos transmembrânicos de α-hélices baseadas na hidrofobicidade da seqüência de resíduos.

Em [FPP95] RNAs são utilizadas para predizer a distância entre alfa-carbonos a partir da seqüência de resíduos.

3.5 Métodos Baseados em Homologia

Stephen Altschul *et all* desenvolveram em 1990 a ferramenta BLAST (*Basic Local Alignment Search Tool*) [AGM+90] para determinação de estrutura tridimensional a partir de alinhamento com seqüências homólogas.

O algoritmo BLAST baseia-se em uma medida de similaridade local para definir estruturas homólogas e portanto a possível estrutura terciária de uma seqüência de resíduos. Os algoritmos *Gapped* BLAST e PSI-BLAST [AMS+97], desenvolvidos 7 anos depois pelo mesmo autor utilizam estatísticas e heurísticas para gerar

alinhamentos faltantes e aumentar a sensibilidade sobre similaridades fracas mas biologicamente relevantes.

3.6 Sistemas Híbridos e Ganho de Informação

Neste trabalho desenvolveremos um método que se vale da informação sobre a estrutura secundária oriunda de previsões de RNAs para guiar algoritmos de MC em direção à conformação nativa de proteínas.

Em [SR04] há um estudo sobre o efeito do conhecimento sobre a estrutura secundária na determinação da estrutura tridimensional de uma proteína. Os autores utilizam teoria da informação para gerar distribuições de ângulos diedrais a partir da estrutura secundária conhecida e da estrutura secundária prevista. O efeito da informação sobre a estrutura secundária na predição da conformação nativa tridimensional é obtido através do cálculo da mudança de entropia, e é mostrado como o grau de precisão dos métodos de previsão de estrutura secundária afetam a previsão da estrutura terciária.

Ainda de acordo com os autores, apenas uma pequena fração da incerteza sobre os ângulos diedrais da conformação nativa (14 a 38%, com resolução entre 20 e 90º) são resolvidos com o conhecimento exato da estrutura secundária. Se a informação sobre a estrutura secundária é proveniente de métodos de previsão a perda de informação é tal que um método com o grau máximo atual de 75% de acurácia retém apenas um terço da informação estrutural codificada na estrutura secundária. Por outro lado, os autores mostram que o ganho de informação aumenta exponencialmente com o aumento na precisão dos métodos de predição da estrutura secundária.

Capítulo 4

Metodologia

O método Monte Carlo realiza uma busca aleatória no espaço de parâmetros de uma função, baseada em uma distribuição de probabilidade. Na aplicação do dobramento de proteínas, os parâmetros são os ângulos diedrais Φ e Ψ de cada aminoácido presente na seqüência da proteína em análise. Teoricamente, na técnica Monte Carlo se poderia produzir valores aleatórios de forma a cobrir todo o hiper-espaço de parâmetros com pontos. Na prática é necessário reduzir a complexidade do problema restringindo as regiões visitadas àquelas de maior interesse de modo a tornar a simulação tratável computacionalmente. Iremos restringir a região para a qual pontos devem ser gerados em função do Mapa de Ramachandran e das RNAs.

A metodologia proposta para o método MC-RNA é aplicar Monte Carlo com distribuição de probabilidade construída a partir das previsões de RNAs em conjunto com o Mapa de Ramachandran. Enquanto o Mapa de Ramachandran restringe o espaço amostral às regiões de ângulos permitidos fisicamente, a previsão das RNAs limita a busca às regiões do Mapa de Ramachandran onde ocorre a estrutura predita.

Para o cálculo da energia livre das conformações criadas pelo modelo, utilizaremos o campo de força MM3Pro implementado no pacote de mecânica molecular TINKER. O campo de força MM3Pro, que é baseado no campo de força MM3 [AYL89a], é um conjunto de parâmetros e expressões ajustados para o cálculo das forças de interação entre átomos de uma proteína.

4.1 Redução do espaço de busca

As abordagens para redução do espaço de conformações tridimensionais de proteínas dos algoritmos propostos neste trabalho baseiam-se em dois métodos: Mapa de Ramachandran e RNAs para predição de estrutura secundária.

O Mapa de Ramachandran pode ser utilizado como ferramenta para a redução do espaço de estados, pois separa as combinações de ângulos diedrais Φ e Ψ em regiões permitidas e proibidas. Nas regiões proibidas encontram-se combinações de ângulos que colocariam os átomos a distâncias menores que seus raios de Van der Walls. As regiões permitidas são ainda divididas por tipo de estrutura secundária. Além desta divisão, é possível ainda calcular a superfície de energia do Mapa a partir dos pares de ângulos diedrais. Podemos portanto dividir a abordagem de redução do espaço de conformações por Monte Carlo em duas: sem e com conhecimento sobre a estrutura secundária.

FIGURA 4.1 – Segmento da previsão de estrutura secundária para a Mioglobina obtida pelo método de B. Rost. Na primeira linha a seqüência de resíduos da Mioglobina, na segunda linha a estrutura ($H = \alpha$-hélice), na terceira linha a probabilidade de acerto da previsão da estrutura secundária, e na quarta linha a estrutura secundária com probabilidade $p >= 0,5$.

Para escolher ângulos diedrais de um resíduo quando não temos informações a priori da estrutura secundária à qual ele pertence, utilizamos a região permitida do Mapa de Ramachandran. Como a região proibida ocupa a maior parte do Mapa de Ramachandran, se utilizarmos apenas a região permitida reduzimos automaticamente o espaço de estados a serem explorados por um algoritmo de busca.

Caso tenhamos conhecimento sobre a estrutura secundária a qual o resíduo pertence, podemos nos valer desta informação para restringir ainda mais o espaço de estados. A região permitida é dividida em sub-regiões conforme a estrutura secundária formada pelos pares de ângulos. Para este trabalho utilizamos as predições de estruturas secundárias de RNAs através do Método PHDsec [RS93, Ros96] disponíveis *on-line*[1]. As RNAs recebem como entrada a seqüência de aminoácidos e fornecem o tipo de estrutura secundária (α-hélice, folha-β ou *coil*) e o grau de confiança na previsão para cada aminoácido da cadeia (Figura 4.1).

Para escolha randômica, simplesmente geram-se dois números randômicos que são mapeados para a distribuição de ângulos da área permitida do mapa ou para a sub-região correspondente à estrutura secundária.

4.2 Método MC-RNA - Aplicado ao Dobramento de Proteínas

Como dito na introdução, O método MC-RNA é um modelo de simulação estocástico que utiliza RNAs para a redução do espaço de busca. O modelo consiste em simulação estocástica pelo método MC, com a probabilidade de transição$P(\mu \rightarrow \nu)$ determinada pela combinação das probabilidades $g(\mu \rightarrow \nu)$ e $A(\mu \rightarrow \nu)$, estas por sua vez determinadas respectivamente pela RNA e pela diferença de energiaΔE entre os estados ν e μ.

O algoritmo geral inicial para a simulação MC aplicada ao problema de dobramento de proteínas é o seguinte:

- Inicialização: criar estado inicial μ composto por seqüência de pares de ângulos diedrais.
- Calcula a energia total da proteína E_ν.
- Enquanto o sistema não atinge equilíbrio:

[1] *The PredictProtein server* em http://www.embl-heidelberg.de/predictprotein/submit_def.html#top.

  - Escolhe aleatoriamente um resíduo.
  - Cria aleatoriamente estado ν a partir de μ, escolhendo aleatoriamente um novo par de ângulos diedrais para resíduo escolhido.
  - Calcula a energia total E_μ da proteína com a nova conformação.
  - calcula $A(\mu \to \nu) = \begin{cases} e^{-\beta(E_\nu - E_\mu)} & \text{se } E_\nu - E_\mu > 0 \\ 1 & \text{caso contrário .} \end{cases}$
  - Efetua transição $\mu \to \nu$ com probabilidade $P(\mu \to \nu)$.

- O sistema atinge o equilíbrio quando a taxa de transições efetuadas ficar abaixo de limite arbitrado ou for zero.

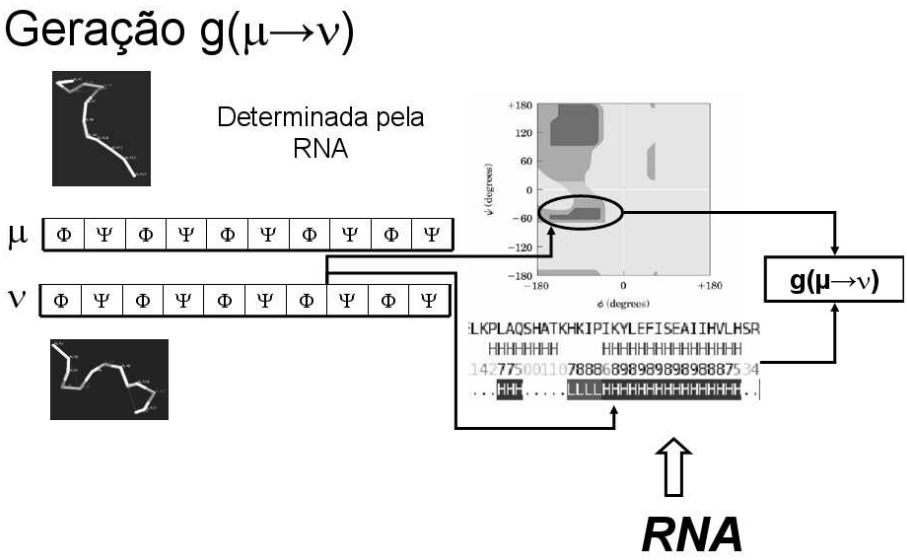


FIGURA 4.2 – O novo estado ν é obtido alterando-se um par de ângulos diedrais de um resíduo. Para ser gerada a transição $\mu \to \nu$ é necessário que o novo par de ângulos pertença à região permitida do mapa de Ramachandran. Se a região do mapa coincidir com a classificação da RNA para a estrutura secundária, então $g(\mu \to \nu)$ equivale ao grau de confiança na previsão da rede.

No método MC-RNA o papel da RNA é determinar a quantidade $g(\mu \to \nu)$, ou seja, a probabilidade de se gerar um novo estado ν a partir de um estado μ. O novo estado ν é o resultado de duas escolhas aleatórias: primeiro ocorre a escolha de um entre todos os resíduos que compõem a proteína, e em seguida a escolha de novos valores para o par de ângulos diedrais deste resíduo. Feito isto duas ações são tomadas: verifica-se se o par de ângulos está na região permitida do mapa de Ramachandran, e se a região do mapa corresponde à estrutura prevista pela RNA para aquele resíduo, naquela posição da proteína. Então a probabilidade de geração da transição para o estado ν é calculada conforme o grau de confiança na previsão da RNA (Figura 4.2). O algoritmo inicial para a simulação do método MC-RNA é então o seguinte:

- Inicialização: criar estado inicial μ composto por seqüência de pares de ângulos diedrais.

- Calcula a energia total E_μ da proteína.
- Enquanto o sistema não atinge equilíbrio:
  - Escolhe aleatoriamente um resíduo.
  - Cria aleatoriamente estado ν a partir de μ escolhendo aleatoriamente um novo par de ângulos diedrais para o resíduo escolhido
  - Se $\nu \in$ { região proibida do mapa de Ramachandran (MR)}:
    - $g(\mu \to \nu) = 0$
  - Se RNA não classifica resíduo como pertencente à α-hélice ou à folha-β:
    - $g(\mu \to \nu) = 1$
  - Caso contrário:
    - Se $\nu \in$ { região do MR correspondente à classificação da RNA }:
      - $g(\mu \to \nu) =$ grau de confiança na classificação da RNA
    - Caso Contrário:
      - $g(\mu \to \nu) = 1-$ grau de confiança na classificação da RNA
  - Se $g(\mu \to \nu) \neq 0$
    - Calcula a energia total E_ν da proteína com a nova conformação.
    - calcula $A(\mu \to \nu) = \begin{cases} e^{-\beta(E_\nu - E_\mu)} & \text{se } E_\nu - E_\mu > 0 \\ 1 & \text{caso contrário .} \end{cases}$
  - Efetua transição $\mu \to \nu$
    com probabilidade $P(\mu \to \nu) = g(\mu \to \nu)\ A(\mu \to \nu)$.
- O sistema atinge o equilíbrio quando a taxa de transições efetuadas ficar abaixo de limite arbitrado ou for zero.

O algoritmo descrito acima é fiel ao método MC-RNA, mas ainda é pouco eficiente em termos computacionais. Em vez de escolher aleatoriamente pares de ângulos em todo o espaço de estados, podemos simplesmente escolher os ângulos diretamente de um espaço previamente reduzido. Este espaço reduzido pode ser obtido em uma primeira etapa eliminando as combinações de ângulos proibidas pelo mapa de Ramachandran. Sobre o espaço de ângulos permitidos, podemos utilizar a classificação de estrutura secundária da RNA e eliminar as combinações de ângulos das áreas que não correspondem à estrutura secundária prevista. Para diminuir ainda mais a complexidade do modelo, podemos restringir o número de resíduos passíveis de sofrerem alteração, formando blocos rígidos de α-hélices ou segmentos de folhas-β. Todas estas otimizações são implementadas no algoritmo proposto, que é baseado no algoritmo inicial descrito anteriormente, e será visto em detalhes a seguir.

O algoritmo proposto para simular o método MC-RNA implementa uma abordagem espacial e paralela. A implementação é realizada em duas fases: (1) geração de estruturas por MC com consulta às RNAs, e (2) minimização e clusterização.

Na primeira fase do algoritmo é gerada uma amostra de N conformações da proteína em estudo. A geração das conformações é feita de acordo com a classificação das RNAs e com o grau de confiança na classificação. Por exemplo, se para

um determinado resíduo em uma determinada posição na seqüência de resíduos a classificação da RNA é α-hélice com grau de confiança de $0,7$, então podemos esperar que em média 70% das conformações da amostra apresentem ângulos típicos de α-hélice no resíduo citado.

A segunda fase do algoritmo consiste em minimização de energia por descida de gradiente e posterior clusterização das conformações minimizadas. A minimização de energia serve para acomodar átomos em posições mais estáveis de maneira a se poder avaliar a real energia potencial da proteína. A clusterização é o método que resultará na escolha das conformações mais próximas da conformação nativa.

4.2.1 Fase 1: Geração de conformações

Nesta seção são explicados com detalhes os passos para se formar os conjuntos iniciais de conformações. A primeira fase do método de simulação proposto constitue-se na geração de um conjunto de conformações tridimensionais aleatoriamente, com a utilização em maior ou menor grau de informação sobre a estrutura secundária das seqüências funcionando como viés.

As proteínas escolhidas para este estudo foram retiradas da lista de proteínas alvo dos experimentos CASP[2]. Para os experimentos foram escolhidas quatro estruturas relativamente pequenas: $1i74$ (domínio 2), $1kkg$, $1g7d$ domínio C-terminal e $1j8b$.

Para cada uma das seqüências foram produzidos grupos de conformações de acordo com 3 algoritmos, que chamaremos de MC (Monte Carlo), MC-RNA (Monte Carlo com RNAs) e MC-Ideal (Monte Carlo com RNA *ideal*). O método MC-RNA utiliza o método MC com informação oriunda da previsão de estrutura secundária da seqüência por RNAs, juntamente com a probabilidade de acerto da RNA. Os outros dois métodos foram utilizados a título de comparação: o MC que consiste em Monte Carlo sem informação sobre estrutura secundária, e o MC-Ideal, que se utiliza de informação real sobre a estrutura secundária das seqüências analisadas.

Fase 1, 1^a Parte: Banco de Dados EVA

O projeto EVA (*EValuation of Automatic protein structure prediction*) é tocado por pesquisadores do grupo CUBIC (*Columbia University Bioinformatics Center*) da Universidade de Columbia, em conjunto com os grupos SALI-lab da Universidade UCFS (Califórnia, São Francisco), PDG (Protein Design Group) da Universidade de Madrid, e demais colaboradores. O objetivo principal do EVA é "Fornecer uma análise contínua, totalmente automatizada e estatisticamente significativa dos servidores de predição de estruturas". Entre outras atividades, o projeto EVA mantém um banco de dados continuamente atualizado de proteínas não homólogas. O banco de dados de proteínas não homólogas EVA é um subconjunto do universo de proteínas de estrutura conhecida. As proteínas destes subconjunto são ditas não homólogas por que não há entre elas nenhum par de proteínas com mais de33% de resíduos idênticos em seqüências de alinhamento de mais de100 resíduos [Ros99].

A lista de proteínas deste subconjunto encontra-se disponível para download no site do projeto. O trabalho realizado utilizou a lista de proteínas do banco de

[2]http://predictioncenter.gc.ucdavis.edu/

dados EVA de 11 de agosto de 2005. Esta lista contém 3419 proteínas ou sub-cadeias de proteínas não homólogas.

A preparação dos dados da lista de proteínas não homólogas EVA constituiu-se das seguintes fases:

1. ***Download* dos arquivos PDB -** Os arquivos PDB (*Protein Data Bank* - http://www.rcsb.org) são arquivos que contém entre outros dados a estrutura primária da proteína (a seqüência de resíduos) e as coordenadas atômicas de átomos. Arquivos PDB de todas as 3419 proteínas foram baixados do banco de dados de proteínas PDB. Do total de proteínas da lista EVA, 44 proteínas não estavam disponíveis e foram substituídas por similares.

2. **Processamento dos arquivos PDB**

   As estruturas tridimensionais das proteínas descritas nos arquivos PDB podem ser visualizadas com o auxílio de softwares de visualização com o RASMOL. Embora a visualização tridimensional de uma proteína a partir de um arquivo PDB baixado do repositório corresponda à conformação nativa da proteína, se utilizamos as informações das coordenadas atômicas de átomos presentes neste arquivo PDB sem nenhum pré-processamento não podemos obter medidas confiáveis sobre a proteína descrita. Os arquivos PDB são gerados a partir de observações oriundas de processos físicos para análise de proteínas como Difração por Raio-X e Ressonância Nuclear Magnética. Por mais precisos que sejam estes métodos, pequenos erros nas coordenadas atômicas dos arquivos PDB podem levar softwares de analise a erros grotescos. Os exemplos mais comuns são determinação de energia potencial excessivamente alta devido à distâncias entre átomos artificialmente pequenas, ou mesmo devido à sobreposição de átomos (energia Infinita). Para melhorar a usabilidade e testar a confiabilidade dos arquivos PDB, todos passaram por um pré-processamento, descrito a seguir.

   (a) ***PDB -> XYZ***

   A ferramenta utilizada nesta etapa é o PDBXYZ, do pacote de modelagem molecular Tinker. Esta ferramenta recebe como entrada um arquivo PDB e produz um arquivo XYZ. Os arquivos XYZ são o tipo básico de arquivo de coordenadas cartesianas do Tinker, e contém a descrição dos átomos. um por linha, em função de suas coordenadas cartesianas. Além do nome do átomo e de suas coordenadas cartesianas cada linha contém uma lista dos átomos aos quais o átomo está ligado, e o número que representa o tipo de átomo no campo de força escolhido. Neste trabalho o campo de força utilizado foi o campo MM3Pro, que é desenvolvido especificamente para o cálculo das forças de interação eletrostáticas levando em consideração características peculiares às proteínas.

   Entre os parâmetros do comando PDBXYZ estão o campo de força utilizado e a cadeia a ser transcrita. Muitas seqüências listadas na lista EVA não correspondem a proteínas completas, mas cadeias específicas de uma proteína. Nestes casos o arquivo XYZ resultante do comando contém apenas as coordenadas dos átomos da cadeia selecionada.

   (b) ***XYZ -> INT***

Os arquivos INT contém uma representação interna da estrutura molecular das proteínas. O formato é semelhante ao dos arquivos XYZ. Porém, ao invés das coordenadas cartesianas dos átomos, as coordenadas internas consistem em uma distância a um átomo previamente determinado, e dois ângulos de ligação ou um ângulo de ligação e um ângulo diedral com átomos predecessores. Para transformar os arquivos XYZ em arquivos INT foi utilizada a ferramenta XYZINT e o campo de força MM3Pro.

3. **Minimização**

   Após a preparação dos arquivos INT, um para cada proteína da lista EVA, foi feita minimização de energia de cada proteína com a ferramenta MINIROT do pacote Tinker. A ferramenta MINIROT realiza minimização de energia por descida de gradiente no espaço formado por ângulos diedrais. Para tanto o software recebe como entrada o arquivo INT a ser minimizado e o critério de parada, na forma do *rms (root mean square)* do gradiente em $kcal/mole/\AA$.

   O critério de parada das minimizações utilizado foi parar quando o*rms* do gradiente de energia fosse menor do eu $1kcal/mole/\AA$: o suficiente para acomodar os átomos em posições realistas de acordo com suas interações eletrostáticas.

   O resultado das minimizações são arquivos INT_2, com os ângulos ligeiramente alterados em relação aos arquivos INT originais pela descida de gradiente.

4. **Pós processamento** Após a minimização das proteínas, o pós-processamento dos dados foi feito para formar um banco de dados de resíduos, respectivas estruturas secundárias e ângulos diedrais. Nesta fase, além do pacote Tinker também foi utilizado o software DSSP [KS83a]. Os ítens a seguir descrevem o DSSP e o método utilizado para a formação do banco de dados.

   (a) **INT_2 -> DSSP_2**

   Como os ângulos dos arquivos INT foram alterados pela descida de gradiente, a conformação espacial das proteínas descritas por eles pode ter sofrido alguma alteração. Para extrairmos a estrutura secundária destas proteínas após a minimização, precisamos gerar novos arquivos PDB, e utilizarmos o programa DSSP para a leitura da estrutura secundária.

   O programa DSSP é a implementação do artigo de1983 de Kabsch e Sander [KS83b] por eles mesmos sobre descrição de estrutura secundária de proteínas a partir de reconhecimento de padrões de pontes de hidrogênio e formas geométricas. O programa DSSP não prevê estrutura secundária, mas dado um arquivo de coordenadas atômicas no formato PDB, define a estrutura secundária a que pertencem os resíduos, algumas características geométricas e a superfície de exposição ao solvente.

   Destas características utilizamos aqui apenas a capacidade de descrever a estrutura secundária a partir de arquivos do tipo PDB. O processo é simples: para cada proteína minimizada geramos a partir do arquivo de ângulos INT_2 o arquivo correspondente XYZ_2 com a ferramenta do Tinker INTXYZ. Em seguida utilizamos outra ferramenta do Tinker, o XYZPDB para gerarmos o arquivo PDB_2 correspondente. Em seguida

utilizamos o arquivo DSSP para gerarmos a partir dos arquivos PDB e PDB_2 os arquivos DSSP e DSSP_2. Os arquivos DSSP e DSSP_2 contém respectivamente a estrutura secundária da proteína antes da minimização e depois da minimização.

(b) **Lista de Resíduos**

Em cada um dos passos de pré-processamento e minimização, proteínas foram descartadas. Na fase de pré-processamento os descartes ocorreram em função de falhas nos arquivos PDB originais que impedissem passos do pré-processamento, como falta de determinados átomos ou resíduos. Já entre as proteínas que ultrapassaram a fase de pré-processamento, muitas não obtiveram sucesso durante o processo de minimização. Ao final, das 3419 proteínas e sub-cadeias de proteínas listadas no EVA, restaram2327 que terminaram as suas respectivas minimizações por atingirem o critério de parada.

Para este trabalho foi desenvolvido um conjunto de ferramentas para manipular dados de estrutura secundária e ângulos diedrais denominado DSTK (*Diedral angles and Secondary strucure TollKit*). O comando SANITIZEDSSP é uma ferramenta do DSTK que extrai dos arquivos DSSP informações sobre os ângulos diedrais dos resíduos e o tipo de estrutura a que pertencem. O comando ainda recebe como parâmetros o arquivo DSSP, um arquivo texto com a lista de proteínas EVA que são formadas por mais de uma cadeia, e a opção de desprezar ou não os resíduos da extremidade da seqüência. O SANITIZEDSSP então lê o arquivo DSSP, extrai os dados apenas da cadeia de interesse (indicada por uma letra no nome da proteína EVA), e mapeia os 8 tipos de estrutura secundária para 3, conforme a Tabela 4.1.

DSSP		DSTK	
Estrutura	Símbolo	Símbolo	Estrutura
4-hélice (α-hélice)	H	A	α-hélice
3-hélice (3_{10}-hélice)	G		
5-hélice (π-hélice)	I		
segmento isolado de folha-β	B	B	folha-β
folha-β	E		
curva (*turn*) com ponte de H	T	C	*coil*
centro de curva com 5 resíduos	S		
coil	C		

TABELA 4.1 – Mapeamento da representação de estrutura secundária do DSSP para a representação utilizada neste trabalho

O comando SANITIZEDSSP recebeu todos os arquivos DSSP_2 e produziu o arquivo EVA_2.DAT, contendo 377540 linhas, cada uma com as informações sobre um resíduo das 2327 proteínas efetivamente minimizadas da lista EVA. O arquivo EVA_2.DAT tem o formato

. . . . . .

```
 .   .  .     .        .    .
 .   .  .     .        .    .
25   V  A  -66.2  -19.8 2bem_A
26   Q  A  -67.7  -17.3 2bem_A
27   Y  A  -99.8   -3.2 2bem_A
28   E  C -159.5   65.6 2bem_A
29   P  A  -68.4  -16.8 2bem_A
30   Q  A  -77.6  -11.8 2bem_A
31   S  A -108.8   29.2 2bem_A
32   V  C  -80.4   75.8 2bem_A
33   E  B -111.7  130.0 2bem_A
34   G  B -143.8 -160.4 2bem_A
35   L  B  -82.9  149.6 2bem_A
36   K  C  -89.0  177.8 2bem_A
 .   .  .     .        .    .
 .   .  .     .        .    .
 .   .  .     .        .    .
```

onde a primeira coluna contém o número do resíduo na seqüência da cadeia da proteína, a segunda coluna a identificação do resíduo, a terceira coluna o tipo de estrutura secundária a que o resíduo pertence, as colunas 4 e 5 são os ângulos diedrais Φ e Ψ, e a quinta coluna contém o código PDB de 4 letras da proteína seguido de uma letra indicando a cadeia.

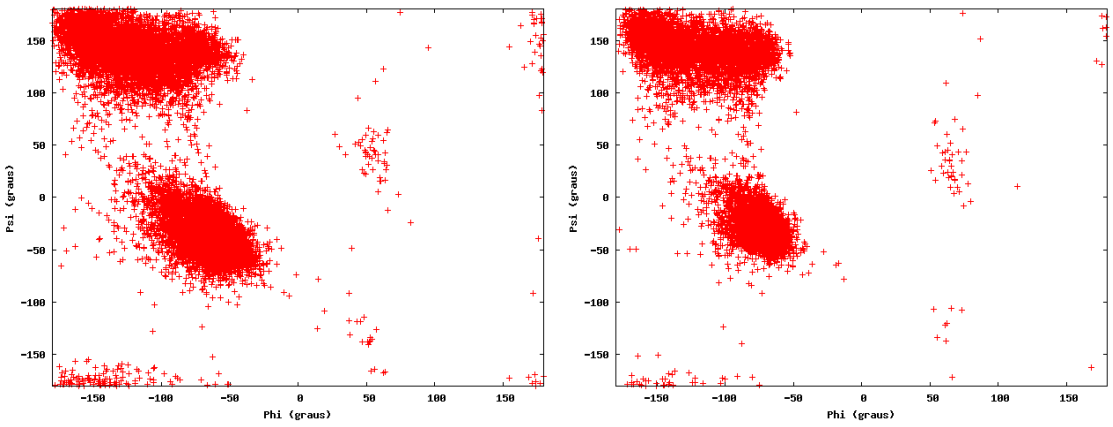


FIGURA 4.3 – Mapas de Ramachandran representando os ângulos diedrais de todos os resíduos alanina da lista EVA antes (esq.) e depois (dir.) da minimização. No eixo horizontal o ângulo diedral Φ e no vertical o ângulo diedral Ψ.

As demais ferramentas do DSTK utilizam este arquivo como fonte para a produção de mapas de Ramachandran específicos para tipos de resíduos e estruturas secundárias. A título de ilustração, a Figura 4.3 mostra os mapas de Ramachandran específicos para o resíduo alanina quando participante de estruturas α-hélice ou folha-β. Os mapas foram feitos selecionando nos arquivos DSSP.DAT e DSSP_2.DAT apenas os ângulos das linhas com os símbolos A

(alanina) na coluna 2 e A ou B (α-hélice ou folha-β) na coluna 3. A confecção do arquivo DSSP.DAT com dados das proteínas pré-minimização teve o objetivo de permitir a comparação com os ângulos preferenciais após a minimização. Como mostra a Figura 4.3, a concentração de ângulos nas regiões atribuídas às duas estruturas secundárias aumentou após a minimização.

Fase 1, 2^a Parte: MC-RNA

O método de geração de estruturas MC-RNA utiliza a previsão da estrutura secundária para a redução do espaço de busca. Ou seja, geramos um conjunto de estruturas a partir da seqüência de resíduos de uma proteína escolhendo os seus ângulos diedrais por Monte Carlo, mas restringindo os ângulos que podem ser escolhidos com informação proveniente das RNAs. O método de geração de estruturas MC-RNA é descrito a seguir passo a passo.

1. **Previsão RNA**

   O primeiro passo é consultar o servidor *Predict Protein* para obter a previsão da estrutura secundária. Como explicado na seção 4.1, o servidor utiliza RNAs treinadas para prever a estrutura secundária a partir de seqüências de resíduos. A seqüências de resíduos da proteína é submetida ao servidor, e este envia a previsão da estrutura secundária para cada resíduo junto com o grau de confiança da previsão, que varia de resíduo para resíduo.

   Partindo desta previsão, o *script* do DSTK PROF2LISTARES cria o arquivo LISTA_RES.DAT. O segmento a seguir é a parte do arquivo LISTA_RES.DAT da proteína $1j8b$ que vai do resíduo 47 ao 64 e que está em destaque na Figura 4.4.

```
 .  .  .  .  .  .  .  .
 .  .  .  .  .  .  .  .
 .  .  .  .  .  .  .  .
47 R B 0.7 B AC
48 R B 0.8 B AC
49 I B 0.8 B AC
50 D B 0.7 B AC
51 I B 0.3 X X
52 D C 0.6 C AB
53 P A 0.2 X X
54 S A 0.0 X X
55 L A 0.2 X X
56 M A 0.4 X X
57 E A 0.4 X X
58 D C 0.5 C AB
59 D C 0.6 C AB
60 K A 0.8 A BC
61 E A 0.8 A BC
62 M A 0.9 A BC
63 L A 0.9 A BC
64 E A 0.9 A BC
```

. . . . . .
. . . . . .
. . . . . .

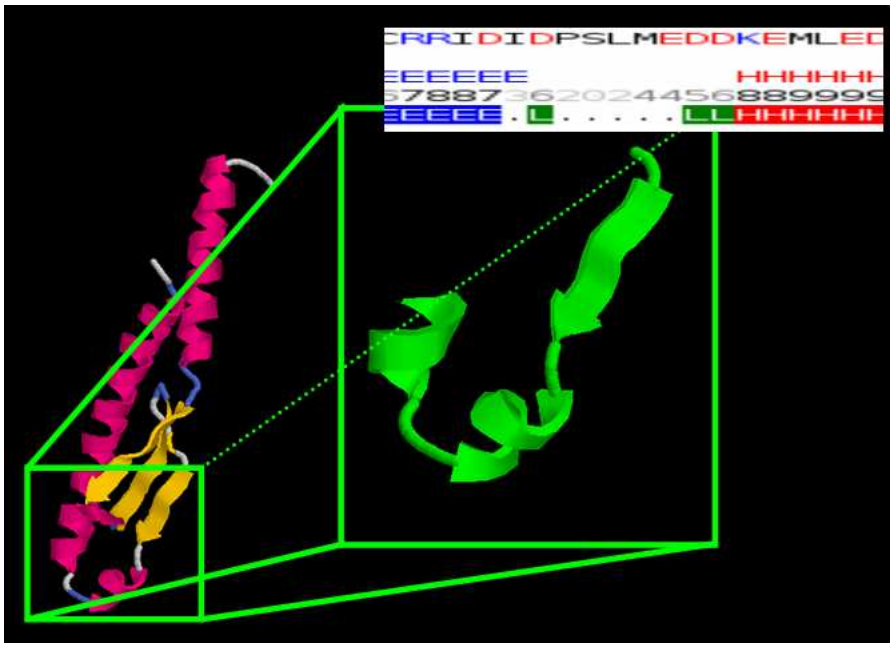


FIGURA 4.4 – À esquerda a proteína $1j8b$. À direita em destaque o segmento que vai do resíduo 47 ao 64. Em cima à direita a previsão das RNAs para o segmento ($E = \beta$, $L = turn/coil$, $H = \alpha$).

A primeira coluna é o índice do resíduo na seqüência, a segunda o tipo de resíduo, a coluna 3 contém a estrutura real da proteína[3], na coluna 4 estão os graus de confiança nas previsões das RNAs, na quinta coluna a estrutura prevista pela RNA, e por fim na sexta coluna as estruturas não previstas para a aquele resíduo. Quando o grau de confiança da previsão (coluna4) é < 0.5, as ultimas duas colunas recebem o símbolo X que funciona como coringa: indica que o resíduo pode pertencer a qualquer estrutura. A previsão das RNAs do servidor *Predict Protein* para cada resíduo pode ser α-hélice (H), folha-β (B) ou outros (C).

2. Geração da amostra de estruturas

   Neste passo as N estruturas da amostra são geradas por MC, escolhendo os ângulos em regiões específicas do mapa de Ramachandran de acordo com as previsões das RNAs. Esta etapa é realizada pelo programa do DSTK DIHEDRAL_CHANGE2 em 3 passos, que são descritos a seguir.

   (a) **Parâmetros**

[3] A coluna 3 contém a estrutura secundária conhecida da proteína, se disponível, extraída de um arquivo DSSP. Os dados desta coluna não são utilizados em nenhuma computação, sendo apenas informativos e servindo de comparação com a estrutura prevista pelas RNAs.

O executável DIHEDRAL_CHANGE2 carrega para a memória 3 arquivos: o arquivo EVA_2.DAT que contém a lista de resíduos, estruturas secundárias e ângulos diedrais de todas as seqüências da lista eva, o arquivo da proteína em análise no formato INT do Tinker, contendo a lista de resíduos e representação interna por distâncias e ângulos, e o arquivo LISTA_RES.DAT contendo as previsões das RNAs para aquela seqüência. Os demais parâmetros do executável são, além de parâmetros para conferência e validação dos arquivos carregados, a quantidade de estruturas a serem geradas, e um número inteiro e negativo para o gerados da seqüência randômica.

(b) **Geração de mapas**

Uma vez carregados com os dados e parametrizado, o executável gera mapas de Ramachandran específicos para cada par de resíduo e estrutura(s) secundária(s) do arquivo LISTA_RES.DAT. Por exemplo, a linha do arquivo correspondente ao resíduo 50 da proteína $1j8b$ é 50 D B 0.7 B AC o que significa que o quinquagésimo resíduo da proteína é o Aspartato (D), e que a rede neural prevê com 70% de confiança que o resíduo faz parte de uma folha-β (B, na quinta coluna). O executável criará então duas listas de ponteiros: uma que irá apontar para todos os resíduos Aspartato da lista eva que participem de folhas-β, e outro para o conjunto de Aspartato que estejam em α-hélices, *turns* ou *coil*. As duas listas agora são dois mapas de Ramachandran: um com Aspartatos em folhas-β, e outro com Aspartatos fora de folhas-β.

O procedimento acima á repetido para cada par de tipo de resíduo e subconjunto de estruturas. Ao final deste procedimento formamos um conjunto de mapas de Ramachandran específicos. Este conjunto é o novo espaço de busca reduzido para o algoritmo MC: somente os ângulos deste conjunto serão disponibilizados para a geração randômica de estruturas.

(c) **Criação das estruturas com consulta às RNAs** Uma vez criado o conjunto de mapas de Ramachandran, o executável passa para a fase de geração de estruturas por MC. Inicialmente é criada uma cópia do arquivo INT representando a estrutura. O procedimento então consiste em, para cada resíduo da proteína:

i. gerar um valor aleatório entre 0 e 1,
ii. se o valor gerado for menor ou igual ao grau de confiança da previsão da RNA, escolher randomicamente um par de ângulos diedrais do mapa de Ramachandran correspondente à previsão,
iii. caso contrário, escolher do mapa de Ramachandran contrário à previsão,
iv. alterar os ângulos diedrais correspondentes na representação interna da cópia do arquivo INT, e salvar o arquivo em disco

O procedimento descrito acima é repetido N vezes, gerando N arquivos INT com conformações diferentes. No exemplo dado anteriormente do quinquagésimo resíduo da proteína $1j8b$, a linha que o descreve é 50 D B 0.7 B AC, indicando que para N grande, 70% das conformações geradas terão os ângulos diedrais do seu resíduo de número 50 escolhidos

aleatoriamente do mapa de Ramachandran que contém apenas ângulos de Aspartatos quando participantes de folhas-β. Da mesma forma, os 30% de conformações restantes terão os ângulos retirados dentre os ângulos de Aspartatos fora de folhas-β.

Para verificar a eficácia do método MC-RNA foram realizados métodos controle: os métodos MC e MC-DSSP. O primeiro é o método MC tradicional, sem nenhuma informação sobre a estrutura secundária. O segundo é o método MC com informação da estrutura secundária real da conformação nativa conhecida da proteína. As motivações destes testes são servir de comparativo com o método MC-RNA, e demonstrar o efeito da informação sobre a estrutura secundária na capacidade do método MC de encontrar a conformação nativa da proteína.

Fase 1, Controle: MC

O método de geração de estruturas MC não utiliza a previsão da estrutura secundária para a redução do espaço de busca. Ou seja, geramos um conjunto de estruturas a partir da seqüência de resíduos de uma proteína escolhendo os seus ângulos diedrais diretamente por Monte Carlo. O objetivo deste método é comparativo: demonstrar o ganho de performance do método MC-RNA em relação ao método MC tradicional. A única restrição do espaço de busca realizada neste método é a escolha dos ângulos a partir de mapas de Ramachandran contendo ângulos permitidos para o tipo de resíduo específico.

O método de geração de estruturas MC é descrito a seguir passo a passo.

1. **Lista de resíduos**

   A lista de resíduos LISTA_RES.DAT, que no método MC-RNA é criada a partir da previsão das RNAs é alterada de modo a que o método MC não obtenha informações que permitam a restrição do espaço de busca baseado nas informações. A única restrição de espaço de busca é feita em relação ao tipo de resíduo.

   Para ilustrar o procedimento utilizaremos a mesma fração da seqüência da mesma proteína utilizada como exemplo no método MC-RNA: a proteína1$j8b$, resíduos 47 ao 64:

```
 .  .  .  .  .  .  .  .
 .  .  .  .  .  .  .  .
 .  .  .  .  .  .  .  .
47  R  B  0.7  X  X
48  R  B  0.8  X  X
49  I  B  0.8  X  X
50  D  B  0.7  X  X
51  I  B  0.3  X  X
52  D  C  0.6  X  X
53  P  A  0.2  X  X
54  S  A  0.0  X  X
55  L  A  0.2  X  X
56  M  A  0.4  X  X
```

```
57 E A 0.4 X X
58 D C 0.5 X X
59 D C 0.6 X X
60 K A 0.8 X X
61 E A 0.8 X X
62 M A 0.9 X X
63 L A 0.9 X X
64 E A 0.9 X X
 . . . . . .
 . . . . . .
 . . . . . .
```

A diferença para a lista original é que as colunas 6 e 7 aonde vai a previsão da estrutura secundária e o complemento da previsão são preenchidos com um X, significando qualquer estrutura.

2. Geração da amostra de estruturas

   Neste passo as N estruturas da amostra são geradas por MC, escolhendo os ângulos em mapas de Ramachandran específicos para cada tipo de resíduo. A diferença em relação ao método MC-RNA é que os mapas de Ramachandran específicos contém ângulos diedrais do resíduo independente do tipo de estrutura secundária a que pertence.

   (a) **Parâmetros**

   O executável DIHEDRAL_CHANGE2 carrega os mesmos parâmetros do método MC-RNA: o arquivo EVA_2.DAT, o arquivo da proteína em análise no formato INT do Tinker, e o arquivo LISTA_RES.DAT modificado.

   (b) **Geração de mapas**

   Uma vez carregados com os dados e parametrizado, o executável gera mapas de Ramachandran específicos para cada tipo de resíduo presente no arquivo LISTA_RES.DAT. Isto significa que, ao contrário do método MC-RNA onde tínhamos inúmeras possibilidades de mapas com combinações de resíduos e estruturas secundárias, teremos aqui no máximo 20 mapas, um para cada tipo de resíduo.

   Para a o exemplo da linha do arquivo da proteína $1j8b$ 50 D B 0.7 X X, o executável criará apenas uma lista de ponteiros que irá apontar para todos os resíduos Aspartato da lista eva, não importando a estrutura secundária à que pertençam.

   (c) **Criação das estruturas** Uma vez criado o conjunto de mapas de Ramachandran, o executável passa para a fase de geração de estruturas por MC:

      i. Escolher randomicamente um par de ângulos diedrais do mapa de Ramachandran correspondente ao tipo de resíduo

      ii. alterar os ângulos diedrais correspondentes na representação interna da cópia do arquivo INT, e salvar o arquivo em disco

O procedimento descrito acima é repetido N vezes, gerando N arquivos INT com conformações diferentes. Para N grande, a distribuição dos ângulos do resíduo de número nas conformações geradas seguirá a distribuição do Aspartato na lista Eva.

Fase 1, Controle: MC-DSSP

O método de geração de estruturas MC-DSSP utiliza a informação da estrutura secundária real da conformação nativa conhecida da proteína para a redução do espaço de busca. Ou seja, realizamos a restrição do espaço de busca como no método MC-RNA, mas nos valemos de informação privilegiada, obtida passando a ferramenta DSSP no arquivo PDB baixado do servidor. Assim utilizamos a informação da estrutura secundária verdadeira da proteína como se fosse a previsão de uma hipotética RNA perfeita, capaz de prever a estrutura secundária de proteínas com índice de acerto de 100%.

Este método, assim como o método MC descrito anteriormente, foi realizado antes do método MC-RNA, e cumpriu o objetivo de demonstrar o ganho de performance do método MC para o dobramento de proteínas quando auxiliado com informações sobre a estrutura secundária. O método MC-DSSP é o seguinte:

1. **DSSP**

   Como explicado anteriormente, o software DSSP não prevê estrutura secundária, mas é capaz de extrair de uma arquivo no formato PDB a estrutura secundária a que pertencem os resíduos, baseado nas coordenadas atômicas presentes no arquivo.

   A lista de resíduos LISTA_RES.DAT, que no método MC-RNA é criada a partir da previsão das RNAs é alterada da seguinte maneira: a previsão das RNAs é substituída pela estrutura real da proteína *lida* pelo DSSP do arquivo PDB, e o grau de confiança da *previsão* é substituído por 1.0.

   A fração da seqüência da proteína $1j8b$, resíduos 47 ao 64 fica então:

```
. . . . . .  .
. . . . . .  .
. . . . . .  .
47 R B 1.0 B X
48 R B 1.0 B X
49 I B 1.0 B X
50 D B 1.0 B X
51 I B 1.0 B X
52 D C 1.0 C X
53 P A 1.0 A X
54 S A 1.0 A X
55 L A 1.0 A X
56 M A 1.0 A X
57 E A 1.0 A X
58 D C 1.0 C X
59 D C 1.0 C X
60 K A 1.0 A X
```

```
61 E A 1.0 A X
62 M A 1.0 A X
63 L A 1.0 A X
64 E A 1.0 A X
. . . . . .  .
. . . . . .  .
. . . . . .  .
```

o que garante que os ângulos serão escolhidos para a estrutura real em 100% dos casos.

2. Geração da amostra de estruturas

   Neste passo as N estruturas da amostra são geradas por MC, escolhendo os ângulos em regiões específicas do mapa de Ramachandran de acordo com a estrutura secundária real obtida pelo DSSP do arquivo PDB original.

   (a) **Parâmetros** O executável DIHEDRAL_CHANGE2 carrega os mesmos parâmetros do método MC-RNA: o arquivo EVA_2.DAT, o arquivo da proteína em análise no formato INT do Tinker, e o arquivo LISTA_RES.DAT modificado.

   (b) **Geração de mapas**

   Como no método MC-RNA, uma vez carregados com os dados e parametrizado, o executável gera mapas de Ramachandran específicos para cada par de resíduo e estrutura(s) secundária(s) do arquivo LISTA_RES.DAT. Ao contrário do MC-RNA, porém, não são gerados mapas para os grupos de resíduos não *previstos* pelo DSSP.

   No exemplo da linha do arquivo lista_res.DAT da proteína $1j8b$ correspondente ao resíduo 50 da proteína $1j8b$, a linha 50 D B 1.0 B X o executável criará apenas uma lista de ponteiros para Aspartatos pertencentes à folhas-β.

   (c) **Criação das estruturas com consulta às RNAs** Como no MC-RNA, criamos a partir dos ângulos dos mapas de Ramachandran reduzidos:

      i. Escolher randomicamente um par de ângulos diedrais do mapa de Ramachandran correspondente ao tipo de resíduo e ao tipo de estrutura secundária.

      ii. alterar os ângulos diedrais correspondentes na representação interna da cópia do arquivo INT, e salvar o arquivo em disco

   O procedimento descrito acima é repetido N vezes, gerando N arquivos INT com conformações diferentes. Independente do tamanho de N, garantimos que 100% das conormações geradas para a proteína $1j8b$ terão os ângulos do resíduo 50 escolhidos do mapa de ângulos específicos de Aspartatos que pertençam à folhas-*beta*.

Finda a etapa de geração de proteínas por MC-RNA, MC e MC-DSSP, dispomos de conjuntos de N arquivos INT, cada um representando uma conformação diferente da proteína em estudo. Porém, todos os arquivos forma gerados fazendo-se alterações em ângulos diedrais de um arquivo matriz, sem preocupações em efeitos

colaterais como proximidade excessiva entre átomos ou sobreposições decorrentes destas alterações. Na próxima seção descrevemos a fase dois do algoritmo, onde as estruturas são reacomodadas através de minimização de energia, e os resultados são clusterizados.

4.2.2 Fase 2: Minimização e Clusterização.

Nesta seção são explicados o processo de minimização e clusterização dos conjuntos de conformações obtidos na fase 1 do método. A segunda fase do método de simulação proposto constitue-se na minimização de energia das conformações geradas por MC, e posterior clusterização.

Cada amostra de conformações geradas por MC é constituída de N arquivos INT, cada um com a representação interna da proteína em estudo na forma de coordenadas. Como explicado na subseção 4.2.1 (Fase1: Geração de conformações), as coordenadas presentes no arquivo INT são distância a um átomo pré-determinado, e dois ângulos de ligação ou um ângulo de ligação e um ângulo diedral em relação á átomos predecessores. Como o método MC somente altera os ângulos diedrais, é necessário um método para realizar alterações nos demais ângulos e distâncias entre átomos representados no arquivo INT, de forma a chegar a uma representação mais realista em termos energéticos. Em outras palavras, cada vez que se modificam ângulos entre átomos de uma proteína, faz-se necessária uma acomodação de todos os demais ângulos das ligações atômicas da estrutura. De outra maneira não seria possível fazer qualquer tipo de comparação entre as conformações que levassem em conta a energia potencial, e a clusterização das conformações não traria informações úteis.

Fase 2, Parte 1: Minimização

Como na minimização das proteínas da lista EVA na fase de pré-processamento, para a minimização das conformações das proteínas pós método MC foi utilizado o método de descida de gradiente no espaço formado por ângulos diedrais. A ferramenta utilizada também foi a mesma: o MINIROT do pacote Tinker. Como explicado na subseção 4.2.1, a ferramenta MINIROT realiza minimização de energia em estruturas descritas por um arquivo INT, ou seja, minimiza a energia de uma proteína, dada um conformação inicial descrita através de ângulos de ligações entre átomos em um arquivo INT.

Ao contrário porém das minimizações realizadas no pré-processamento das proteínas da lista EVA, o critério *rms* foi mais elevado. Enquanto as minimizações das proteínas da lista EVA eram interrompidas quando a o gradiente de energia baixava de $1kcal/mole/\AA$, o critério de parada adotado para as minimizações de energia por descida de gradiente para as conformações pós MC foi de$10kcal/mole/\AA$. A adoção da interrupção precoce da minimização em $10kcal/mole/\AA$ adotada para as conformações pós MC se deve à dificuldade relativa da realização de descida de gradiente encontrada nestas estruturas. Como as proteínas da lista EVA já se encontram em conformações próximas à conformação nativa, o tempo gasto em minimização de energia é muito menor do que para as estruturas geradas por MC. Estas últimas podem se encontrar em posições muito diversas da conformação nativa, e até mesmo em estruturas inviáveis, exigindo grande quantidade de tempo para a estabilização em um mínimo local por descida de gradiente. Na verdade, as proteínas

geradas por MC precisaram de mais tempo em média para serem minimizadas com critério de parada em $10kcal/mole/\AA$ do que as proteínas da lista EVA com o critério de parada mais refinado de $10kcal/mole/\AA$. A título de comparação, enquanto as minimizações das proteínas da lista EVA levaram em média108 passos de descida de gradiente para atingir o gradiente de $1kcal/mole/\AA$ e serem interrompidas, as médias para as estruturas geradas estão entre 110 e 3683 passos por estrutura (Tabela 4.2). Se levarmos em consideração que o número médio de resíduos das proteínas da lista EVA é de aproximadamente 162 resíduos por proteína, e que todas as seqüências testadas são menores que a média, fica mais claro como a desorganização estrutural de uma proteína afeta os tempos de minimização.

Proteína (número de resíduos)	Método	número médio de passos	tempo médio (minutos)
$1i74$ (108)	MC	1153	14
	MC-RNA	922	10
	MC-DSSP	117	$1,6$
$1kkg$ (108)	MC	1611	25
	MC-RNA	1718	20
	MC-DSSP	453	$5,4$
$1g7d$ (77)	MC	794	5
	MC-RNA	705	$4,5$
	MC-DSSP	208	$1,3$
$1j8b$ (92)	MC	870	$7,4$
	MC-RNA	616	4
	MC-DSSP	110	$0,51$

TABELA 4.2 – A Tabela mostra a média de passos de minimização e de tempo de simulação por conformação gerada por três métodos: MC. MC-RNA e MC-DSSP. A quantidade de informação aumenta no sentido MC->MC-RNA->MC-DSSP, e o tempo de minimização tende a diminuir no mesmo sentido. Isto é um indício de que quanto maior a informação disponível sobre a estrutura secundária, mais próximas à conformação nativa estarão as conformações geradas pelo MC. Os tempos foram obtidos em computadores Intel(R) Xeon(TM) CPU $2.40GHz$, com $1MB$ ou $2MB$ de memória e dedicação exclusiva.

Após a minimização, são tiradas 4 medidas para cada uma das conformações obtidas: energia, superfície acessível total, superfície acessível das cadeias laterais, e distância RMS com a conformação final. A medida da energia nada mais é do que a energia final obtida no processo de minimização com a ferramenta MINIROT. A superfície acessível é a superfície formada pela rolagem de uma esfera de $1,4\AA$ por sobre as partes externas da proteína às quais a esfera consiga ter acesso. A superfície acessível das cadeias laterais é a contribuição das cadeias laterais dos resíduos hidrofóbicos à superfície total, e a distância RMS é a raiz quadrada da soma dos quadrados das distâncias entre átomos da conformação obtida e da conformação nativa conhecida.

As superfícies acessíveis da proteína e de suas cadeias laterais são obtidas através do seguinte procedimento:

1. **INT_2->PDB_2**

   Cada conformação INT gerada pelo MC e posteriormente minimizada é armazenada em um arquivo INT_2. A partir destes arquivos e com as ferramentas do Tinker INTXYZ e XYZPDB geramos arquivos no formato *Protein Data Bank* com o prefixo PDB_2, contendo as proteínas minimizadas descritas através de seus átomos, tipo de resíduo a que pertencem e coordenadas cartesianas.

2. **Cálculo da superfície acessível total** A ferramenta SPACEFILL do Tinker calcula entre outros a superfície acessível de moléculas. Para tanto a ferramenta utiliza uma versão modificada da descrição analítica original da superfície de moléculas de Connolly [Con83]. A superfície é particionada em seus componentes geométricos e decomposta em contribuições convexas para cada átomo individualmente. O executável SPACEFILL recebe como entrada um arquivo PDB_2 e calcula a superfície total e a contribuição de cada átomo para a superfície total.

3. **Cálculo da superfície acessível das cadeias laterais** O interesse na responsabilidade das cadeias laterais na superfície total é a identificação da posição de resíduos hidrofóbicos: se estão corretamente voltados para o interior da proteína, a contribuição de suas cadeias laterais para a formação da superfície acessível á próxima a zero. Como resíduos hidrofóbicos voltados para fora não são facilmente encontrados na natureza, superfícies acessíveis de cadeias laterais deste tipo de resíduo indicam conformações pouco prováveis.

   Para o cálculo da contribuição das cadeias laterais à área total são considerados neste trabalho os resíduos hidrofóbicos ALA, CYS, CYH, CSS, CYX, PHE, ILE, LEU, VAL, e TRP. Como esta medida tem fins comparativos, as contribuições dos átomos de carbono das cadeias laterais é suficiente, e os demais átomos são excluídos do cálculo para reduzir o custo computacional.

   Para obtermos a contribuição dos átomos da cadeia lateral à superfície acessível total, é preciso identificar os átomos que pertencem à essas cadeias. A ferramenta ANALYZE do DSTK recebe como entrada o arquivo PDB da conformação em análise e a saída do arquivo SPACEFILL. O ANALYZE então percorre o arquivo PDB_2, identifica todos os átomos pertencentes aos resíduos hidrofóbicos e cria uma lista com os índices de todos os átomos de carbono destes resíduos que não sejam o carbono α. Então identifica na saída do SPACEFILL os carbonos da lista e as suas respectivas contribuições convexas à área total. A saída da ferramenta ANALYZE é a superfície total calculada pelo SPACEFILL e a soma das contribuições de todos os carbonos pertencentes às cadeias laterais dos resíduos hidrofóbicos

Para o próximo e último passo da metodologia, a clusterização, são necessários ainda os ângulos diedrais das proteínas minimizadas. Estes ângulos são extraídos dos arquivos INT_2 através da ferramenta GET_ANGLE_RES do DSTK, e são unificados pela ferramenta ANALYZE em um arquivo contendo o nome da proteína, o índice da sua conformação obtida por MC com ou sem RNA, um par de Ângulos diedrais para cada resíduo da seqüência, a energia potencial da conformação, a superfície total acessível, a superfície acessível das cadeias laterais hidrofóbicas, e a distância da estrutura para a conformação nativa conhecida da proteína.

A seguir é explicado como foi feita a clusterização nestes dados e quais os critérios utilizados para a escolha de variáveis do processo.

Fase 2, Parte 2: Clusterização

Para a análise das estruturas resultantes das simulações MC foi utilizado o algoritmo de *clusterização* particional *K-means* em função da sua baixa complexidade.

Até aqui a implementação do método MC-RNA foi descrita utilizando a variação de energia como parâmetro. No entanto, as diferenças de energia entre a conformação nativa que se quer atingir e mínimos locais é muito sutil. Muitas vezes estruturas completamente diversas podem se encontrar no mesmo patamar de energia potencial. Para melhorar a performance do modelo, é interessante utilizar não apenas a energia, mas outras propriedades das proteínas que as possam diferenciar de maneira mais eficaz.

A função objetivo de qualquer algoritmo de busca deve se basear em parâmetros que indiquem o grau de proximidade entre a conformação gerada e a provável estrutura tridimensional da proteína. Os parâmetros de comparação tipicamente utilizados para comparar duas conformações tridimensionais são: Erro Médio Quadrado entre os átomos, energia potencial, hidrofobicidade, compactação e até semelhança visual. A minimização da energia potencial e a maximização da compactação não leva em conta efeitos como hidrofobicidade, que pode indicar estruturas menos compactas e de maior energia como sendo as conformações preferenciais na natureza. Afim de balancear as deficiências de cada parâmetro, propomos utilizar uma função objetivo mista baseada em 2 parâmetros: compactação (medida da superfície exposta ao solvente) e energia potencial.

Se assumirmos que a compactação de uma proteína é proporcional à área acessível ao solvente, podemos utilizar esta informação extra para aceitar ou rejeitar uma transição entre estados de mesma energia. Para o cálculo da compactação foi utilizada a implementação do algoritmo de Connolly [Con83] do pacote de mecânica molecular Tinker. Este algoritmo calcula a parcela da superfície de Van der Waal de uma molécula que é acessível ao solvente. O processo consiste em modelar a molécula do solvente por uma esfera, e rolar esta esfera sobre a molécula. A superfície gerada desta maneira é composta de partes de esferas e toros, e portanto tem sua área passível de ser calculada analiticamente.

Para reduzir a complexidade, a clusterização por *K-means* foi realizada sobre apenas alguns ângulos diedrais das proteínas. Mais precisamente foram escolhidos ângulos de resíduos sobre os quais as RNAs não conseguem fazer um predição satisfatória. Como o método *K-means* é sensível às condições iniciais, o método é repetido algumas vezes para se obter o conjunto de clusters mais provável. Terminada a clusterização, os clusters são examinados quanto à presença de conformações com as seguintes características: baixa energia, pequena superfície total, pequena superfície hidrofóbica e pequeno RMS com a original. De posse destas informações é possível determinar quais clusters contem as conformações mais condizentes com a conformação nativa, e determinar as características comuns das conformações deste cluster. Deste modo conseguimos isolar os ângulos diedrais que nos aproximam mais da conformação nativa da proteína em estudo.

No próximo Capítulo, conforme mostramos os resultados do método, exemplificamos a escolha de ângulos e de clusters descrita acima.

Capítulo 5

Resultados

Neste Capítulo descrevemos os resultados obtidos para as proteínas utilizadas nos experimentos: $1i74$ (domínio 2), $1kkg$, $1g7d$ domínio C-terminal e $1j8b$. Para cada proteína foram realizados um experimento com o algoritmo MC-RNA e um experimento com cada um dos algoritmos de controle MC-DSSP e MC, totalizando 12 simulações. Cada simulação gerou uma amostra contendo 1000 conformações, e todas sofreram minimização de energia conforme as regras explicadas no Capítulo da metodologia. Em todos os experimentos parte das estruturas não pôde ser minimizada (o executável MINIROT do Tinker interrompia a minimização com erro) ou acabou em mínimos locais muito altos, e foi eliminada. As eliminações de configurações em mínimos locais altos foi necessária para que pudéssemos trabalhar com médias significativas, e portanto o critério utilizado para a eliminação foi a aproximação entre médias e medianas. Este critério se mostrou válido se observarmos que as distribuições têm desvios padrão compatíveis com as médias para distribuições que aproximam a normal. Como a clusterização por *K-means* é dependente dos pontos escolhidos como clusters iniciais, e o número de clusters deve ser determinado pelo usuário, foram necessários exaustivos testes com o algoritmo de clusterização *K-means*, variando-se o número de número de *clusters* e verificando o número de inicializações aleatórias necessárias para a repetição de configurações de ângulos entre rodadas. Após experimentos com as quatro proteínas, o número de 5 *clusters*, com 5 inicializações aleatórias cada um, foi considerado o conjunto de parâmetros ótimo, e utilizado em todos os experimentos deste trabalho. Nas quatro seções a seguir são apresentados os resultados para as quatro proteínas, uma por sessão. Em cada seção sempre é mencionado o número total de conformações efetivamente utilizado nas clusterizações, que é o número de proteínas que tiveram a sua energia minimizada com sucesso. Há também uma tabela em cada seção apresentando uma análise das 5 rodadas e os 5 clusters obtidos em cada rodada.

5.1 Proteína $1j8b$

A proteína $1j8b$ é composta por 92 resíduos, uma folha-β central composta por 3 segmentos, duas α-hélices longas e uma curta. A Tabela abaixo mostra a seqüência de resíduos da proteína na primeira linha, acompanhada da estrutura secundária da conformação nativa (lida pelo DSSP) e pela previsão da estrutura secundária feita pelas RNAs respectivamente na segunda e terceira linhas.

```
01-60: LGGLMKQAQQMQEKMQKMQEEIAQLEVTGESGAGLVKITINGAHNCRRIDIDPSLMEDDK
DSSP : CCCAAAAAAAAAAAAAAAAAATTTBBBBBBAAATBBBBBBTTCCBBBBBBCAAAAATCA
RNA  : C..AAAAAAAAAAAAAAAAAAA.BBB..CCC.BBBBBB.CC..BBBBB.C.....CCA

61-92: EMLEDLIAAAFNDAVRRAEELQKEKMASVTAG
DSSP : AAAAAAAAAAAAAAAAAAAAAAAAAAAAATCC
RNA  : AAAAAAAAAAAAAAAAAAAAAAAAA....CC
```

Os pontos na linha da previsão das RNAs indica que para aquele resíduo a RNA não conseguiu fazer um previsão com mais de 50% de confiança. Para a clusterização das conformações obtidas a partir da seqüência da proteína $1j8b$ foram selecionados os ângulos dos resíduos identificados por estes pontos, descartados os próximos às extremidades. Ou seja, os ângulos diedrais dos resíduos para os quais não se tem previsão através de RNAs.

Para a proteína $1j8b$ temos então 26 ângulos de 13 resíduos selecionados. Pode-se verificar no alinhamento da estrutura secundária real lida pelo DSSP e da estrutura secundária predita pelas RNAs que estas obtiveram sucesso em identificar as duas maiores α-hélices e a existência de três segmentos de folha-β. A pequena α-hélice formada pelo segmento $PSLME$ (resíduos 53 a 57) porém não foi identificada.

Para melhor entender a influência da informação sobre a estrutura secundária no desempenho do método Monte Carlo, a Tabela 5.1 lista uma série de valores comparativos para a Energia. As primeiras 3 colunas mostram respectivamente a quantidade de conformações descartadas devido à impossibilidade de minimização por descida de gradiente, o número de estruturas descartadas por energia demasiado alta, e o número de conformações efetivamente utilizado na fase final de clusterização. Os demais valores são medidas de energia: energia mínima, máxima, média, mediana e desvio padrão da amostra de N conformações.

Método	# $E = \infty$	# E alta	N	E min	E max	E média	E mediana	DP
MC	95	128	777	1415	2618	1723	1713	133
MC-RNA	67	86	847	1427	2468	1672	1666	108
MC-DSSP	19	20	961	1380	1919	1602	1601	51

TABELA 5.1 – Medidas de energia das amostras de conformações geradas para a seqüência da proteína $1j8b$. As três primeiras colunas contém respectivamente o número de conformações impossíveis de minimizar, o número de conformações cortadas da cauda à direita da distribuição e o número N de conformações destinadas à clusterização. As cinco colunas restantes são as menores e maiores energias da amostra de N conformações, a média, a mediana e o desvio padrão. Todos os valores se referem às conformações após a fase de minimização por descida de gradiente.

Os métodos estão ordenados de modo que a quantidade de informação sobre a estrutura secundária fornecida ao MC cresce de cima para baixo: MC tem zero informação e MC-DSSP tem 100% de informação. Desta forma, de acordo com a Tabela 5.1 podemos inferir que quanto mais informação sobre a estrutura secundária, menor a ocorrência de conformações impossíveis ou improváveis, menor a energia média e menor a variância das conformações geradas pelo método MC. A diminuição da energia média das conformações minimizadas e da variância da distribuição desta energia pode ser visualizado nos gráficos da Figura 5.1.

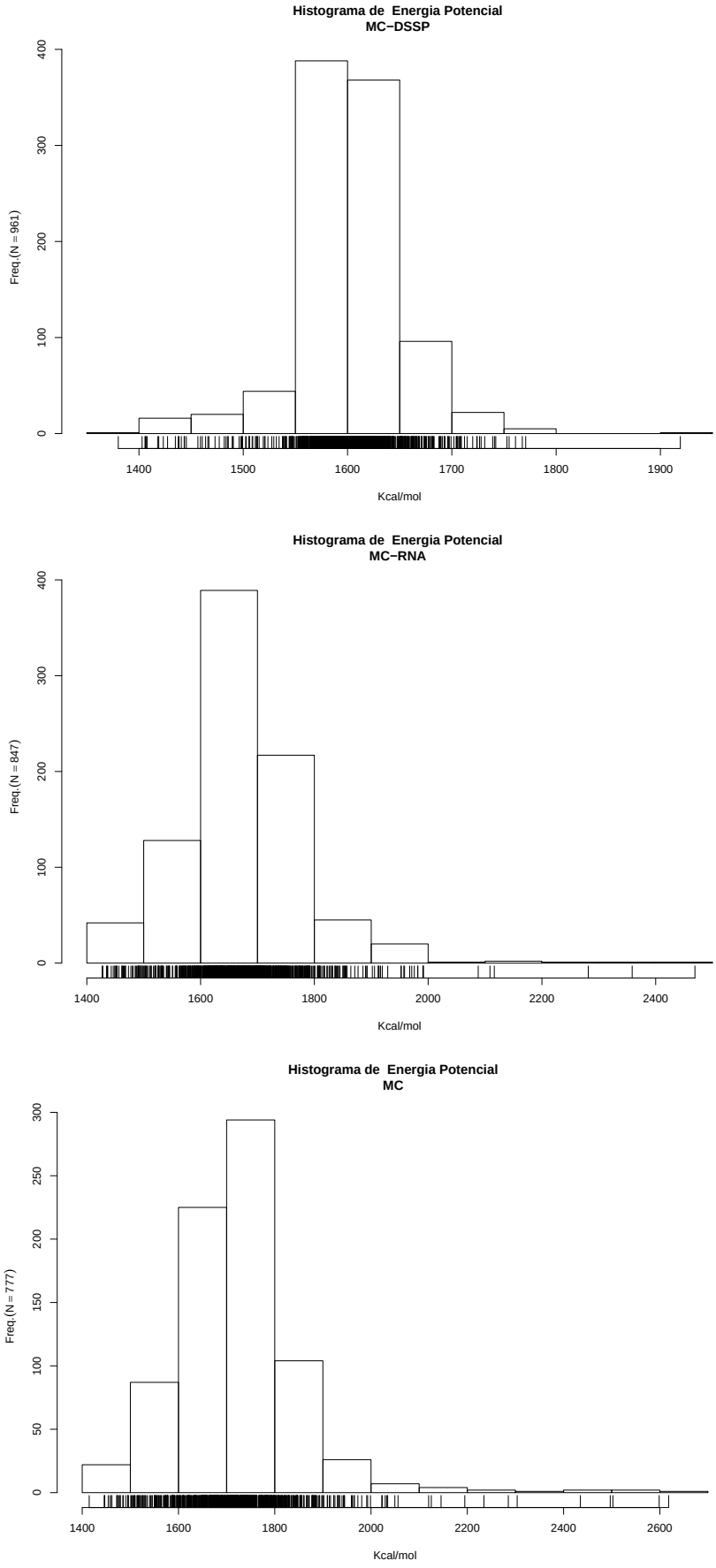


FIGURA 5.1 – Distribuição de energia potencial das amostras para as conformações geradas pelos métodos MC-DSSP (topo), MC-RNA e MC, para a proteína $1j8b$. À medida que evoluímos do método MC para o MC-DSSP a quantidade de informação aumenta, e a média e a variância da distribuição de energia diminui.

5.1.1 Resultados da clusterização para $1j8b$

Para cada conjunto de conformações gerado pelos métodos Monte Carlo foi realizada clusterização por *K-means*. Para cada método Monte Carlo, o conjunto de conformações sobre o qual se aplicou a clusterização foi extraido do total de 1000 conformações gerada pelo método, excluídas as proteínas que não conseguiram ser minimizadas e as que encerram a minimização de energia em mínimos locais relativamente altos. Para os métodos Monte Carlo (MC), Monte Carlo com previsão de estrutura secundária por RNAs (MC-RNA) e Monte Carlo com informação sobre estrutura secundária real (MC-DSSP), os conjuntos contém 777, 847 e 961 conformações respectivamente.

Cada conformação dos conjuntos contém $2n$ ângulos diedrais, e valores de energia, superfície total exposta ao solvente, parcela da superfície referente à elementos hidrofóbicos, e distância RMS em relação à conformação nativa. Para reduzir a complexidade dos dados a serem clusterizados, foram escolhidos os ângulos dos resíduos para os quais não há previsão de estrutura secundária pela RNA. Esta heurística parte da presunção que, se a previsão das RNAs para os demais resíduos fosse perfeita, poderíamos descrever a estrutura tridimensional apenas com a definição dos ângulos de resíduos intermediários. Ou seja, se assumimos que as RNAs foram capazes de prever a estrutura secundária de certos segmentos da proteína, podemos descrever a estrutura terciária dizendo como estes segmentos se dobram uns sobre os outros. Esta informação é extraída dos ângulos diedrais dos resíduos intermediários entre uma e outra estrutura secundária prevista. Seguindo o mesmo raciocínio, excluímos os ângulos não previstos de resíduos nas extremidades da proteína, quando ocorrem, visto que agregam pouca informação sobre a estrutura tridimensional e aumentam a complexidade da clusterização.

O número de clusters escolhido para esta e para as demais proteínas foi, conforme explicado na introdução deste capítulo, 5 Após a clusterização, os clusters foram analisados quanto à concentração de conformações com menor energia, superfícies de exposição total e hidrofóbica, e a distância RMS com a conformação nativa. Deste exame se constatou que quando a clusterização consegue reunir simultaneamente em um dos cinco clusters as maiores proporções de conformações com menor energia, com menor superfície de exposição ao solvente total, e com menor superfície de exposição hidrofóbica, este cluster reúne também a maior proporção de conformações com menor distância RMS em relação à conformação nativa da proteína. Em outras palavras, se a concentração de conformações com menores valores para as três variáveis pode ser encontrado em um mesmo cluster, é grande a probabilidade de este cluster estar mais próximo à conformação nativa. O contrário é verdadeiro: se as concentrações de conformações com menores valores para as três variáveis se encontram dispersos entre os clusters, não podemos dizer qual cluster se aproxima mais da conformação nativa.

	MC				MC-RNA				MC-DSSP			
Cl	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS
Rodada no. 1												
1	**27**	**27**	**25**	**26**	17	16	21	15	9	11	10	9
2	7	10	9	10	17	16	15	18	**41**	**40**	**42**	**41**
3	19	19	23	19	**27**	**20**	**23**	**22**	5	5	2	5
4	14	12	10	11	16	19	16	18	20	19	16	16
5	10	9	10	11	7	13	9	11	21	21	26	25
Rodada no. 2												
1	**25**	**23**	**22**	**21**	**29**	23	**28**	24	7	9	9	8
2	11	13	12	16	23	**29**	21	26	5	5	3	5
3	21	20	19	19	9	9	9	7	34	35	33	33
4	13	11	16	11	12	15	16	12	**40**	**37**	**42**	**40**
5	7	10	8	10	11	8	10	15	10	10	9	10
Rodada no. 3												
1	10	13	12	15	13	13	16	14	5	5	2	5
2	21	16	17	18	10	10	10	6	5	5	3	5
3	6	10	8	10	**23**	**30**	**22**	27	39	**40**	40	38
4	16	14	**19**	16	**23**	13	15	22	**40**	37	**42**	40
5	**24**	**24**	21	18	15	18	21	15	7	9	9	8
Rodada no. 4												
1	20	17	**23**	14	9	8	10	6	20	19	16	16
2	16	13	13	19	21	17	19	18	8	9	9	8
3	11	13	12	16	**23**	**30**	**21**	**25**	**46**	**45**	**44**	**46**
4	6	10	8	10	16	14	18	18	17	19	20	23
5	**24**	**24**	21	18	15	15	16	17	5	4	7	3
Rodada no. 5												
1	**25**	**26**	**25**	**22**	10	12	12	6	21	21	26	25
2	21	19	16	19	14	19	19	15	7	6	10	6
3	6	10	8	10	15	15	19	22	9	11	10	9
4	9	8	9	10	16	11	12	16	20	19	16	16
5	16	14	19	16	**29**	**27**	**22**	**25**	**39**	**39**	**34**	**40**

TABELA 5.2 – Clusters das conformações da proteína $1j8b$. Para cada método há 5 colunas: o número do cluster, e as freqüências de presença dentro de cada cluster de conformações com menor Energia, menor Superfície Total, menor Superfície Hidrofóbica e menor distância RMS. As rodadas de 1 a 5 referem-se a cinco inicializações com sementes aleatórias diferentes. Os maiores valores entre os clusters de cada rodada estão grifados, e quando um cluster contém simultaneamente o maior número de conformações com baixos valores para as3 medidas, o valor RMS também é grifado.

Na Tabela 5.2 estão dispostos dados sobre as conformações que apresentam menores valores em uma de três medidas: energia, superfície total ou superfície hidrostática, e a distribuição destas conformações entre os clusters. A tabela é constituída pela união de 15 tabelas menores, cada uma representando os 5 clusters de uma rodada ou inicialização específica de um dos métodos MC, MC-RNA e MC-DSSP. Para cada método temos 5 colunas: a primeira representa o número do cluster (de 1 a 5), as três colunas seguintes contém a quantidade de ocorrências de conformações com valores pequenos para as 3 medidas energia (E), superfície total (ST) e superfície hidrofóbica (SH), e a quinta e última coluna a ocorrência de conformações com baixas distâncias RMS em relação à conformação nativa.

Para entendermos melhor a Tabela 5.2 usaremos um exemplo de sua construção. A aplicação do método MC-RNA para a geração de 1000 conformações da proteína $1j8b$ resultou, após o processo de minimização, em uma amostra contendo 847 conformações minimizadas. Em seguida, a amostra sofreu 5 rodadas de clusterização por *k-means*, cada rodada com uma semente aleatória diferente. Findo o processo, temos cinco conjuntos de clusters para cada um dos três métodos MC-DSSP, MC-RNA e MC. Feito isto, foram identificadas 84 conformações (ou aproximadamente 10%) do total da amostra com os menores valores para cada medida: energia, superfície total, superfície hidrofóbica e para a distância RMS. Por fim, a distribuição deste subconjunto de conformações com os menores valores, ou valores mínimos, foi verificada contando quantas ocorreram em cada cluster.

Agora podemos verificar, olhando para a Tabela 5.2, linha 3, coluna $MC-RNA$, que na rodada número 1 a clusterização produziu 5 clusters, e que o cluster de número 3 contém concomitantemente a maior *concentração* de conformações com valores de energia, superfície total, superfície hidrofóbica e distância RMS próximos aos mínimos da amostra para estas medidas. Para cada conjunto de 5 clusters a maior ocorrência de conformações de energia e de medidas de superfície mínimas é assinalada em negrito. Observando os demais conjuntos de 5 clusters das outras rodadas e dos outros métodos, chega-se à conclusão de que quando ocorre a concomitância de concentrações de mínimos como no cluster assinalado acima, ocorre neste mesmo cluster a maior concentração de distâncias RMS próximas ao mínimo. O inverso nos leva a concluir que quando as concentrações de valores próximos aos mínimos para as medidas de energia e áreas estão dispersos entre dois ou mais clusters, nada se pode afirmar quanto à localização do cluster com maior concentração de conformações de distância RMS mínima.

Por exemplo, a segunda rodada de clusterização da amostra do método MC-RNA resultou em 29 conformações entre as 84 de menor energia no cluster 1, e 23 no cluster 2. Mas como o cluster 2 contém 29 das 84 conformações mais compactas (menor superfície total exposta ao solvente) contra 23 do cluster 1, a distribuição de configurações com distâncias RMS em relação à conformação nativa é quase igual entre os dois. De fato, o cluster 1 perde neste quesito para o cluster 2 por 24 a 26. Em uma situação real de estudo de uma proteína nova, não temos acesso à distância RMS com a estrutura nativa, simplesmente porque não conhecemos a estrutura nativa da proteína. Neste caso, para tentar identificar qual o cluster que mais se aproxima da estrutura nativa poderíamos escolher o critério de concentração simultânea de mínimos de energia, superfícies total e hidrofóbicas. Por isto identificamos com negrito a quantidade de conformações de mínimos RMS pertencentes a um cluster apenas quando este mesmo cluster contiver simultaneamente concentrações altas em relação aos demais clusters de valores mínimos de energia, superfície total e superfície hidrofóbica. Outro fato que podemos depreender dos valores da tabela é que quanto maior for a concentração de conformações com valores mínimos para as três medidas em um mesmo cluster, maior será a concentração de conformações com distâncias RMS mínimas em relação à conformação nativa.

Se olharmos novamente para a Tabela 5.2, agora comparando os métodos, podemos verificar que, a medida que avançamos da esquerda para a direita indo do método MC para o MC-DSSP, o número de ocorrências de clusters com concentração de mínimos para as três medidas aumenta, e também aumenta a diferença relativa entre as concentrações nestes clusters em relação aos demais.

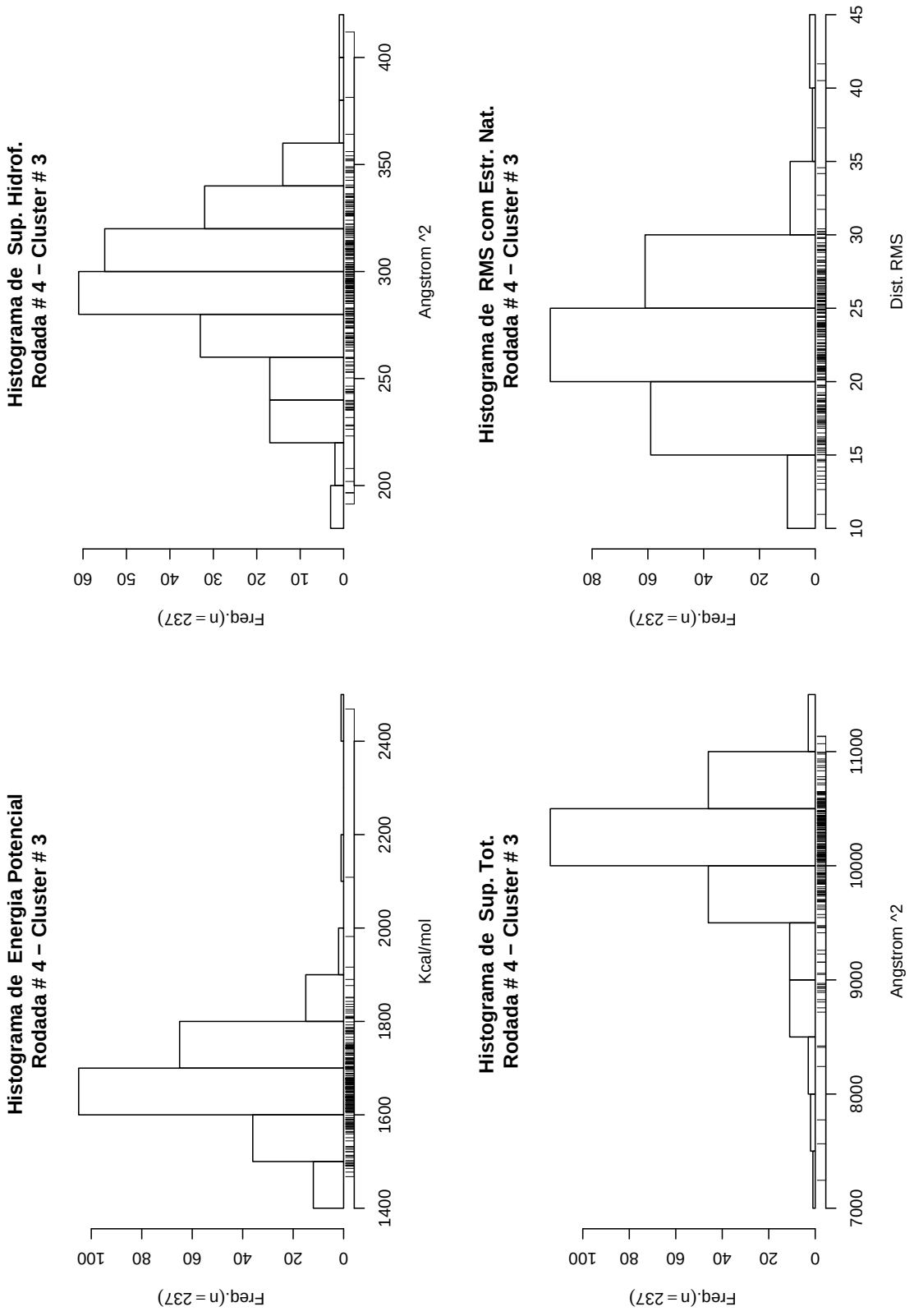
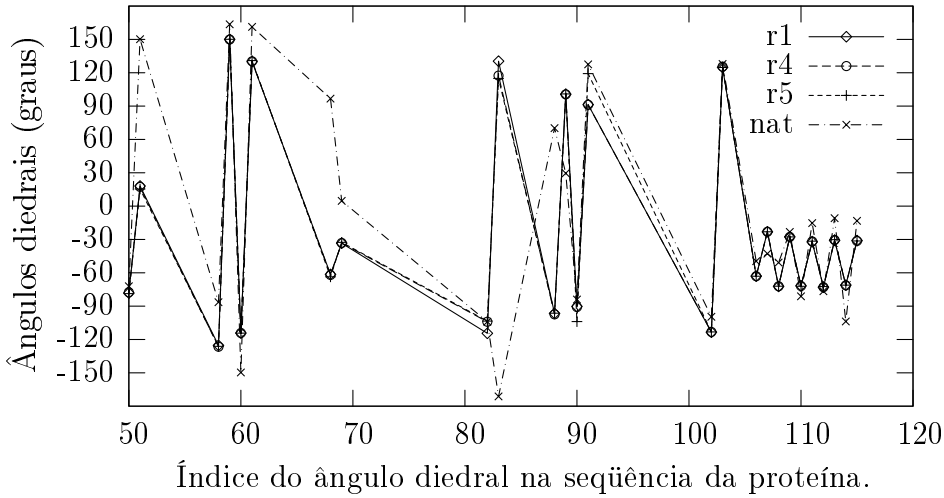


FIGURA 5.2 – Distribuição da energia, superfície total, superfície hidrofóbica e distância RMS à conformação nativa para o cluster 3 da quarta rodada de clusterização para a proteína $1j8b$, conformações geradas por MC-RNA.

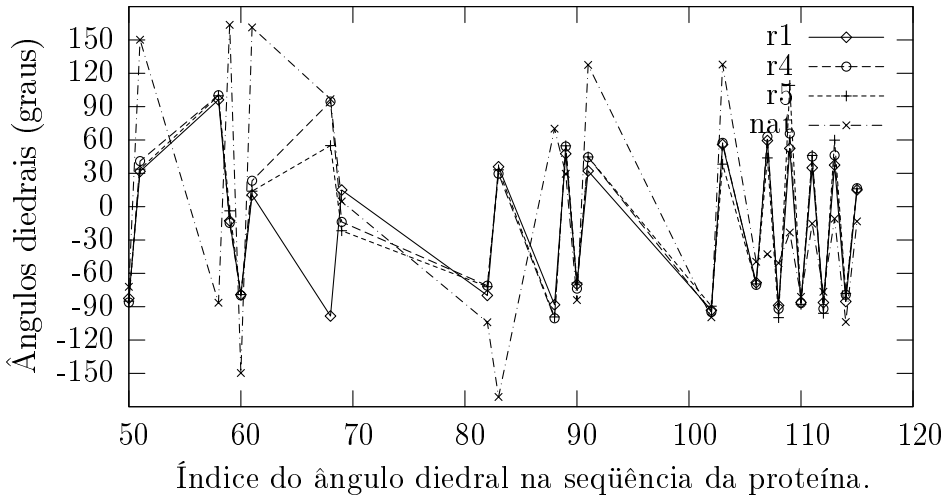
Este fato indica que o aumento de informação que ocorre do método MC para o MC-RNA, e deste para o MC-DSSP acarreta aumento no número de conformações geradas mais próximas ou semelhantes à conformação nativa, e que isto é identificado pelo método de clusterização *k-means* que tende a concentrar estas configurações em um único cluster.

Os histogramas da Figura 5.2 representam as distribuições de probabilidade do cluster 3 da quarta rodada de clusterização, da amostra obtida com o método MC-RNA. A forma de sino das distribuições e a variância pequena indicam que o cluster realmente identifica um padrão de conformações com características comuns. Podemos observar também a concentração proporcionalmente maior de conformações de baixa energia em relação a distribuição do total das conformações mostrada anteriormente na Figura 5.1. Se estivéssemos utilizando o método para de fato inferir algo sobre a estrutura nativa da proteína, escolheríamos este cluster e os clusters 3 da rodada 1 e 5 da rodada 5. Todos os três são cluster com concentração de mínimos, e não por coincidência se plotarmos os vetores de ângulos que os representam verificamos que são bem semelhantes entre si. Não apenas isto, mas são próximos ao vetor feito com os ângulos retirados da estrutura nativa da proteína. Podemos observar esta proximidade nos três gráficos, novamente um para cada método, da Figura 5.3. Nestes gráficos os ângulos dos clusters com concentração simultânea dos mínimos dos parâmetros energia e superfícies são plotados como pontos unidos por linhas para facilitar a visualização. No eixo horizontal dos três gráficos estão os índices dos ângulos diedrais utilizados na clusterização. Neste caso são os ângulos dos resíduos de número 25, 29, 30, 34, 41, 44, 45, 51, 53, 54, 55, 56 e 57 na seqüência da proteína, e que foram escolhidos por não terem estrutura prevista pelas RNAs. Como cada resíduo têm um par de ângulos, os índices dos ângulos representados no gráfico começam com 50 e 51, 58 e 59, até os últimos ângulos de número 114 e 115. Nos gráficos temos ainda, além dos clusters do MC-DSSP, pontos unidos pela linha tracejada mais fraca que representam os ângulos da estrutura nativa. Como podemos observar, a semelhança entre os ângulos da estrutura nativa e do método MC-DSSP é significativa. Observa-se também que para o MC-DSSP e para o MC-RNA, os clusters que mais se aproximam da estrutura nativa (cluster 3 da rodada $r4$ para ambos MC-DSSP e MC-RNA, e cluster 1 da rodada $r2$ para o MC) são os clusters com maior concentração de mínimos entre as rodadas para os métodos MC-DSSP e MC-RNA, mas isto não se observa com o método MC. A Tabela 5.3 mostra, para cada um dos três métodos, para cada uma das cinco rodadas de clusterização, os ângulos centrais dos clusters com maior concentração de mínimos para os parâmetros energia, superfície total e superfície hidrofóbica. Para facilitar a comparação, os ângulos da conformação nativa estão replicados para cada método. Se observarmos os ângulos de índice 50 a 83 (correspondentes aos resíduos de índice 25, 29, 30, 34 e 41) podemos verificar que o método MC tem maior dificuldade em convergir para um cluster do que o método MC-RNA. Os valores dos ângulos dos resíduos pertencentes ao outro extremo da proteína por sua vez mostram que o MC-DSSP formou um cluster mais próximo da conformação nativa do que o método MC-RNA, como era de se esperar.

MC-DSSP: Ângulos da estrutura nativa e dos clusters das rodadas1, 4 e 5 (clusters 2, 3 e 5)



MC-RNA: Ângulos da estrutura nativa e dos clusters das rodadas1, 4 e 5 (clusters 3, 3 e 5)



MC: Ângulos da estrutura nativa e dos clusters das rodadas1, 2 e 5 (clusters 1, 1 e 1)

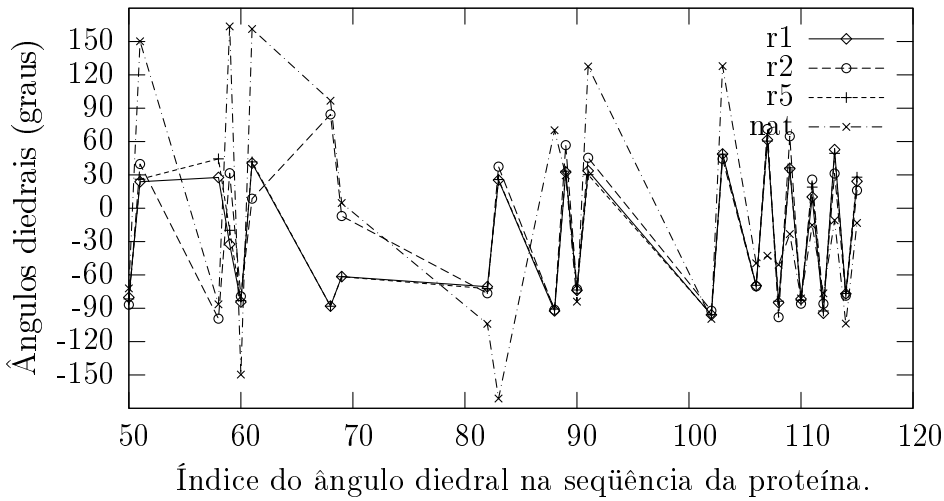


FIGURA 5.3 – Os gráficos mostram os ângulos da estrutura nativa e dos clusters que concomitantemente tem a maior concentração de estruturas com menor energia, superfície total e superfície hidrofóbica exposta ao solvente, para3 rodadas de clusterização para cada um dos 3 métodos, para a seqüência da proteína $1j8b$.

rodada	cluster	Ângulos Diedrais centrais dos *clusters*: método MC-DSSP												
Índices Φ		50	58	60	68	82	88	90	102	106	108	110	112	114
estr. nat		-72	-86	-149	96	-104	70	-84	-99	-49	-50	-81	-76	-103
1	2	-78	-125	-114	-61	-114	-96	-90	-113	-62	-72	-71	-72	-71
2	4	-84	5	-114	-59	-105	-94	-89	-112	-63	-72	-71	-72	-70
3	3	-77	-125	-114	-62	-114	-95	-89	-113	-62	-72	-71	-72	-70
3	4	-84	3	-114	-59	-105	-93	-89	-112	-63	-72	-71	-72	-70
4	3	-77	-126	-114	-61	-104	-97	-90	-113	-63	-72	-71	-72	-71
5	5	-78	-126	-114	-63	-103	-96	-103	-113	-63	-72	-72	-72	-71
Índices Ψ		51	59	61	69	83	89	91	103	107	109	111	113	115
estr. nat		150	163	161	4	-171	29	127	127	-42	-23	-15	-10	-13
1	2	18	149	129	-33	130	100	91	124	-23	-27	-32	-30	-30
2	4	17	-156	133	-32	119	98	100	127	-21	-27	-33	-30	-32
3	3	18	149	130	-33	129	101	106	124	-23	-27	-31	-30	-31
3	4	18	-155	133	-32	118	98	98	127	-21	-27	-33	-31	-32
4	3	17	149	130	-33	117	100	91	125	-23	-27	-31	-30	-31
5	5	15	150	130	-32	114	100	119	124	-23	-27	-31	-30	-31

rodada	cluster	Ângulos Diedrais centrais dos *clusters*: método MC-RNA												
Índices Φ		50	58	60	68	82	88	90	102	106	108	110	112	114
estr. nat		-72	-86	-149	96	-104	70	-84	-99	-49	-50	-81	-76	-103
1	3	-82	96	-79	-98	-79	-88	-69	-93	-68	-88	-86	-86	-85
2	1	-83	92	-78	-97	-80	-90	-70	-92	-68	-89	-86	-87	-87
2	2	-85	99	-79	96	-69	-100	-73	-94	-70	-91	-87	-91	-79
3	3	-86	99	-79	96	-69	-99	-72	-94	-70	-90	-87	-91	-79
4	3	-86	100	-79	94	-71	-100	-73	-94	-70	-91	-86	-91	-79
5	5	-84	99	-79	54	-71	-99	-68	-89	-69	-99	-87	-95	-78
Índices Ψ		51	59	61	69	83	89	91	103	107	109	111	113	115
estr. nat		150	163	161	4	-171	29	127	127	-42	-23	-15	-10	-13
1	3	31	-12	10	14	35	47	32	55	59	52	35	37	15
2	1	31	-14	15	16	35	50	31	57	61	49	37	37	18
2	2	40	-11	22	-15	29	53	42	56	62	63	44	44	16
3	3	41	-11	22	-13	29	53	42	56	62	67	43	44	18
4	3	41	-14	23	-13	29	54	44	57	62	65	45	46	16
5	5	33	-3	13	-21	33	54	44	38	43	109	46	59	16

rodada	cluster	Ângulos Diedrais centrais dos *clusters*: método MC												
Índices Φ		50	58	60	68	82	88	90	102	106	108	110	112	114
estr. nat		-72	-86	-149	96	-104	70	-84	-99	-49	-50	-81	-76	-103
1	1	-80	27	-84	-88	-70	-92	-73	-95	-69	-84	-81	-94	-76
2	1	-86	-99	-79	84	-76	-91	-72	-92	-70	-98	-85	-85	-78
3	5	-87	-93	-79	101	-76	-86	-70	-93	-70	-98	-84	-86	-78
4	5	-87	-93	-79	101	-75	-86	-69	-93	-70	-98	-85	-86	-78
5	1	-81	44	-83	-88	-72	-90	-73	-95	-69	-83	-82	-92	-76
Índices Ψ		51	59	61	69	83	89	91	103	107	109	111	113	115
estr. nat		150	163	161	4	-171	29	127	127	-42	-23	-15	-10	-13
1	1	23	-32	40	-61	25	32	33	48	61	35	10	52	24
2	1	39	31	8	-7	37	56	45	44	71	64	25	31	16
3	5	37	16	16	-8	42	52	40	47	64	59	26	31	17
4	5	37	16	15	-8	41	53	41	48	64	59	25	32	17
5	1	26	-19	41	-61	25	33	30	48	60	36	19	49	27

TABELA 5.3 – A tabela mostra os ângulos dos cluster que concomitantemente têm a maior concentração de estruturas com menor energia, superfície total e superfície hidrofóbica exposta ao solvente, para cada uma das cinco rodadas de clusterização, para cada um dos três métodos, para a seqüência da proteína $1j8b$. A linha no topo de cada método contém os ângulos da estrutura nativa conhecida da proteína.

Por fim, evidenciando o aumento significativo da probabilidade de gerarmos estruturas mais próximas à conformação nativa proporcionado pela informação sobre a estrutura secundária, colocamos lado a lado na Figura 5.4 a conformação nativa da proteína $1j8b$ e as estruturas de menor distância RMS produzidas por cada um dos algoritmos. Estas imagens foram feitas com o software de visualização *Rasmol*. Este software tem por característica atribuir a cor amarela para segmentos de folha-β apenas se estes estão próximos o suficiente para formarem pontes de hidrogênio. Na estrutura gerada pelo método MC-DSSP, apesar dos ângulos beta aparecerem nos locais corretos, há sutis diferenças de ângulos que impedem a aproximação maior dos segmentos e a identificação da folha-β. O método MC-RNA teve mais dificuldade em tratar a região da folha-β e em formar as α-hélices, mas está a caminho da forma tridimensional correta. O método MC por sua vez apresenta dificuldade maior, inclusive com a formação das α-hélices.

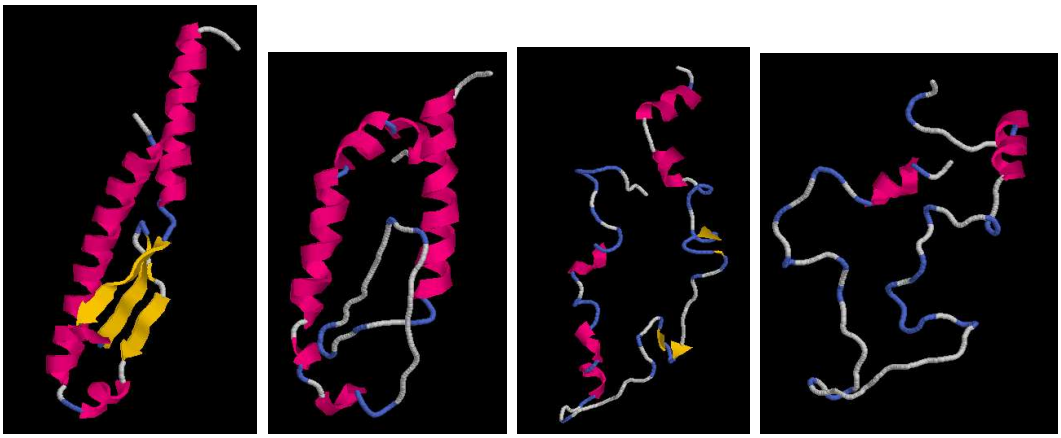


FIGURA 5.4 – Da esquerda para a direita: Conformação nativa da proteína $1j8b$, e conformações de menor distância RMS com a conformação nativa obtidas pelos métodos MC-DSSP, MC-RNA e MC. As energias são respectivamente de1087 1443, 1784, 1620 $Kcal/mole$.

5.2 Proteína $1g7d$, domínio C-terminal

A proteína $1g7d$ é uma proteína globular composta por 77 resíduos, em quatro α-hélices ligadas por *coils* e *turns*. Apesar de ser composta apenas por α-hélices (4, ligadas por pequenos segmentos *coil* ou *turn*), e ser portanto teoricamente um problema mais fácil, a $1g7d$ tem uma característica crucial: contém um núcleo hidrofóbico muito bem definido. Embora esta característica não pareça tão relevante à primeira vista, muitos métodos que ignoram a variável entrópica da hidrofobia falham ao tentar encontrar a conformação nativa deste tipo de proteína. Um deste métodos é o de mecânica molecular no vácuo. O pacote Tinker, usado durante todo este trabalho para a minimização de proteínas por descida de gradiente, implementa a mecânica molecular no vácuo. Para proteínas com forte núcleo hidrofóbico, a característica hidrofobia acaba por ter um papel mais importante do que a própria energia da molécula na determinação da estrutura nativa. Por isto esta proteína é importante para os testes do método MC-RNA.

A tabela abaixo mostra a seqüência de resíduos da proteína na primeira linha, acompanhada da estrutura secundária da conformação nativa (lida pelo DSSP) e pela previsão da estrutura secundária feita pelas RNAs respectivamente na segunda e terceira linhas.

```
01-60: PGCLPAYDALAGQFIEASSREARQAILKQGQDGLSGVKETDKKWASQYLKIMGKILDQGE
DSSP : CCCCTAAAAAAAAAAACCTAAAAAAAAAAAAAATTTTCTTTAAAAAAAAAAAAAAAATCT
RNA  : CCCC..AAAAAAAAA..C.AAAAAAAAAAAAA....C...AAAAAAAAAAAAAAA.CC

61-77: DFPASELARISKLIENK
DSSP : AAAAAAAAAAAAAAACC
RNA  : CC.AAAAAAAAAA.CC
```

Para a proteína $1g7d$, seguindo os critérios descritos para a proteína $1j8b$, selecionamos os ângulos Φ e Ψ de 12 resíduos, os resíduos 16, 17, 19, 34, 35, 36, 37, 39, 40, 41, 58 e 63.

A partir das amostras de conformações geradas pelos métodos MC-DSSP, MC-RNA e MC foram realizados cortes na cauda à direita das 3 distribuições de energia. As distribuições após os cortes tem o formato dos histogramas da Figura 5.5. Tanto pelos histogramas quanto pelos dados da Tabela 5.4 pode-se verificar o aumento da variância a medida que descemos do gráfico relativo ao método MC-DSSP em direção ao método MC, e passando pelo MC-RNA. A variância pode ser encarada como a medida de especialização de um cluster, e deve ser baixa para indicar que aquele cluster representa realmente um grupo de características comuns de seus membros.

Método	# $E = \infty$	# E alta	N	E min	E max	E média	E mediana	DP
MC	110	163	727	1282	1931	1500	1490	80
MC-RNA	100	157	743	1279	2311	1471	1456	87
MC-DSSP	38	78	890	1297	2155	1433	1424	56

TABELA 5.4 – Medidas de energia das amostras de conformações geradas para a seqüência da proteína $1g7d$. As três primeiras colunas contém respectivamente o número de conformações impossíveis de minimizar, o número de conformações cortadas da cauda à direita da distribuição e o número N de conformações destinadas à clusterização. As cinco colunas restantes são as menores e maiores energias da amostra de N conformações, a média, a mediana e o desvio padrão. Todos os valores se referem às conformações após a fase de minimização por descida de gradiente.

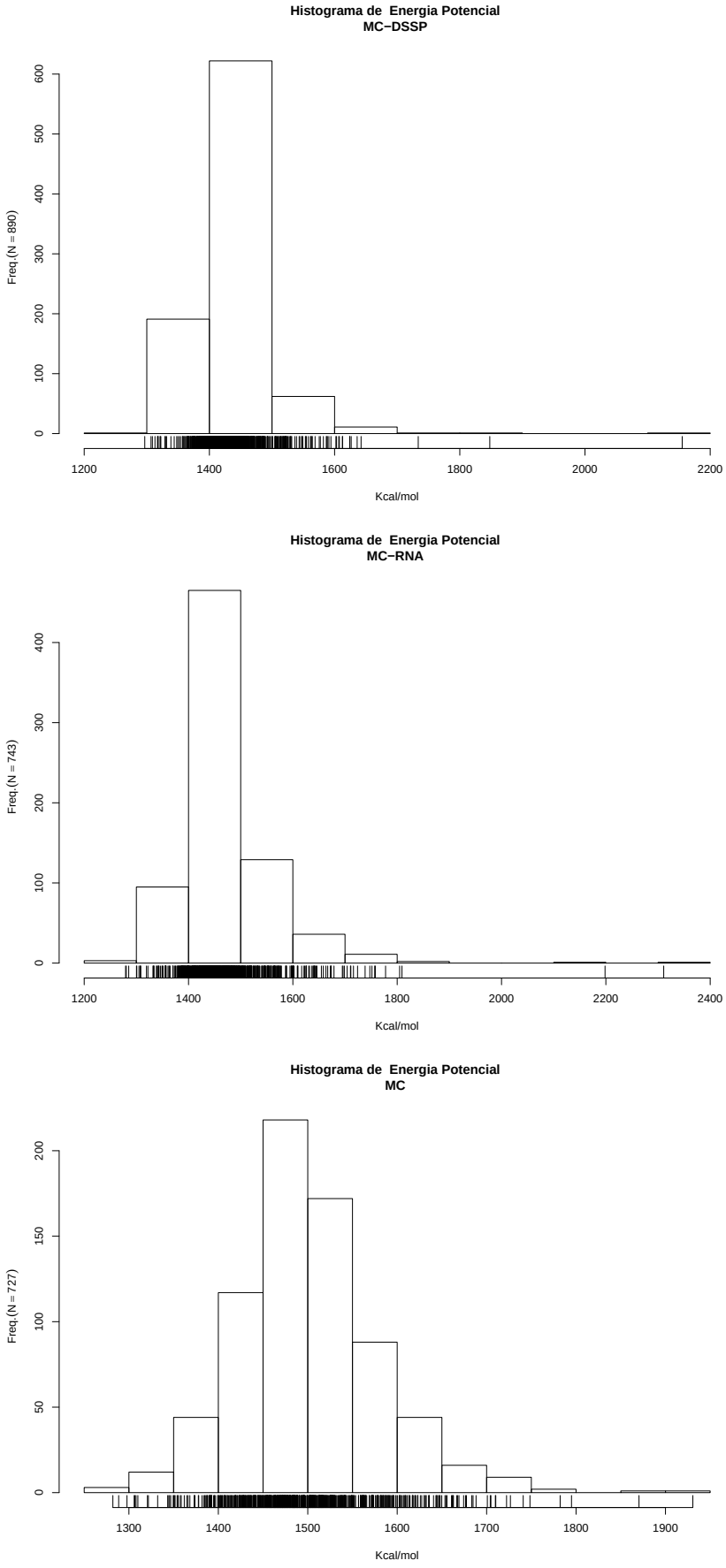


FIGURA 5.5 – Distribuição de energia potencial das amostras para as conformações geradas pelos métodos MC-DSSP (topo), MC-RNA e MC, para a proteína $1g7d$.

	MC				MC-RNA				MC-DSSP			
Cl	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS
Rodada no. 1												
1	11	10	14	12	**26**	15	**20**	**24**	**26**	**25**	24	19
2	12	12	10	6	9	**17**	9	9	14	14	12	14
3	**24**	**27**	**27**	**25**	15	16	22	19	21	18	16	17
4	10	14	9	17	15	16	14	11	22	23	**27**	**31**
5	15	9	12	12	9	10	9	11	6	9	10	8
Rodada no. 2												
1	**18**	15	12	10	15	9	19	18	21	18	17	17
2	14	9	12	12	**27**	**25**	**26**	**27**	6	10	10	8
3	**18**	16	16	11	9	10	7	10	14	13	9	13
4	11	**20**	12	16	9	17	9	9	22	22	**27**	**29**
5	11	12	**20**	23	14	13	13	10	**26**	**26**	26	22
Rodada no. 3												
1	**21**	**24**	**25**	**22**	**21**	15	**25**	**24**	20	17	15	14
2	16	12	20	20	8	12	15	11	14	13	11	13
3	15	9	12	13	20	14	16	19	14	16	12	16
4	6	14	5	9	15	14	9	11	17	19	**27**	**25**
5	14	13	10	8	10	**19**	9	9	**24**	**24**	24	21
Rodada no. 4												
1	**21**	**21**	**24**	**21**	9	**17**	8	8	14	14	12	14
2	19	15	13	11	**21**	12	**19**	15	21	19	17	18
3	9	20	12	17	17	**17**	15	13	7	8	9	7
4	8	7	11	11	13	12	18	17	21	23	**27**	**31**
5	15	9	12	12	14	16	14	21	**26**	**25**	24	19
Rodada no. 5												
1	14	9	12	12	10	16	10	9	19	12	12	11
2	17	15	14	13	9	9	9	9	19	**25**	**27**	**30**
3	14	14	14	19	14	15	11	11	4	6	13	6
4	9	13	13	15	15	13	17	21	**26**	**25**	25	21
5	**18**	**21**	**19**	13	**26**	**21**	**27**	**24**	21	21	12	21

TABELA 5.5 – Clusters das conformações da proteína $1g7d$. Para cada método há 5 colunas: o número do cluster, e as freqüências de presença dentro de cada cluster de conformações com menor Energia, menor Superfície Total, menor Superfície Hidrofóbica e menor distância RMS. Para esta proteína específica, na coluna MC-DSSP, as concentrações de mínimos RMS são determinados pela concentração de mínimos de superfície hidrofóbica exposta ao solvente.

A Tabela 5.5 de clusters da proteína $1g7d$ mostra a distribuição entre os cluster da ocorrência de valores mínimos para as medidas de energia, superfície total exposta ao solvente e superfície hidrofóbica exposta ao solvente. Um aspecto que merece destaque nesta Tabela é a importância da última medida.

Para determinar os clusters de maior concentração de mínimos de distancia RMS das configurações geradas pelo método MC-DSSP, ao contrário do que ocorre com as demais proteínas estudadas neste trabalho e com os métodos MC-RNA e MC para a própria $1g7d$, não é necessária a concentração de mínimos para as três medidas no mesmo cluster. Ao contrário, para um cluster poder ser declarado possuidor da maior concentração de mínimos de distância RMS com a conformação nativa basta que ele contenha a maior concentração de mínimos de superfície hidrofóbica exposta ao solvente. Nem mesmo o número de mínimos de energia não necessita ser superior

ao dos demais clusters. E isto acontece para as 5 rodadas de clusterização das configurações geradas pelo método MC-DSSP para esta proteína.

No método MC-RNA a influência da componente entrópica inserida pela medida da superfície hidrofóbica ainda está presente mas não é mais determinante para a determinação do cluster de mínimos de RMS. Na rodada de clusterização de número 4 do método MC-RNA, os cluster 2 e 3 contém relativamente mais configurações com superfícies hidrofóbicas mínimas (19 e 18) do que os demais, mas contém apenas ambos contém apenas 12 mínimos de superfície total das 74 configurações possíveis (para a amostra de 743 configurações, são consideradas mínimas as 10% que tem os menores valores). O cluster vencedor contém concentrações respectivamente de 16 e 14 mínimos de superfície total e hidrofóbica. A influência da presença de mínimos de superfície hidrofóbica em clusters diminui ainda mais para o método MC, onde nem mesmo um cluster com concentração concorrente de mínimos das três medidas consegue reunir maioria de mínimos de distância RMS com a conformação nativa.

A partir destas constatações e comparações podemos chegar a importante conclusão que a quantidade de informação influencia diretamente no aparecimento de características específicas detectáveis pelo método de clusterização. No caso de proteínas globulares, aparentemente a informação detectada pela adição de informação sobre estrutura secundária promovida pelas RNAs é que a posição dos resíduos hidrofóbicos é extremamente importante para a determinação da estrutura nativa. Nas Figuras 5.6 e 5.7 encontram-se os histogramas das distribuições do segundo cluster da segunda rodada de clusterização das conformações geradas por MC-RNA, e do quarto cluster, rodada 1, das configurações geradas por MC-DSSP. É fácil ver que a superfície hidrofóbica está com média flagrantemente menor na distribuição de probabilidade gerada pelo MC-DSSP.

A Figura 5.8 mostra a comparação dos ângulos dos centros de alguns clusters com os mesmos ângulos da conformação nativa da proteína $1g7d$. Os cluster escolhidos são os de maior concentração de mínimos de superfície hidrofóbica exposta ao solvente, e a proximidade com os ângulos da conformação nativa demonstra mais uma vez a influência deste fator na determinação da estrutura tridimensional.

Por fim, a Figura 5.9 mostra uma comparação visual entre a conformação nativa e algumas das conformações geradas pelos métodos. As conformações mostradas oriundas do método MC-RNA não foram porém as de menor superfície hidrofóbica, mas as de menor energia, com níveis baixos de superfície hidrofóbica exposta. Há um compromisso entre a compactação de estruturas, exposição de elementos hidrofóbicos ao solvente e energia potencial que estabelece limites abaixo dos quais o decréscimo de uma medida acarreta na elevação de outras. As imagens mostram que o método MC-RNA gerou conformações visualmente muito similares à estrutura nativa, só não encontrando a conformação nativa por defeitos nos ângulos de alguns resíduos dos segmentos coil que interligam algumas α-hélices.

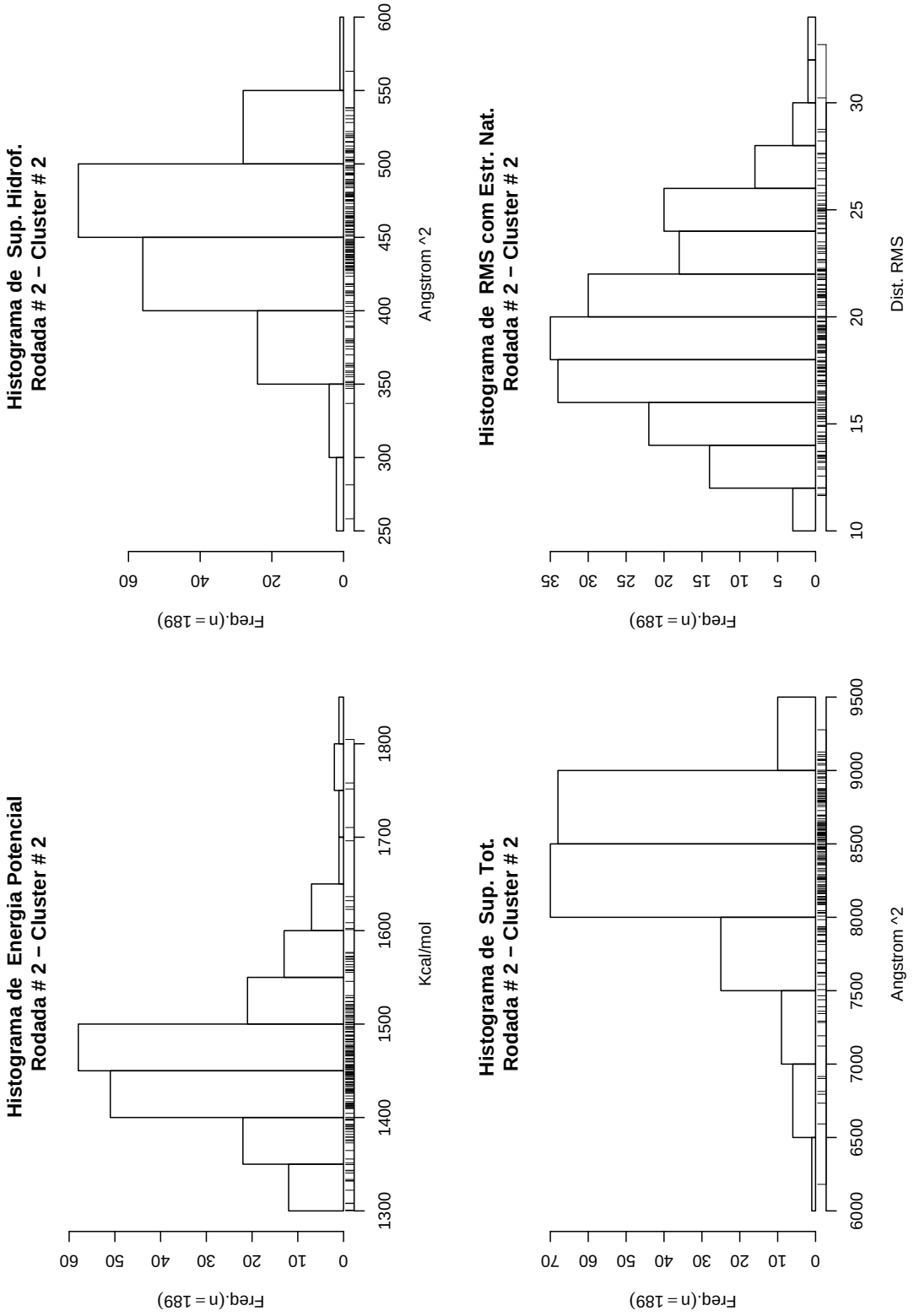


FIGURA 5.6 – Distribuição da energia, superfície total, superfície hidrofóbica e distância RMS à conformação nativa para o cluster 2 da segunda rodada de clusterização para a proteína $1g7d$, conformações geradas por MC-RNA.

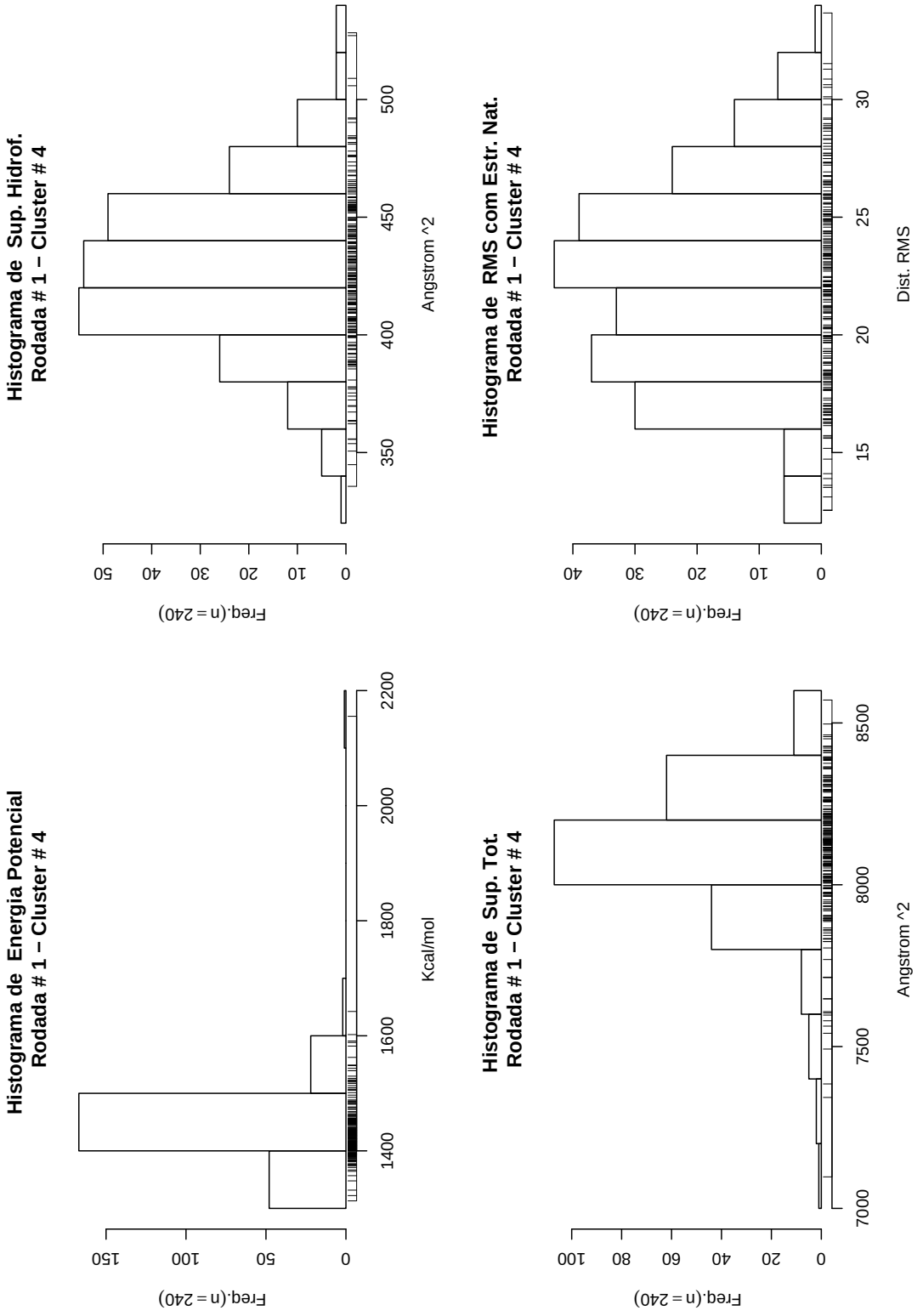
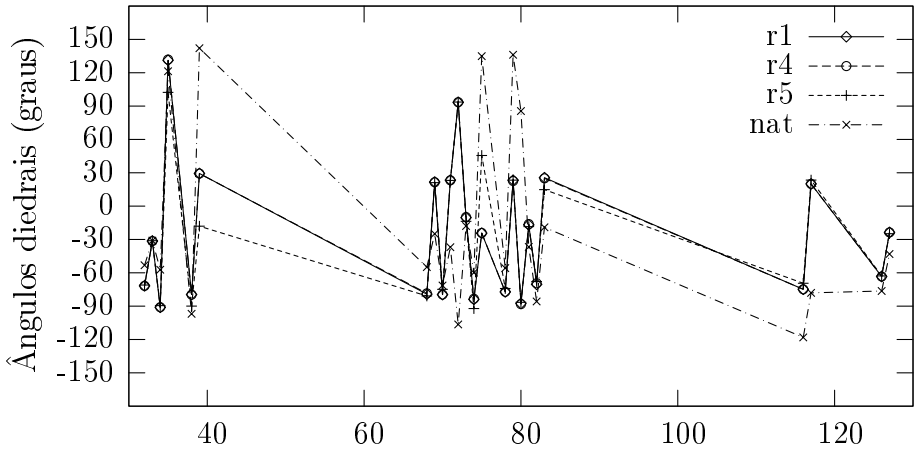
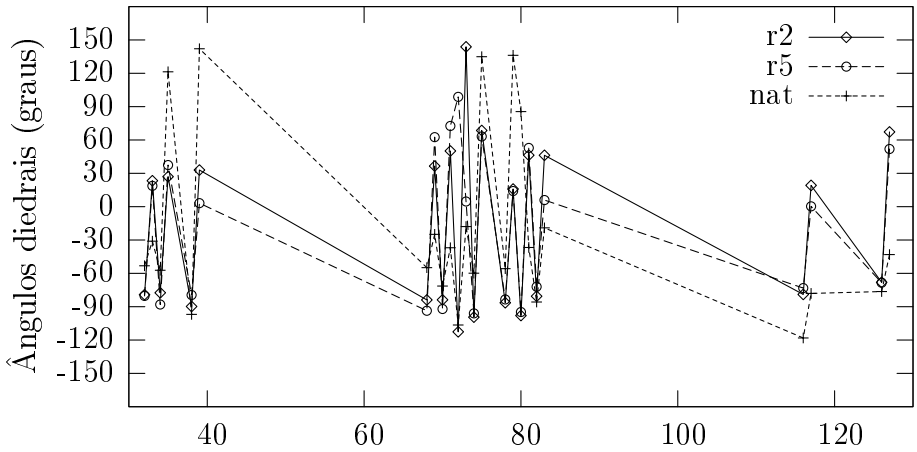


FIGURA 5.7 – Distribuição da energia, superfície total, superfície hidrofóbica e distância RMS à conformação nativa para o cluster 1 da segunda rodada de clusterização para a proteína $1g7d$, conformações geradas por MC-DSSP.

MC-DSSP: Ângulos da estrutura nativa e dos clusters das rodads1, 4 e 5 (clusters 4, 4 e 2)



MC-RNA: Ângulos da estrutura nativa e dos clusters das rodadas2 e 5 (clusters 2 e 5)



MC: Ângulos da estrutura nativa e dos clusters das rodadas1, 3 e 4 (clusters 3, 1 e 1)

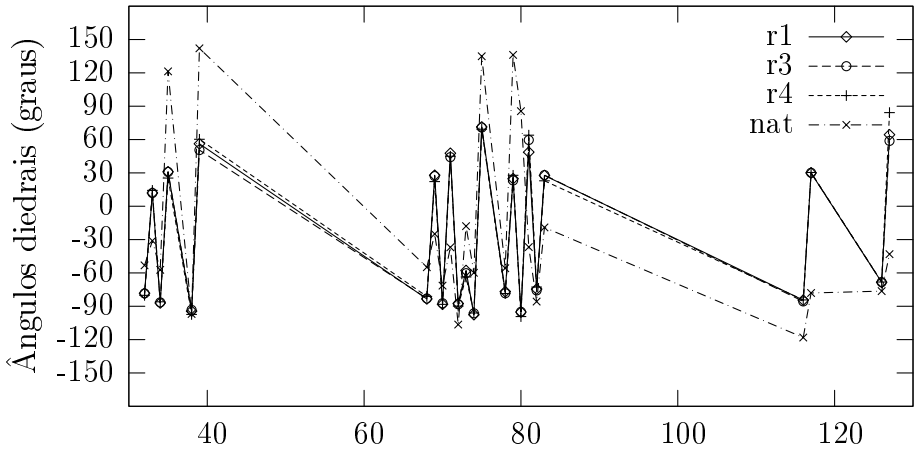


FIGURA 5.8 – Os gráficos mostram os ângulos do cluster que tem a maior concentração de estruturas com menor superfície hidrofóbica exposta ao solvente para cada um dos três métodos, para seqüência da proteína $1g7d$. A título de comparação, a linha de rótulo *nat* corresponde aos ângulos da estrutura nativa. No eixo horizontal, os índices dos ângulos

FIGURA 5.9 – Da esquerda para a direita e de cima para baixo: Conformação nativa da proteína $1g7d$, conformação de menor distância RMS com a conformação nativa pelo método MC-DSSP, as três conformações de menor *energia* pelo método MC-RNA, e em baixo à direita a conformação de menor distância RMS pelo método MC. As energias e RMS de cada uma são respectivamente (em $Kcal/mole$, Å): (877,0), (1425, 10.8), (1279, 12.3), (1281, 13.9), (1285, 15) e (1387, 10.4).

5.3 Proteína $1i74$, domínio 2

A proteína $1i74$ é composta por 108 resíduos que formam uma seqüência de segmentos de segmentos folhas-β e α-hélices intercaladas. Uma peculiaridade e um desafio para a determinação da estrutura nativa desta proteína é a existência de folhas-β formadas por segmentos não contíguos na seqüência. Por exemplo, o segmento β que vai do resíduo 11 ao 20 forma pontes de hidrogênio com outro segmento que vai do resíduo 56 ao 63 em uma folha-*beta* paralela. Entre o primeiro e o segundo há a maior α-hélice do domínio, e o próximo segmento forma uma folha-β antiparalela com o último. Portanto não é uma estrutura trivial de se determinar, porém há várias proteínas que se assemelham a ela, e por isto vale a pena investigar.

A tabela abaixo mostra a seqüência de resíduos do domínio 2 da proteína $1i74$ na primeira linha, acompanhada da estrutura secundária da conformação nativa (lida pelo DSSP) e pela previsão da estrutura secundária feita pelas RNAs respectivamente na segunda e terceira linhas.

```
01-60 : IDAKTFELNGSQVRVAQVNTVDINEVLERQNEIEEAIKASQAANGYSDFVLMITDILNSN
DSSP  : CCBBBBBBTTBBBBBBBBBBTCAAAAAAAAAAAAAAAAAAAAATCTBBBBBBBBTTTTB
RNA   : CC......CC..BBB.BB......AAAAAAAAAAAAAAAA..CCCC.BBBBB....CCC

61-108: SEILALGNNTDKVEAAFNFTLKNNHAFLAGAVSRKKQVVPQLTESFNG
```

```
DSSP  : BBBBBBBTTAAAAAAAATCCCBTTBBBBTTCCCAAAATAAAAAAAACC
RNA   : ..BBB..CC.AAAAA......CCC.BBB.................CC
```

Para a escolha dos resíduos cujos ângulos diedrais fizeram parte da clusterização foram considerados os resíduos sem previsão por parte das RNAs e distantes das extremidades. Como as RNAs não foram capazes de fazer nenhuma previsão sobre os resíduos do primeiro segmento β ou da última α-hélice, esse resíduos foram descartados da clusterização. Foram selecionados28 resíduos: 11, 12, 16, 19, 20, 21, 22, 23, 24, 42, 43, 48, 54, 55, 56, 57, 61, 62, 66, 67, 70, 76, 77, 78, 79, 80, 81 e 85.

A Tabela 5.6 mostra os valores para tamanho de amostras e medidas de energia para os três métodos. Podemos perceber que, em comparação com as duas primeiras proteínas estudadas neste trabalho, o número de estruturas que falharam em minimizar a energia ou pararam em mínimos locais muito alto aumentou. Isto acontece devido ao tamanho relativamente maior da seqüência. Com o aumento das seqüências o número de combinações possíveis de ângulos a serem escolhidos por um método Monte Carlo que não causem colisões entre moléculas decresce. Quando moléculas próximas causam energias que tendem ao infinito a descida de gradiente pode ficar inviabilizada. É interessante notar que o acréscimo de informação sobre a estrutura secundária diminui o risco destas mal-formações, como mostra a diminuição de descartes do método MC-RNA me relação ao MC, e do MC-DSSP em relação ao MC-RNA.

Método	# $E = \infty$	# E alta	N	E min	E max	E média	E mediana	DP
MC	151	171	678	1438	2869	1775	1778	143
MC-RNA	116	167	717	1483	2699	1769	1769	125
MC-DSSP	19	38	943	1442	1912	1702	1706	60

TABELA 5.6 – Medidas de energia das amostras de conformações geradas para a seqüência da proteína $1i74$. As três primeiras colunas contém respectivamente o número de conformações impossíveis de minimizar, o número de conformações cortadas da cauda à direita da distribuição e o número N de conformações destinadas à clusterização. As cinco colunas restantes são as menores e maiores energias da amostra de N conformações, a média, a mediana e o desvio padrão. Todos os valores se referem às conformações após a fase de minimização por descida de gradiente.

Como mostram os histogramas da Figura 5.10, que são traduzidos em números na Tabela 5.6, as distribuições de energia para os três métodos Monte Carlo são muito semelhantes, sendo o desvio padrão praticamente a única diferença entre os métodos. Isto indica que talvez seja necessário aumentar o tamanho da amostra para compensar a complexidade embutida nos graus de liberdade de seqüências maiores. Mas isto também tem um limite, pois a complexidade cresce exponencialmente com o aumento da estrutura.

À excessão da rodada de clusterização 2 o resultado das clusterizações foi indefinido para o método MC-RNA, como mostra a Tabela 5.7. Os gráficos das distribuições de medidas da Figura 5.11 também corroboram esta impressão, dado que a distribuição de energia para o cluster é similar à distribuição do total da amostra. Apesar disto, na comparação dos ângulos do cluster3 da segunda rodada da coluna

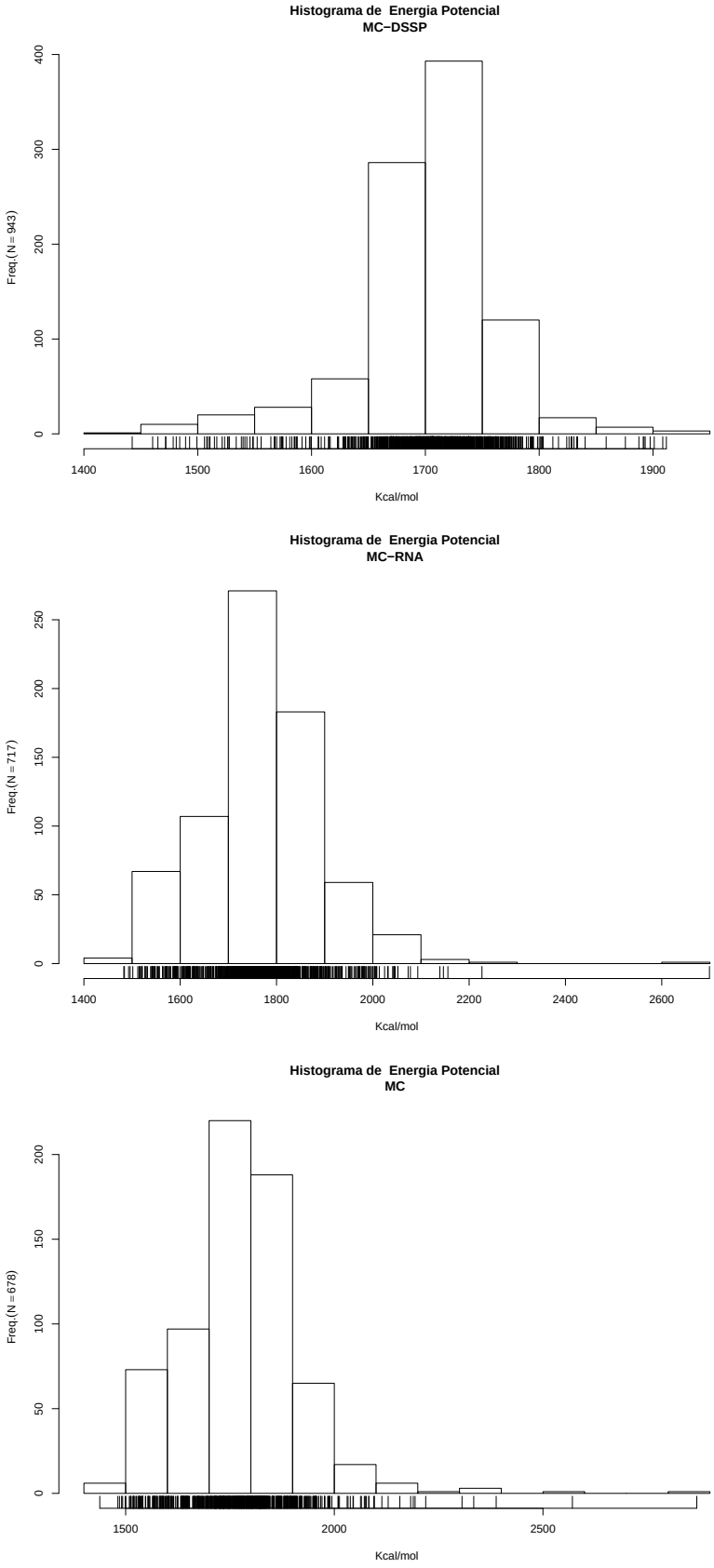


FIGURA 5.10 – Distribuição de energia potencial das amostras de conformações geradas pelos métodos MC-DSSP (topo), MC-RNA e MC, para a proteína1i74.

MC-RNA com os ângulos da estrutura nativa, o método MC-RNA ainda se saiu um pouco melhor do que o método MC (Figura 5.12). Em compensação a proximidade dos ângulos dos clusters do método MC-DSSP com a conformação nativa salta aos olhos, evidenciando a influência da informação sobre estrutura secundária para auxiliar o método Monte Carlo no dobramento de proteínas.

	MC				MC-RNA				MC-DSSP			
Cl	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS
Rodada no. 1												
1	12	12	12	9	13	11	9	9	1	4	1	3
2	14	17	13	14	**19**	18	**19**	21	17	15	14	11
3	6	7	8	10	6	10	9	9	12	10	14	11
4	**20**	**18**	**20**	**20**	18	**23**	**19**	19	25	22	31	31
5	15	13	14	14	15	9	15	13	**39**	**43**	**34**	**38**
Rodada no. 2												
1	13	11	13	9	14	10	16	15	**32**	**35**	28	34
2	15	15	15	17	13	13	15	16	10	8	12	9
3	9	13	10	10	**20**	**25**	**20**	**20**	25	21	**31**	31
4	11	12	**16**	16	12	9	8	9	18	16	15	12
5	**19**	**16**	13	15	12	14	12	11	9	14	8	8
Rodada no. 3												
1	11	12	12	13	13	15	11	15	30	24	24	31
2	**16**	**15**	15	18	15	16	14	13	6	15	11	7
3	13	13	11	12	**18**	12	**18**	15	5	4	4	5
4	13	12	12	9	13	**18**	15	15	**42**	**37**	**46**	**43**
5	14	**15**	**17**	15	12	10	13	13	11	14	9	8
Rodada no. 4												
1	13	13	14	10	16	14	13	14	10	16	12	11
2	**18**	**17**	**16**	**16**	5	8	6	8	29	27	22	27
3	10	14	11	11	14	14	**17**	18	12	10	14	12
4	12	13	11	15	17	13	**18**	13	**42**	**37**	**45**	**41**
5	14	10	15	15	**19**	**22**	**17**	18	1	4	1	3
Rodada no. 5												
1	15	15	14	12	15	16	15	14	7	6	5	5
2	**18**	**16**	**16**	**17**	11	10	11	13	26	22	32	32
3	11	14	12	15	**17**	11	**16**	10	9	14	8	8
4	10	9	14	11	13	12	**15**	19	**34**	**36**	**34**	**37**
5	13	13	11	12	15	**22**	14	15	18	16	15	12

TABELA 5.7 – Clusters das conformações da proteína $1i74$. Para cada método há 5 colunas: o número do cluster, e as freqüências de presença dentro de cada cluster de conformações com menor Energia, menor Superfície Total, menor Superfície Hidrofóbica e menor distância RMS. Para esta proteína específica, na coluna MC-DSSP, as concentrações de mínimos RMS são determinados pela concentração de mínimos de superfície hidrofóbica exposta ao solvente.

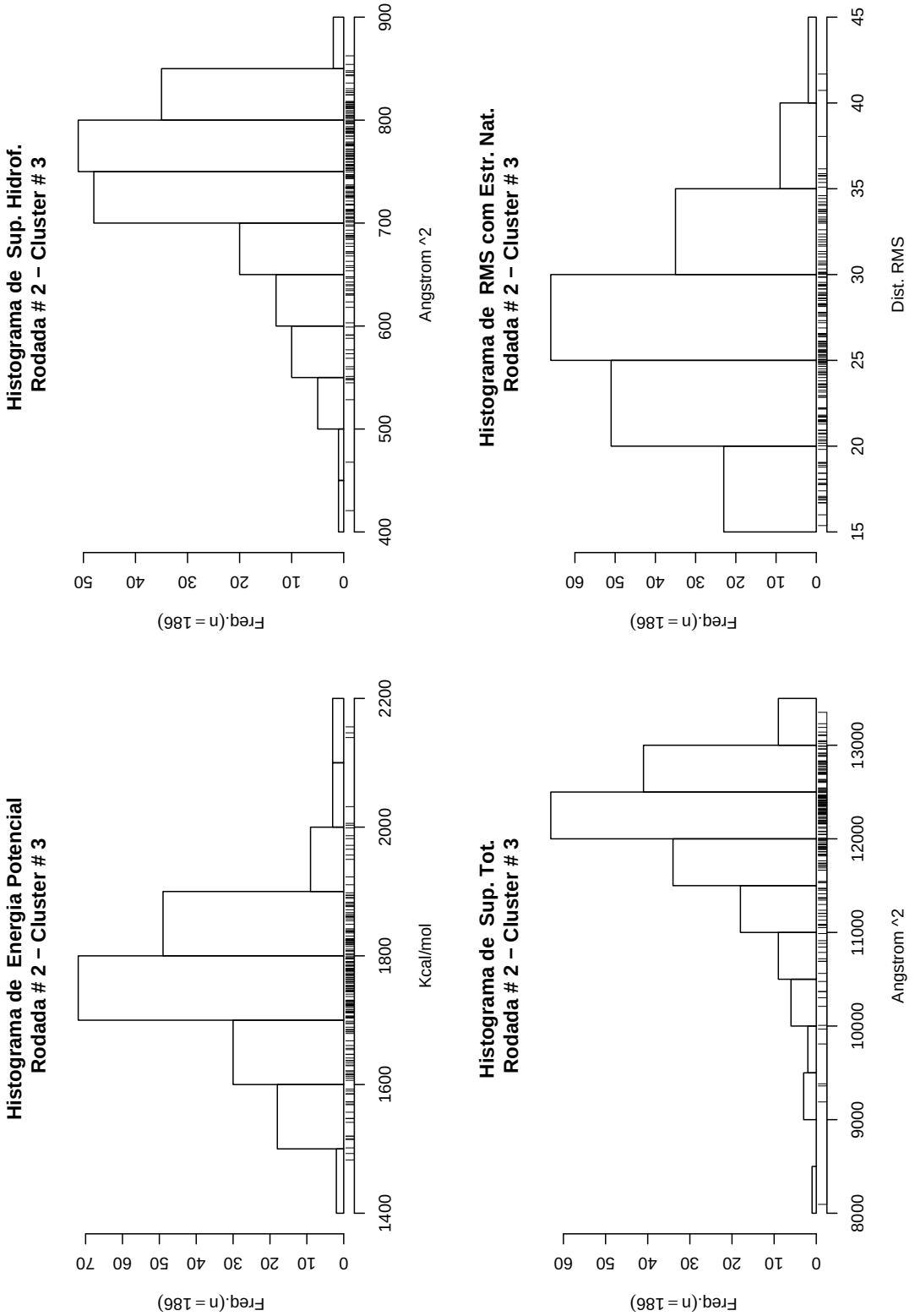
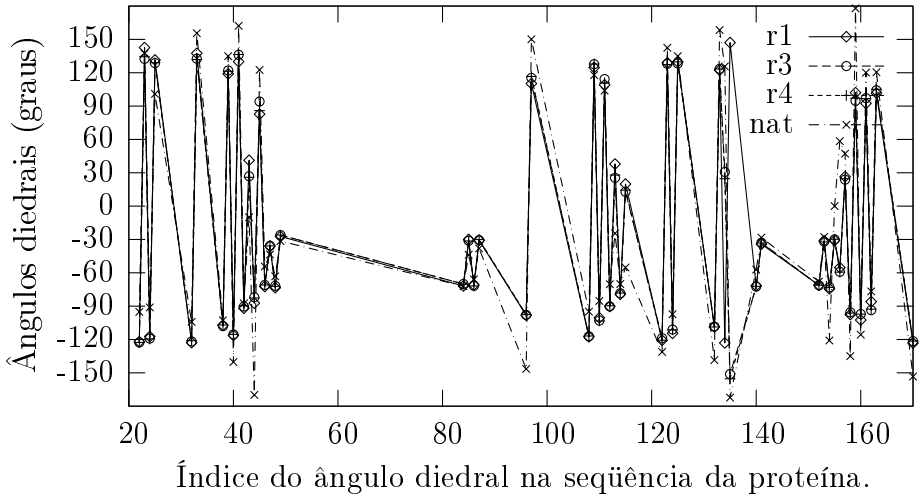
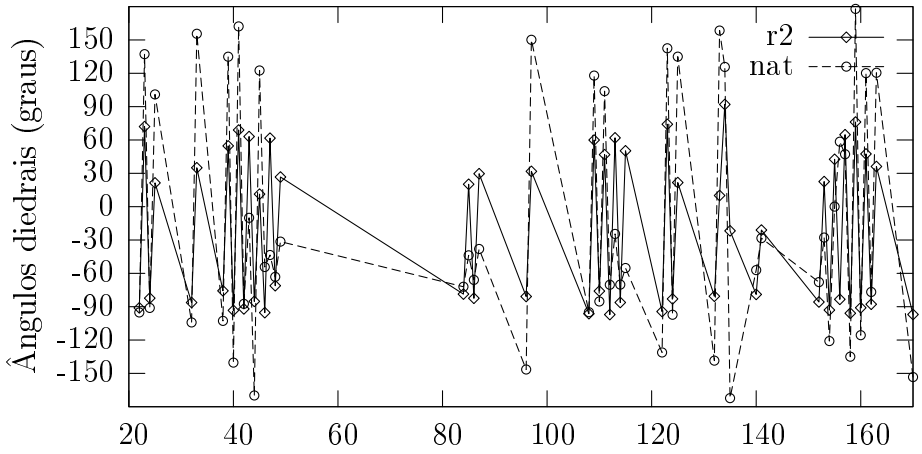


FIGURA 5.11 – Distribuição da energia, superfície total, superfície hidrofóbica e distância RMS à conformação nativa para o cluster 3 da segunda rodada de clusterização para a proteína 1*i*74, conformações geradas por MC-RNA.

MC-DSSP: Ângulos da estrutura nativa e dos clusters das rodadas 1, 3 e 4 (clusters 5, 4 e 4)



MC-RNA: Ângulos da estrutura nativa e do cluster 3, rodada 2



MC: Ângulos da estrutura nativa e dos clusters das rodadas 1, 4 e 5 (clusters 4, 2 e 2)

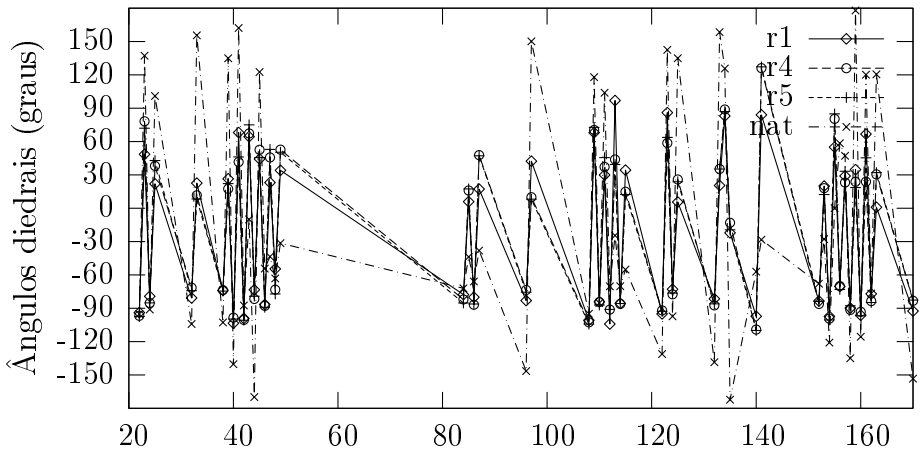


FIGURA 5.12 – Os gráficos mostram os ângulos do cluster que concomitantemente tem a maior concentração de estruturas com menor energia, superfície total e superfície hidrofóbica exposta ao solvente, para três rodadas de clusterização para cada um dos três métodos, para a seqüência da proteína 1*i*74. A título de comparação, a linha de rótulo *nat* corresponde aos ângulos da estrutura nativa.

Na comparação visual com a estrutura nativa, é visível a semelhança entre esta e a conformação de menor RMS do método MC-DSSP (Figura 5.13). A conformação do método MC-DSSP tem na ordem certa e muitas vezes na posição espacial correta praticamente todos os elementos da estrutura da conformação nativa. Apenas algumas diferenças sutis de ângulos de resíduos em *turns* ou *coils* que conectam dois segmentos de folha-β impedem que se aproximem completamente e sejam reconhecidos com a folha-*beta* da proteína. O método MC-RNA não conseguiu bons resultados nas estruturas secundárias das extremidades porque as RNAs falharam em prever que eram um folha-β, e não conseguiu na sua estrutura de menor RMS acertar todos os segmentos que teriam ângulos de folha-β. Em defesa do método porém podemos dizer que: (1) a estrutura nativa é de difícil determinação dada a formação de estruturas intercaladas como a folha-β paralela intercalada por uma α-hélice de folhas-β, sendo que mínimas imperfeições no escolha dos ângulos pelo método provocam facilmente sobreposição de átomos, e (2) o método MC-RNA se saiu no mínimo tão bem quanto o método MC.

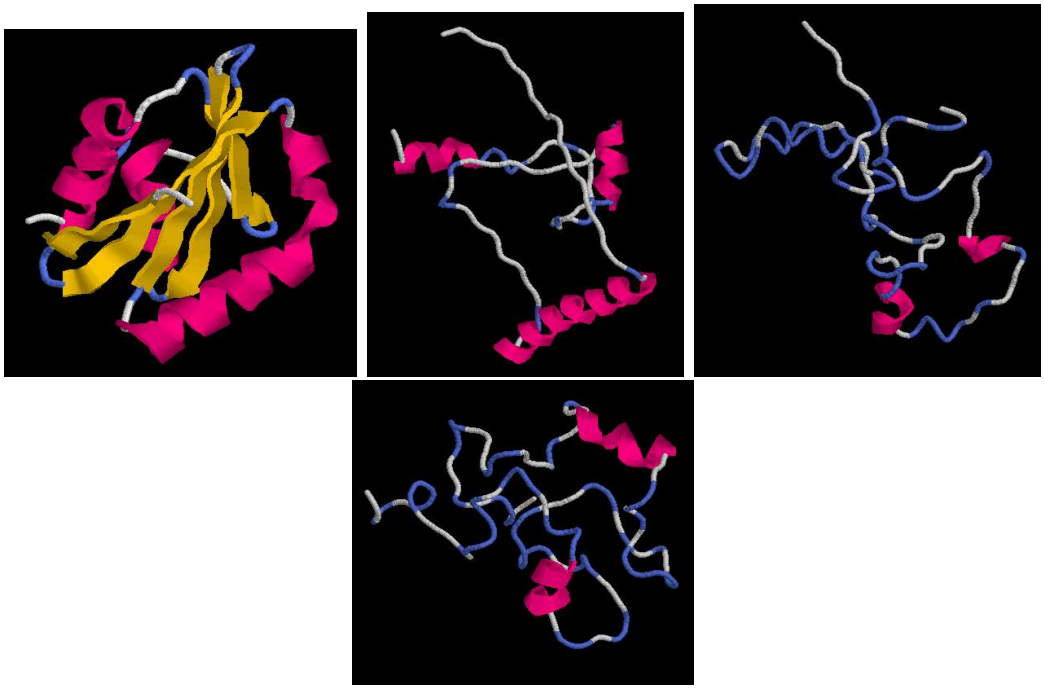


FIGURA 5.13 – Da esquerda para a direita e de cima para baixo: Conformação nativa da proteína $1i74$, conformações de menor distância RMS com a conformação nativa pelos métodos MC-DSSP, MC-RNA, e em baixo pelo método MC. As energias são respectivamente de 1194 1460, 1608 e 1520 $Kcal/mole$

5.4 Proteína $1kkg$

A proteína $1kkg$ tem α-hélices e folhas-β intercaladas em uma estrutura tridimensional em forma de espiral. É composta de 108 resíduos, dispostos na listagem abaixo na primeira linha. A estrutura secundária da conformação nativa (lida pelo DSSP) e pela previsão da estrutura secundária feita pelas RNAs estão respectivamente na segunda e terceira linhas.

```
01-60 : MAKEFGRPQRVAQEMQKEIALILQREIKDPRLGMMTTVSGVEMSRDLAYAKVYVTFLNDK
DSSP  : CCCCTTTAAAAAAAAAAAAAAATTTTTTTAAATTCCBCTCBBBBTTTTBBBBBBBCTAAA
RNA   : CCC....AAAAAAAAAAAAAAAAAA..CCCCC...BBBBBB..CC....BBBBB..CCC

61-108: DEDAVKAGIKALQEASGFIRSLLGKAMRLRIVPELTFFYDNSLVEGMR
DSSP  : CAAAAAAAAAAAAATAAAAAAAAAAAATTCTCCCBBBBBBCCTTTTTCC
RNA   : C.AAAAAAAAAAA...AAAAAA............BBB.......CCC
```

Podemos observar que a previsão das RNAs para a proteína $1kkg$ foi melhor e mais abrangente do que para a seqüência de mesmo tamanho do domínio 2 da Proteína $1i74$. Portanto, descartando resíduos próximos às extremidades e escolhendo resíduos sem previsão por parte das RNAs, selecionamos para clusterização a seguinte lista de 29 resíduos: 27, 28, 34, 35, 36, 43, 44, 47, 48, 49, 50, 56, 57, 62, 75, 76, 77, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94 e 95.

Apesar da melhora na previsão da estrutura secundária por parte das RNAs, a Tabela 5.8 nos mostra que, como ocorreu com o domínio 2 da proteína $1i74$, as distribuições são muito parecidas, principalmente as geradas pelos métodos MC e MC-RNA. A diferença para a distribuição MC-DSSP fica novamente por conta do menor desvio padrão. A impressão se confirma se compararmos o gráfico da distribuição de energia da Figura 5.15 com a distribuição de energia para o conjunto das conformações geradas pelo método MC-RNA (Figura 5.14). O cluster em questão é o quinto cluster da quarta rodada da coluna MC-RNA na Tabela 5.9, e é também o único cluster gerado a partir das conformações geradas a partir do método MC-RNA a preencher o pré-requisito de concentrar simultaneamente a maior parcela de mínimos de energia, superfície total e superfície hidrofóbica relativamente aos demais. Se formos analisar os gráficos de comparação entre ângulos centrais dos clusters, os centros dos clusters do MC-RNA e do método MC praticamente se equiparam. Em compensação os ângulos centrais dos clusters do método MC-DSSP mais uma vez se aproximam muito bem dos ângulos da conformação nativa.

Método	# $E = \infty$	# E alta	N	E min	E max	E média	E mediana	DP
MC	191	238	571	12919	14386	13448	13508	209
MC-RNA	124	234	642	12956	14233	13459	13500	195
MC-DSSP	51	118	831	12937	13826	13502	13529	129

TABELA 5.8 – Medidas de energia das amostras de conformações geradas para a seqüência da proteína $1kkg$. As três primeiras colunas contém respectivamente o número de conformações impossíveis de minimizar, o número de conformações cortadas da cauda à direita da distribuição e o número N de conformações destinadas à clusterização. As cinco colunas restantes são as menores e maiores energias da amostra de N conformações, a média, a mediana e o desvio padrão. Todos os valores se referem às conformações após a fase de minimização por descida de gradiente.

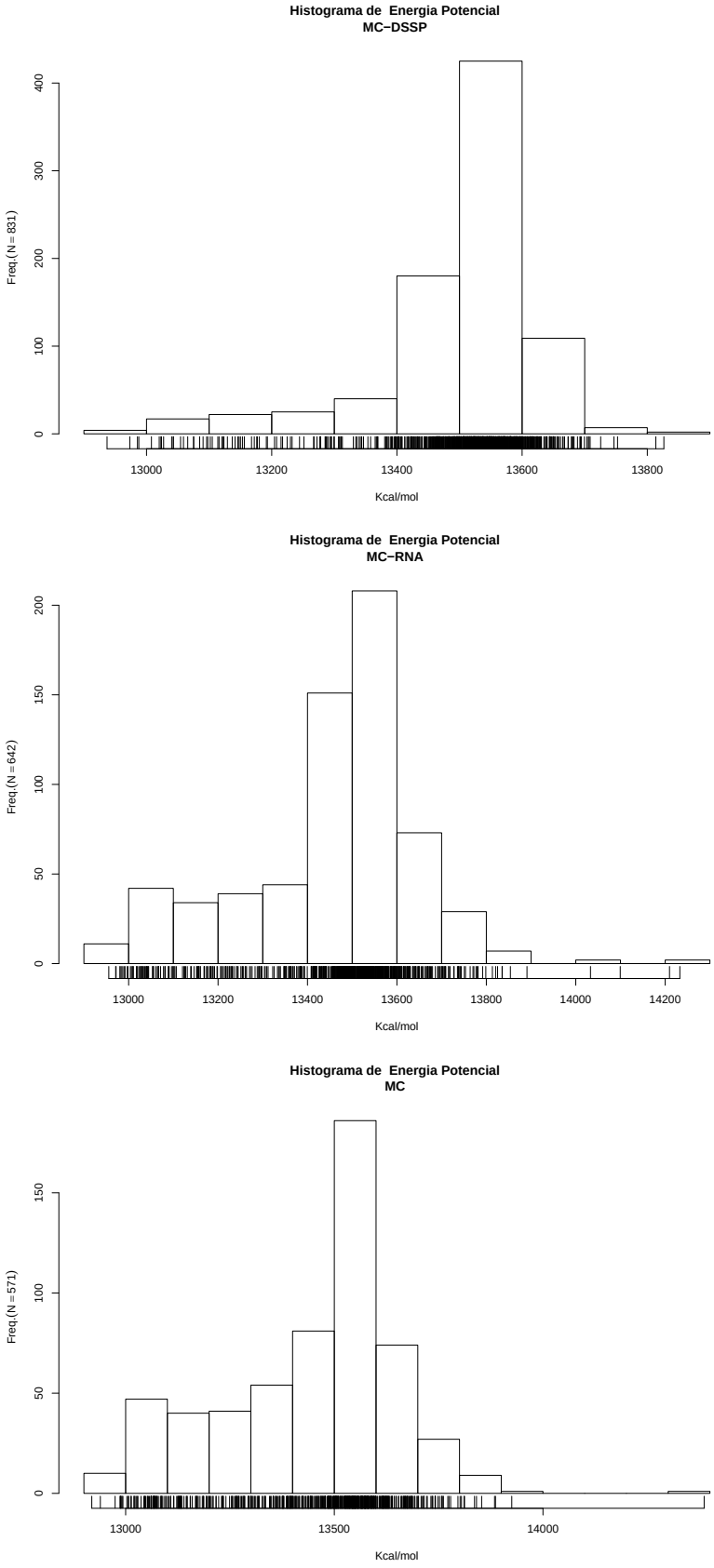


FIGURA 5.14 – Distribuição de energia potencial das amostras de conformações geradas pelos métodos MC-DSSP (topo), MC-RNA e MC, para a proteína $1kkg$.

	MC				MC-RNA				MC-DSSP			
Cl	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS	<E	<ST	<SH	<RMS
Rodada no. 1												
1	3	4	4	4	14	11	10	8	15	13	14	18
2	10	7	12	7	11	11	11	12	4	5	5	3
3	**23**	**25**	**23**	**24**	10	14	10	11	15	12	16	11
4	11	11	8	9	**18**	**16**	**20**	16	**37**	**37**	**34**	**40**
5	10	10	10	13	11	12	13	17	12	16	14	11
Rodada no. 2												
1	8	12	9	11	11	11	8	8	21	**26**	18	20
2	**16**	**15**	13	12	**16**	12	14	13	8	6	7	7
3	10	9	12	9	13	**16**	**16**	17	15	13	18	11
4	**16**	**16**	**17**	19	14	14	**15**	14	13	14	13	18
5	7	5	6	6	10	11	11	12	**26**	24	**27**	27
Rodada no. 3												
1	**19**	**19**	**20**	**19**	**17**	**15**	**19**	14	14	15	13	17
2	2	3	5	4	14	11	10	12	13	16	10	15
3	12	12	14	10	10	11	11	12	15	11	19	13
4	16	15	11	17	13	12	12	12	**24**	**24**	**21**	21
5	8	8	7	7	10	**15**	12	14	17	17	**20**	17
Rodada no. 4												
1	16	11	12	14	16	16	15	12	7	6	7	5
2	7	11	10	10	11	12	8	12	**30**	**30**	**32**	**32**
3	**18**	**18**	**18**	**17**	11	12	12	12	16	13	17	12
4	9	8	7	8	6	8	8	10	22	18	17	26
5	7	9	10	8	**20**	**16**	**21**	**18**	8	16	10	8
Rodada no. 5												
1	14	12	11	11	12	13	11	10	16	13	18	13
2	9	10	11	12	14	13	12	10	8	6	7	7
3	**17**	**19**	**18**	**18**	**17**	**14**	**18**	16	**19**	18	18	19
4	12	11	9	8	12	**15**	13	15	**20**	21	**23**	22
5	5	5	8	8	9	9	10	13	**20**	**25**	17	22

TABELA 5.9 – Clusters das conformações da proteína $1kkg$. Para cada método há 5 colunas: o número do cluster, e as freqüências de presença dentro de cada cluster de conformações com menor Energia, menor Superfície Total, menor Superfície Hidrofóbica e menor distância RMS. Para esta proteína específica, na coluna MC-DSSP, as concentrações de mínimos RMS são determinados pela concentração de mínimos de superfície hidrofóbica exposta ao solvente.

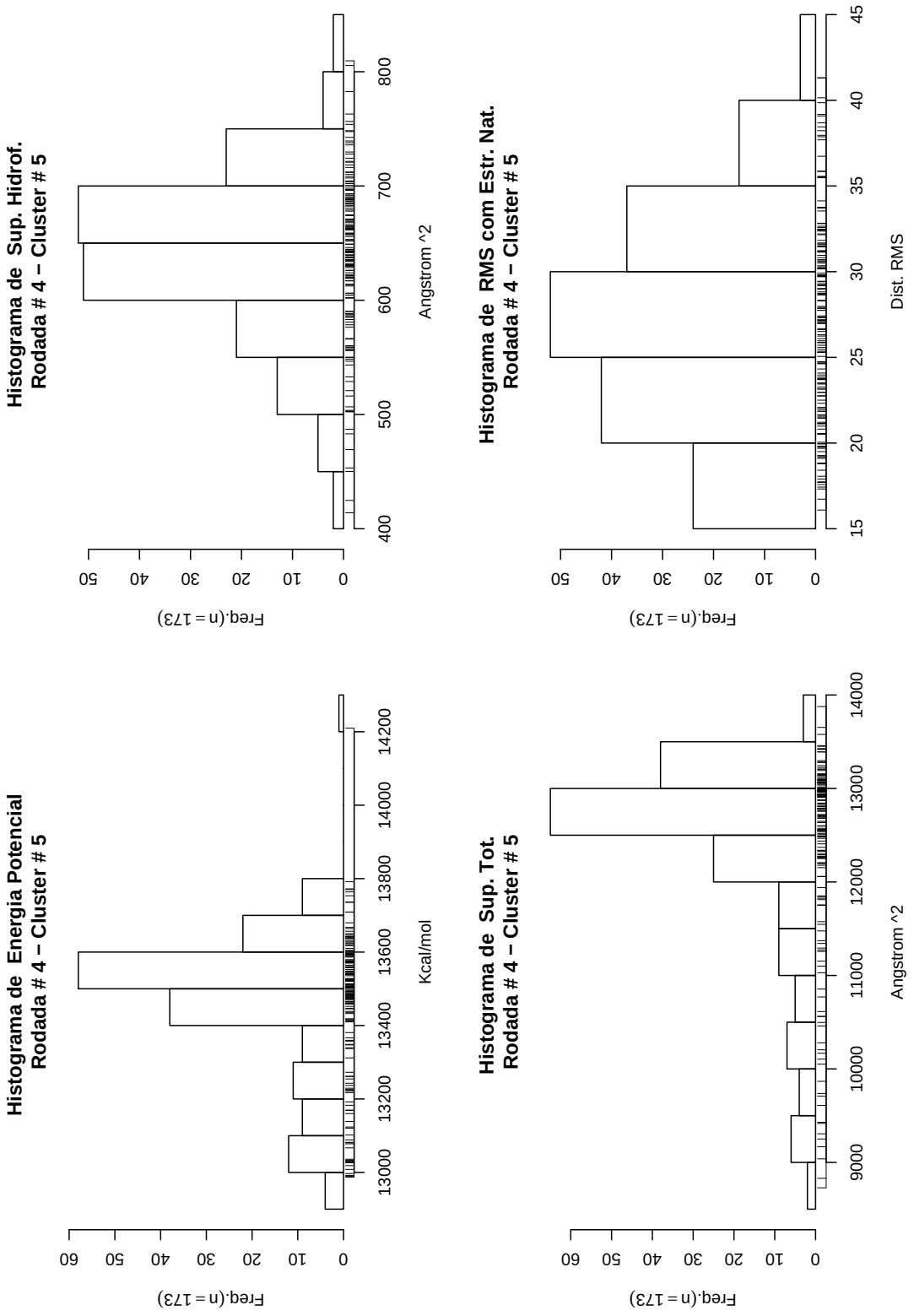
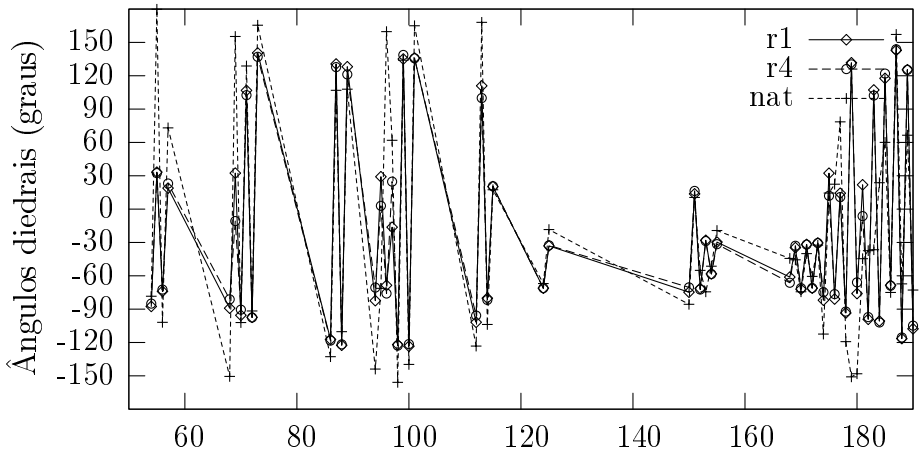
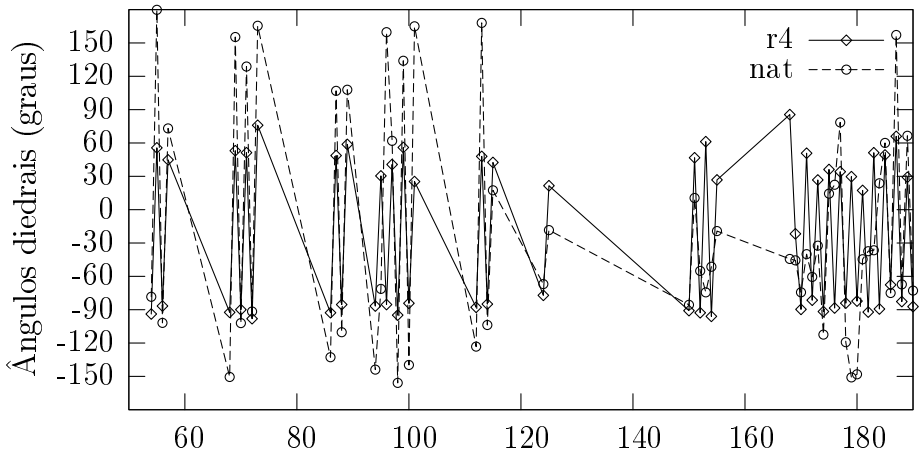


FIGURA 5.15 – Distribuição da energia, superfície total, superfície hidrofóbica e distância RMS à conformação nativa para o cluster 5 da quarta rodada de clusterização para a proteína $1kkg$, conformações geradas por MC-RNA.

MC-DSSP: Ângulos da estrutura nativa e dos clusters das rodadas1 e 4 (clusters 4 e 2)



MC-RNA: Ângulos da estrutura nativa e do cluster 5, rodada 4



MC: Ângulos da estrutura nativa e dos clusters das rodadas 1, 4 e 5 (clusters 3, 3 e 3)

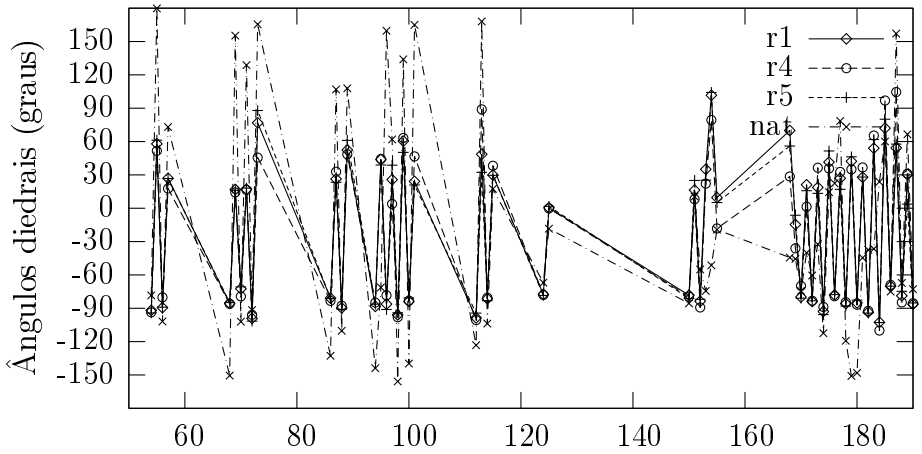


FIGURA 5.16 – Os gráficos mostram os ângulos dos clusters que tem a maior concentração de estruturas com menor superfície hidrofóbica exposta ao solvente para cada um dos três métodos, para a seqüência proteína 1*kkg*. A título de comparação, a linha de rótulo *nat* corresponde aos ângulos da estrutura nativa.

Apesar da estrutura dos ângulos não ser idêntica numericamente à estrutura nativa, como observado na Figura 5.16 MC-DSSP, os ângulos obtidos são uma versão quase em escala da estrutura nativa, indicando que a maior parte das estruturas dentro do cluster 4 se aproxima da estrutura nativa. Observa-se que os ângulos são similares. Comparando o método MC, MC-RNA e MC-DSSP, conclui-se que para esta proteína, os ângulos representativos dos clusters para os métodos MC e MC-RNA são equivalentes e os resultados são semelhantes. Porém, MC-DSSP apresenta resultados significativamente melhores. A diferença assim reside na capacidade de prever a estrutura secundária com exatidão, tornando assim o método limitado ao desempenho do método de previsão da estrutura secundária. Fosse o resultado de previsão da estrutura secundária da RNA semelhante ao DSSP, então a qualidade dos resultados para o método MC-RNA também se distinguiria do MC puro. Apesar disto, considerando o melhor caso para cada um dos métodos (veja Figura 5.17) em termos de energia, MC-DSSP obteve o melhor resultado.

Deve-se levar em conta também o fato que esta proteína representa um caso difícil, pois possui folhas beta no centro da estrutura, que desta forma influenciam a estrutura global da proteína. Qualquer erro na previsão das folhas beta afeta a estrutura de forma global.

A simples comparação visual entre as quatro estruturas na Figura mostra que os resultados MC-DSSP e MC-RNA encontram-se próximos da estrutura nativa. Faltam as pontes de hidrogênio ao longo das folhas-beta para que o programa Rasmol desenhe as estruturas coil no estilo de setas. Porém pode-se verificar que há um alinhamento dos resíduos no local onde a estrutura beta deveria ter se formado, indicando que uma minimização posterior poderia levar à formação das pontes de hidrogênio.

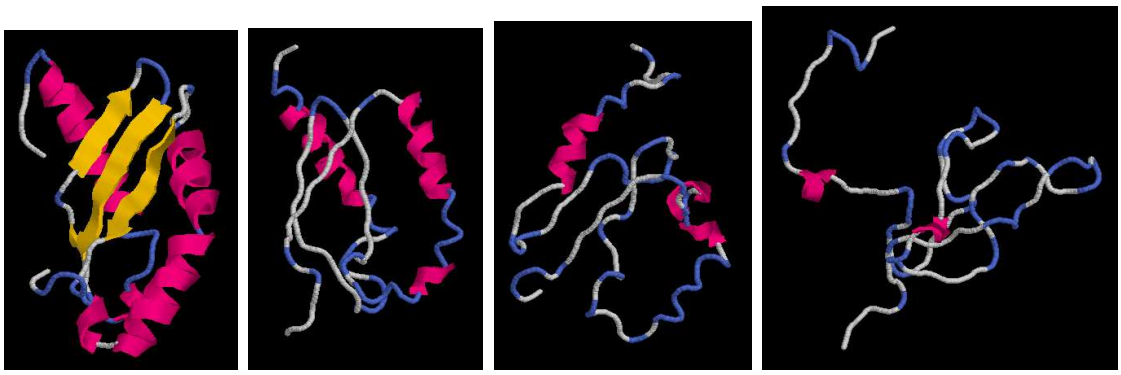


FIGURA 5.17 – Da esquerda para a direita: Conformação nativa da proteína $1kkg$, conformações de menor distância RMS com a conformação nativa pelos métodos MC-DSSP, MC-RNA, e em baixo pelo método MC. As energias são respectivamente de 13259 13137, 13207 e 13205 $Kcal/mole$

Capítulo 6

Considerações Finais

Neste trabalho foi proposto um novo método de simulação estocástico chamado MC-RNA. O método MC-RNA utiliza MC para percorrer o espaço de busca de conformações de proteína, e RNAs para a redução deste espaço de busca.

A utilização das RNAs para a redução do espaço de busca se dá através da previsão da estrutura secundária da proteína. Se utilizássemos o método MC sem nenhuma heurística que permita a redução do espaço de busca de proteínas, o algoritmo passaria a maior parte do tempo percorrendo estados que correspondem à conformações improváveis e distantes da conformação nativa. Com a previsão das RNAs reduz-se o espaço de busca restringindo o acesso à conformações consideradas pouco prováveis. O método MC-RNA acarreta em melhora na capacidade do MC em encontrar conformações mais próximas da conformação nativa. Esta melhora, que significa menor distância entre as conformações criadas pelo MC-RNA e a estrutura nativa, é dependente da qualidade da informação entregue pela RNA ao MC Para testar esta afirmação, foram realizados testes em paralelo para o método MC-RNA e para dois métodos de controle: o método MC e o método MC-DSSP. O método MC-DSSP pode ser considerado um método MC-RNA *ideal*, pois utiliza a informação da estrutura nativa de uma proteína lida pelo software DSSP como se fosse a previsão de um RNA com capacidade de acerto de 100%.

Os métodos de controle foram implementados e aplicados nas proteínas de teste com a motivação de verificar a influência da quantidade de informação no desempenho do método MC para a busca da conformação nativa de proteínas. O método MC-RNA e os métodos de controle foram testados com4 proteínas de domínio público, relativamente pequenas, utilizadas pelo CASP como alvo de simulações: $1j8b$, $1g7d$ domínio C-terminal, $1i74$ (domínio 2) e $1kkg$. Inicialmente foi criado um banco de dados contendo 377540 resíduos retirados de 2327 proteínas constantes da lista de proteínas não homólogas do grupo EVA. O banco de dados assim criado contem, para cada resíduo, o tipo do resíduo, os ângulos diedrais do resíduo quando na proteína original, e a estrutura secundária a qual pertencia. Uma vez construído, o banco de dados passou a ser o espaço de busca para o método MC-RNA e os demais métodos de controle, constituindo-se por si só em um primeiro passo na redução do espaço de busca, uma vez que exclui combinações de ângulos não permitidas no mapa de Ramachandran. O próximo passo foi obter as informações necessárias para os métodos MC-RNA e MC-DSSP. Para o método MC-RNA foram utilizadas as previsão de estrutura secundária do método PROF submetendo-se a seqüência de resíduos das quatro proteínas alvo para o servidor *Predict Protein*. Para obter a

informação real sobre a estrutura secundária necessária para o método MC-DSSP, os arquivos em formato PDB de cada proteína obtidos do servidor *Protein Data Bank* foram submetidos ao software DSSP.

A simulação dos métodos MC-RNA e dos métodos controle consistiram em gerar amostras de 1000 conformações para cada proteína, resíduo a resíduo, escolhendo os ângulos no banco de dados criado previamente. A variável em comum aos três métodos é que todos selecionam ângulos entre resíduos do banco com o mesmo tipo do resíduo da conformação. Os métodos MC-RNA e MC-DSSP utilizam um critério de seleção a mais, baseado em informação sobre a estrutura secundária. A diferença entre eles é que enquanto o método controle MC-DSSP utiliza informações verdadeiras sobre a estrutura nativa conhecida, o método MC-RNA utiliza a previsão de RNAs, e leva em conta o grau de confiança da previsão para a sua utilização. O produto destas simulações, 12 amostras de 1000 conformações, uma para cada proteína e para cada método, sofreu minimização de energia por descida de gradiente e clusterização por *k-means*. Para a clusterização foram selecionados apenas ângulos diedrais de resíduos pertencentes à segmentos coils, devido à sua importância na configuração de dobramento das estruturas secundárias e conseqüente conformação tridimensional. O processo de clusterização foi realizado de maneira a formar cinco clusters, e foi repetido cinco vezes para cada amostra de 1000 conformações com inicializações aleatórias. Findo o processo de clusterização, para cada amostra os clusters das cinco rodadas foram analisados quanto as proporções de quatro medidas em suas proteínas: energia potencial, área da superfície exposta ao solvente, área da superfície exposta correspondente a resíduos hidrofóbicos, e distância média quadrada (RMS) entre os átomos da conformação gerada e os da conformação nativa. Um método de análise que se mostrou de grande valia foi simplesmente selecionar subgrupos de configurações com valores abaixo de um limite para cada medida, e verificar como se distribuíram entre os clusters. Para os testes realizados, o limite arbitrado para determinada variável foi o valor desta variável alto o suficiente apenas para separar uma amostra contendo um décimo da população. Várias rodadas de clusterização foram realizadas com inicialização aleatória, e em todas as rodadas é notória a correspondência entre as proteínas pertencentes a um cluster em termos de baixo RMS e os atributos medidos nesta dissertação, a saber, energia, superfície total e superfície hidrofóbica. Ou seja, clusters com proteínas cujas estruturas tendem a possuir pequena distância RMS também são aqueles clusters onde as proteínas tendem a possuir baixa energia, pequena superfície total e pequena superfície hidrofóbica. Apesar deste resultado atestar a factibilidade e bons resultados da abordagem, deve-se considerar que os resultados para estruturas menores são mais conclusivos neste sentido.

Talvez o mais importante resultado seja o fato de que, dos três atributos considerados, baixos valores de superfície hidrofóbica são melhores indicadores de que se está em um cluster com proteínas de baixo RMS do que baixos valores de energia. A energia potencial continua sendo um parâmetro importante para a determinação da conformação nativa, mas no ambiente natural a energia compete com a superfície hidrofóbica exposta ao solvente, e se cria uma balanço entre variação de energia e entropia. Devido a este efeito, muitos métodos de dobramento de proteínas que não levam em consideração a água acabam por minimizar a proteína além da barreira imposta pela redução da entropia causada, por exemplo, por torções estruturais que exponham resíduos hidrofóbicos ao solvente. Métodos como a Dinâmica Molecular

(DM) podem ser realizados em água e são capazes de simular o efeito entrópico sob o custo extra de simular uma caixa de água ao redor da proteína. O método MC-RNA conseguiu, através da inserção da medida de superfície hidrofóbica na determinação de clusters candidatos, simular o efeito hidrofóbico sem que isto acarretasse em aumento do grau de complexidade do algoritmo. O efeito da água parece ter sido extremamente bem modelado através da superfície hidrofóbica, pois em praticamente todas as rodadas o critério da superfície hidrofóbica indica o cluster com o menor RMS mesmo para as proteínas maiores.

Procurou-se na dissertação considerar diferentes casos de proteínas de estudo. Assim, analisou-se tanto casos compostos somente de estruturas secundárias alfa-hélice, quanto misturas das duas: α-hélices e folhas-β. Proteínas compostas por folhas-β apresentam maior grau de dificuldade para a determinação da estrutura nativa devido basicamente a duas peculiaridades: (1) as características não locais da estrutura secundária, que permitem formar pontes de hidrogênio com resíduos distantes na seqüência de aminoácidos, e são difíceis de prever para métodos que se baseiam majoritariamente em informações locais como as RNAs, e (2) a alta sensibilidade a variações nos ângulos diedrais dos resíduos pertencentes às alças entre folhas-β. Para a energia de uma folha-β composta por dois ou mais segmentos ser minimizada os ângulo diedrais de dois ou três resíduos que participem da alça *coil* que liga cada par de segmentos da folha devem ser tais que os segmentos fiquem paralelos e próximos o suficiente para permitir a formação de pontes de hidrogênio. Métodos Monte Carlo possuem um elevado nível de aleatoriedade nas estruturas geradas. Um pequeno erro na distância e orientação entre as estruturas e o casamento das folhas-beta não ocorre, mesmo que a previsão da estrutura secundária seja perfeita. Ou seja, a sensibilidade da energia potencial em relação aos ângulos diedrais nas alças (estruturas *coil*) é especialmente grande para folhas beta. Para a formação de folhas-beta completas através da formação de pontes de hidrogênio, que possuem uma energia potencial relativamente alta, a movimentação das estruturas pode requerer saltos entre mínimos locais separados por picos de energia. Já a estrutura α-hélice possui as pontes de hidrogênio locais, internas à estrutura, e a interação com as demais estruturas da proteína se dá principalmente por interações de Van der Waal cujas contribuições energética são sensivelmente menores e mais suaves. Os fenômenos descritos acima explicam a dificuldade dos algoritmo MC-RNA e mesmo do MC-DSSP em formar folhas-β apesar de terem obtido resultados bons em termos de RMS e mesmo visuais, como para com a proteína $1kkg$ que obteve a estrutura global tridimensional em forma de espiral com os dois métodos MC-RNA e MC-DSSP, e como para a proteína $1i74$ que atingiu a estrutura espacial e a seqüencia de resíduos correta para o método MC-DSSP, e principalmente para a proteína $1j8b$ one o MC-RNA aproximou a estrutura tridimensional global e o método MC-DSSP não formou a folha-β interna por pequenas imprecisões nos ângulos dos resíduos das alças. As estruturas resultantes de ambos os métodos MC-RNA e MC-DSSP aplicados à proteína globular $1g7d$ são conformações que aliam baixa energia potencial à pequena superfície hidrofóbica, e como para a proteína $1j8b$ não se atingiu a conformação nativa por defeitos nos ângulos dos resíduos nos segmentos *coil* As conformações obtidas estão próximas o suficiente da conformação nativa de cada proteína para permitir que a aplicação de um algoritmo de busca como a DM possam partir destes pontos sub-ótimos e levar adiante o processo de dobramento da proteína em tempo reduzido Com o método de identificação de clusters com

concentração de baixos valores para a energia, superfície total e superfície hidrofóbica, o método MC-RNA pode prover, a título de trabalhos futuros, um conjunto de conformações inicias carregados de informação indicando o caminho provável para o dobramento de proteínas. Métodos caros computacionalmente como a DM em vácuo ou em solvente podem então utilizar o cluster selecionado conforme os critérios de energia e superfície hidrofóbica exposta ao solvente do método MC-RNA e utilizar as conformações do cluster como um grupo de condições iniciais pré otimizado

Por fim, a comparação entre as conformações de menor RMS obtidas por cada método mostrou indubitavelmente a importância da informação sobre a estrutura secundária e o seu grande efeito na aproximação da estrutura terciária nativa por métodos de busca como o Monte Carlo. O método MC-RNA obteve em todos os testes a estrutura mais próxima da conformação nativa em relação ao método MC, e também foi suplantado pelo método MC-DSSP em todos as simulações . Se esta regra for aplicável aos demais métodos de determinação de estrutura terciária, sempre que surgir um método de treinamento de RNAs para prever a estrutura secundária que logre avanços na taxa de acerto, este método poderá estar apto a melhorar a performance de técnicas de dobramento de proteínas.

Bibliografia

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Meyers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215, Issue 3:403–410, 1990.

[AL89] N.L. Allinger and J.-H. Lii. Molecular mechanics. the mm3 force field for hydrocarbons vibrational frequencies and thermodynamics. *J.Am.Chem. Soc*, 111:8566–8575, 1989.

[AMS+97] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Ressearch*, 25(17):3389–3402, 1997.

[Anf93] Christian B. Anfinsen. Nobel lecture: Studies on the principles that govern the folding of protein chains. In Sture Forsén (Lund University), editor, *Nobel Lectures, Chemistry 1971-1980*, pages 55–71. World Scientific Publishing Co., 1993.

[AYL89a] N. L. Allinger, Y. H. Yuh, and J.-H. Lii. Molecular mechanics. the mm3 force field for hydrocarbons. *J.Am.Chem. Soc*, 111:8551–8566, 1989.

[AYL89b] N.L. Allinger, Y. H. Yuh, and J.-H. Lii. Molecular mechanics. the mm3 force field for hydrocarbons. the van der waals' potentials and crystal data for aliphatic and aromatic hydrocarbons. *J.Am.Chem. Soc*, 111:8576–8582, 1989.

[BA91] J. P. Bowen and N. L. Allinger. *Reviews in Computational Chemistry*, volume 2. Verlag Chemie Publishers, New York, 1991.

[BB01] P. Bald and S. Brunak. *Bioinformatics: the machine learning approach*. MIT Press, Cambridge, MA, USA, 2001.

[BBB+90] H. Bohr, J. Bohr, S. Brunak, R. M J. Cotterill, H. Fredholm, B. Lautrup, and S. B. Petersen. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Letters*, 261:43–46, 1990.

[BBO+83] B. R. Brooks, R. E. Broccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[BH65] G. H. Ball and D. J. Hall. Isodata, a novel method of data analysis and classification. Technical report, Stanford University, Stanford, CA, 1965.

[BK00] Oliver Bieri and Thomas Kiefhaber. Kinetic models in proteing folding. In B. D. Hames and D. M. Glover, editors, *Mechanisms of Protein Folding*, chapter 2. Oxford University Press, 2 edition, 2000.

[BMHB03] K. Balali-Mood, T. A. Harroun, and J. P. Bradshaw. Molecular dynamics simulations of a mixed dopc/dopg bilayer. In *EPJ E '03: Proceedings of the 2nd International Workshop on Dynamics in Confinement*, pages S135–S140. Eur. Phys. J., 2003.

[BO27] M. Born and J. R. Oppenheimer. *Ann. Physik*, 84:457, 1927.

[CF74] P. Y. Chou and G. D. Fasman. Prediction of protein conformation. *Biochemistry*, 13, Issue 2:222–245, 1974.

[CLZ+01] Zhongqiang Chen, Qi Liu, Yisheng Zhu, Yixue Li, and Yuhong Xu. A hydrophobicity based neural network method for predicting transmembrane segments in protein sequences. In *EMBS '01: Proceedings of the 23rd Annual EMBS International Conference*, pages 2899–2902, Shangai, China, 2001. Department of Biomedical Engineering. Shanghai Jiaotong University.

[Con83] M. L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, Oct 1983.

[DBS04] Saravanan Dayalan, Savitri Bevinakoppa, and Heiko Schroder. A dihedral angle database of short sub-sequences for protein structure prediction. In *CRPIT '29: Proceedings of the second conference on Asia-Pacific bioinformatics*, pages 131–137, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[FA01] Yukiko Fujiwara and Minoru Asogawa. Prediction of subcellular localizations using amino acid composition and order. *Genome Informatics*, 12:103–112, 2001.

[FPP95] Steve Fairchild, Ruth Pachter, and Ronald Perrin. Protein structure analysis and prediction. *The Mathematica Journal*, 5, 1995.

[GRB96] K. Gunasekaran, C. Ramakrishnan, and P. Balaram. Disallowed ramachandran conformations of amino acid residues in protein structures. *Journal of Molecular Biology*, 264:191–198, 1996.

[GRG91] J.-F. Gibrat, B. Robson, and J. Garnier. Influence of the local amino acid sequence upon the zones of the torsional angles [phi] and [psi] adopted by residues in proteins. *BIOCHEMISTRY*, 30:1578–1586, 1991.

[GS91] K. D. Gibson and H. A. Scheraga. *J. Biomole Struct Dyn.*, 8:1109, 1991.

[Hay01] Simon Haykin. *Redes Neurais: princípios e prática*. Bookman, Porto Alegre, 2 edition, 2001.

[HBvGP84] J. Hermans, H. J. C. Berendsen, W. F. van Gunsteren, and J. P. M. Postma. A consistent empirical potential for water-protein interactions. *Biopolymers*, 23:1, 1984.

[HK89] L. Howard Holley and Martin Karplus. Protein secondary structure prediction with a neural network. In *PNAS*, volume 86, pages 152–156, 1989.

[HMK95] S. R. Holbrook, S. M. Muskal, and S. H. Kim. *Predicting protein structural features with artificial neural networks*, pages 161–194. AAAI Press, Menlo Park, 1995.

[HTdW95] A. Hertz, E. Taillard, and D. de Werra. A tutorial on tabu search. In *Proc. of Giornate di Lavoro AIRO'95 (Enterprise Systems:Management of Technological and Organizational Changes)*, pages 13–24, Italy, 1995.

[HW02] Ji-Tao Huang and Ming-Tao Wang. Secondary structural wobble: the limits of protein prediction accuracy. *Biochemical and Biophysical Research Communications*, 294:621–625, 2002.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[JMTR96] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc*, 118:11225–11236, 1996.

[Jon99] David T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292:195–202, 1999.

[KGV83] S. Kirkpatrick, D. C. Gellat, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671, 1983.

[Kin67] B. King. Step-wise clustering procedures. *J. Am. Stat. Assoc.*, 69:86–101, 1967.

[KS83a] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[KS83b] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–637, Dec 1983.

[KS96] R. D. King and M.J.E. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*, 5:2298–2310, 1996.

[LCN00] Albert L. LEHNINGER, Michael M. COX, and David L. NELSON. *Principles of biochemistry*. Worth, New York, 3rd. edition, 2000.

[Lev97] J. M. Levin. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Engineering*, 10:771–776, 1997.

[Mat75] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. In *Biochim Biophys Acta*, volume 405, pages 442–51, 1975.

[MBMP02] M. A. Moret, P. M. Bisch, K. C. Mundim, and P. G. Pascutti. New stochastic strategy to analyze helix folding. *Biophysics Journal*, 82:1123–1132, 2002.

[MHA95] P. K. Mehta, J. Heringa, and P. Argos. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Science*, 4:2517–2525, 1995.

[MMBS75] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides VII, geometric parameters, partial charges, non-bonded interactions, hydrogen bond interactions and intrinsic torsional potentials for naturally ocurring amino acids. *J. Phys. Chem.*, 79:2361–2381, 1975.

[NB99] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Clarendon Press - Oxford, Oxford, New York, 1999.

[QS88] Ning Qian and Terrence J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202:865–884, 1988.

[Ros96] Burkhard Rost. Phd: predicting one-dimensional protein structure by profile based neural networks. In *Computer Methods for Macromolecular Sequence Analysis*, volume 266, pages 525–539, 1996.

[Ros99] Burkhard Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 85-94:12, 1999.

[Ros01] Burkhard Rost. Protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134:204–218, 2001.

[RS93] Burkhard Rost and Chris Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–599, 1993.

[Sch02] Tamar Schlick. *Molecular Modeling and Simulation: an interdisciplinary guide.* Springer, New York, 2002.

[SNS84] M. J. Sippl, G. Nemethy, and H. A. Scheraga. Intermolecular potentials for crystal data 6. determination of empirical potentials for O-H—O=C hydrogen bonds for packing configurations. *J. Phys. Chem*, 88:6231–6633, 1984.

[SR04] Armando D. Solis and S. Rackovsky. On the use of secondary structure in protein structure prediction: a bioinformatic analysis. *Polymer*, 45:525–546, 2004.

[SS73] P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy*. Freeman, London, UK, 1973.

[SS91] C Sander and R Schneider. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[SS95] Asaf A. Salamov and Victor V. Solovyev. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology*, 247:11–15, 1995.

[VRD+01] S. Vucetic, P. Radivojac, A. K. Dunker, C. J. Brown, and Z. Obradovic. Methods for improving protein disorder prediction. In *International Joint INNS-IEEE Conference on Neural Networks*, volume 4, pages 3030–3034, Orlando, Florida, U.S.A., 2001.

[War63] J. H. Jr. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58:236–244, 1963.

[WK81] P. K. Weiner and P. A. Kollman. Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. *J. Comp. Chem*, 2:287–303, 1981.

[WM00] c. H. Wu and J. W. McLarty. Neural networks and genome informatics. *Methods in Computational Biology ans Biochemistry* 1, 2000.

[Wu96] Cathy H. Wu. Artificial neural networks for molecular sequence analysis. *Computers & Chemistry*, 21:237–256, 1996.

[YL93] Tau-Mu Yi and Eric S. Lander. Protein secondary structure prediction using nearest-neighbor methods. *Journal of Molecular Biology*, 232:1117–1129, 1993.

[YY01] Ikuo Yoshihara and Yoshiyuki Kamimai Moritoshi Yasunaga. Feature extraction from genome sequence using multi-modal neural networks. In *Proc. Genome Informatics 2001*, volume Genome Informatics Series No.12, pages 420–422. Universal Academic Press, 2001.