UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO” - UNESP.

LESSANDRA ELLER

SIMULAÇÃO DO ENOVELAMENTO DE PROTEÍNAS DESNATURADAS UTILIZANDO O MÉTODO DE CRESCIMENTO DE CADEIAS EM REDE TETRAÉDRICA

Tese apresentada para obtenção do grau de Doutor em Biofísica Molecular, Área de Concentração em Biofísica Molecular. Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho” - UNESP.

ORIENTADOR: PROF. DR. MARCO ANTONIO ALVES DA SILVA

São José do Rio Preto - SP

2010

Eller, Lessandra.
Simulação do Enovelamento de Proteínas Desnaturadas Utilizando o Crescimento de Cadeias em Rede Tetraédrica./ Lessandra Eller. - São José do Rio Preto : [s.n.], 2010.
67 f. : 36 il. ; 30 cm.

Orientador: Marco Antonio Alves da Silva
Tese (doutorado) Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas

1. Enovelamento proteico. 2. Crescimento de Cadeias. 3. Rede Tetraédrica. 4. Simulação Computacional. I. daSilva, Marco Antonio Alves. II. Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas. III.Título

CDU -

Resumo

Neste trabalho geramos configurações de proteínas desnaturadas utilizando o modelo de crescimento de cadeia nascente numa rede tetraédrica (ou de diamante) infinita e verificamos a influência das restrições estéricas (interação de esfera rígida) e das armadilhas topológicas no processo de crescimento e determinação estrutural ou conformacional das cadeias geradas. Os procedimentos, resultados, discussões e conclusões do atual estudo são expostos em três capítulos um tanto autosuficientes e descritos sucintamente abaixo.

No primeiro capítulo estabelecemos as bases do atual trabalho, que é baseado em gerar conformações de proteínas desnaturadas por um algoritmo de crescimento de cadeia em rede de diamante com resíduos unidos por forças covalentes e interagindo somente por restrições estéricas. Para este fim partirmos de aspectos globais para especifícos. Para contextualizar o modelo de crescimento de cadeia iniciamos da relação entre DNA e RNA para o crescimento e a síntese de proteína; para a rede de diamante (ou tetraédrica) iniciamos das estruturas cristalinas e redes de Bravais até as redes cúbicas ; para a formação e interação do polímero apresentamos uma exposição ampla, mas de forma resumida desde as interações covalentes e não covalentes até as restrições estéricas.

No segundo capítulo investigamos o papel das restrições estéricas (interações de esfera rígida) em polímeros desnaturados usando um modelo minimalista de algoritmo de crescimento de cadeia numa rede tetraédrica segundo três abordagens: cadeia ideal (**ci**, desprezando os impedimentos estéricos), cadeia com restrição estérica de curto alcance (**ca**, sobreposição proibida entre primeiros vizinhos ao longo da sequência de resíduos, i.e., *i* com *i*-2) e cadeia com restrição estérica de longo alcance (**la**, sobreposição proibida entre quaisquer resíduos) .

Os resultados da simulação mostram que somente a restrição estérica entre primeiros vizinhos é suficientemente eficiente para conduzir a conformação da cadeia (como expresso pela compactação e números de sobreposição) próximo à estruturas reais (cadeias sem sobreposição) e mais afastado de casos menos realístas (cadeia ideal). Tais resultados também revelam a importância das interações de curto alcance e fornecem entendimentos básicos que podem ser proveitosos em várias questões, como os mecanismos do *folding* de proteínas e a formação de estruturas secundárias.

No terceiro capítulo as restrições estéricas totais (**la**) são empregadas em associação com duas abordagens estratégicas para analisar as implicações de configurações proibidas, ou armadilhas topológicas, na compactação da cadeia e tempo de simulação. Pelo método SAWm1 inicia o crescimento da cadeia do primeiro dos quatro resíduos inserido na cadeia e em sobreposição com o resíduo *n*, por outro lado, no método SAWm2 quando encontrado uma armadilha topológica, no resíduo *n* (com outros quatros resíduos), uma nova cadeia é iniciada do começo (do quarto resíduo).

Uma análise comparativa entre SAWm1 e SAWm2 inicialmente pelas armadilhas topológicas através de três dentre quatro grandezas diferentes, mostram que SAWm1 é mais eficiente do que SAWm2, pois enquanto no primeiro a variação destas grandezas em função do número de resíduos tem ajustes logarítmicos, lineares e polinomiais quadráticos, no segundo são tipo log normal, polinomiais quadráticos e exponenciais, respectivamente.

Por outro lado, quanto a compactação, não é possível identificar diferenças significativas nas conformações geradas pelos dois métodos. Pelo tempo de simulação, representado pelo tempo de CPU, em função do comprimento *N* das cadeias nascentes, novamente SAWm1 parece mais vantajoso sobre SAWm2, pois são ajustados por um polinômio de segunda ordem e crescimento exponencial, respectivamente. Portanto pelo menos para cadeias longas ($N \gtrsim 1000$ resíduos) e pelas grandezas utilizadas, parece razoável propormos que o método de retificar armadilhas topológicas de SAWm1 possa substituir SAWm2, além disso agora conhecemos alguns comportamentos e limites dos dois métodos o que facilita implementações adicionais futuras em ambos.

Abstract

In this work we generate configurations of denatured proteins using the growth model of nascent chain on the infinite tetrahedral (or diamond) lattice and ascertain the influence of the steric constrains (hard-core interaction) and the topologic traps in the growth process and structural or conformational determination of the generated chains. The procedures, results, discussions and conclusions of the present study are displayed in three chapter rather self-sufficient and shortly described below.

In the first chapter we establish the bases of the current work, that is focused on generate denatured proteins conformations by a diamond lattice chain growth algorithm with the residues linked by covalent bonds and interplaying only by steric constraints. For this aim we take up from global to specific features. To contextualize the chain growth model we begin from relation between DNA and RNA to the growth and protein synthesis; for diamond (or tetrahedral) lattice we start from crystalline structures and Bravais lattices to cubic ones; for the formation and interaction of the polymer we present a broad explanation, but of brief form since the covalent and noncovalent interactions until the steric constraints.

In the second chapter we investigate the role of steric constrains (hard-core interaction) in denatured polimers using a minimalist model of chain growth algorithm in a tetrahedral lattice under three approaches: ideal chain (**ic**, neglecting steric constraints), chain with steric constraint of short range (**sr**, forbidden overlap between first neighbors along the residue sequence, i.e., i with i - 2) and chain with steric constraint of long range (**lr**, forbidden overlap among whatever residues).

The simulation results show that only the steric constraint between first neighbors is efficient enough to drive the chain conformational (as express by compactness and overlap number) near of real structures (non-overlapping chains) and farther of less realistic case (ideal chain). Such results also reveal the importance of the short-range interactions and supply underlying understanding that can be helpful to several subjects as the mechanisms of protein folding and the formation of secondary structures.

In the third chapter the full steric constrains (**lr**) are employed in association with two strategic approachs to analyze the implications of forbidden configurations, or topologic traps, in the chain compactness and simulation time. By SAWm1 method begins the chain growth from the first of the four residues inserted on the chain and on overlap with the residue n, on the other hand the SAWm2 method once encountered a topologic trap, in the residue n (with other four residues), a new chain is begun from the outset (of the forth residue).

A comparative analysis between the SAWm1 and SAWm2 initially by topologic traps through 3 out 4 different variables show that SAWm1 is more efficient than SAWm2, because while in the first the change of these variable on the chain length have logarithmic, linear and polynomial fit, in the second are log normal, polynomial and exponential ones, respectively.

Though, as regards compactness is not possible to discern meaningful differences in the generated conformations by both methods. By simulation time, depicted by CPU time, as function of the length N of nascent chains, newly SAWm1 seems more useful than SAWm2, because they are fitted by a second order polynomial and growing exponentially, respectively. Therefore, at least for long chains ($N \gtrsim 1000$ residues) and by the variables employed seem reasonable to presume that the method to remove topologic traps of SAWm1 can replace SAWm2, moreover now we know some behaviors and ranges of the two methods what ease more future implementations in both.

Conteúdo

Capítulo 1: Crescimento de cadeia nascente, rede de diamante e interação de esfera rígida
1.1 Introdução 1
1.2 Processo de crescimento de cadeia 3
1.2.1 Gene e cadeia polipeptídica 3
1.2.2 O crescimento da cadeia polipeptídica 4
1.2.2.1 Transcrição: do DNA ao mRNA 4
1.2.2.2 Translação: do mRNA à cadeia polipeptídica 5
1.3 Estruturas cristalinas 7
1.3.1 Propriedades dos sólidos cristalinos e células unitárias paralelepipédica 8
1.3.2 Redes de Bravais 8
1.3.3 Redes cúbicas 11
1.3.4 Rede de diamante 12
1.4 Forças em estruturas de biomoléculas (proteínas) 14
1.4.1 Interações covalentes 14
1.4.2 Interações não covalentes 15
1.4.2.1 Eletrostática 16
1.4.2.1.1 Ligações Iônicas 18
1.4.2.1.2 Ligações de Hidrogênio 18
1.4.2.1.3 Forças de Van der Waals 19
1.4.2.2 Interações Hidrofóbicas 20
1.4.2.3 Forças repulsivas de Van der Waals 21
1.4.3 Argumentos adicionais a favor da utilização apenas da força repulsiva de VDW 23
1.5 Comentários Finais 24

Capítulo 2: Implicações das restrições estéricas para o crescimento e enovelamento de cadeias peptídicas nascentes
2.1 Introdução 27
2.2 Modelo e métodos 28
2.2.1 Descrição do modelo 28
2.2.1.1 Início do crescimento da cadeia 29
2.2.1.2 Continuando o crescimento da cadeia 30
2.2.2 Cadeias nascentes com diferentes níveis de restrições estéricas 32
2.2.2.1 Cadeia Ideal (**ci**) 33
2.2.2.2 Cadeia com restrições estéricas de curto alcance (**ca**) 33
2.2.2.3 Cadeia com restrições estéricas de longo alcance (**la**) 34
2.2.2.3 Medidas de aspectos conformacionais 35
2.2.3.1 Grandezas empregadas nas simulações MC 35
2.2.3.2 Inicialização de grandezas 36
2.3 Resultados e discussões 37
2.4 Conclusões 46

Capítulo 3: Influências das armadilhas topológicas na compactação e tempo de enovelamento de cadeias nascentes longas em rede tetraédrica
3.1 Introdução 48
3.2 Métodos 49
3.3 Resultados e discussões 50
3.3.1 Armadilhas topológicas 51
3.3.1.1 Frequência das Armadilhas topológicas 51
3.3.1.2 Distribuição do número de armadilhas 53

3.3.1.3 Número máximo de armadilhas com frequência não nula 56
3.3.1.4 Número total de armadilhas 56
3.3.2 Influência das armadilhas topológicas nas compactações das cadeias 58
3.3.2.1 Variações dos raios de giração médios, nos métodos SAWm1 e SAWm2 58
3.3.2.2 Determinação do expoente de escalonamento para SAWm1 e SAWm2 60
3.3.3 Tempo de processamento para SAWm1 e SAWm2 61
3.4 Conclusões 62

4 Referências 65

Lista de Símbolos

A: nucleobase (ou base de nucleotídeo / base nitrogenada) adenina
A_{rm}(iniciais de Armadilha): medida da ocorrência (A_{rm}=1) e não ocorrência (A_{rm} = 0) de armadilha topológica
ADP (iniciais em inglês de *Adenosine diphosphate*): Adenosina difosfato
AMP (iniciais em inglês de *Adenosine monophosphate*): Adenosina monofosfato
ATP (iniciais em inglês de *Adenosine triphosphate*): Adenosina trifosfato
C: nucleobase citosina
C: base centrada, ponto adicional no centro de cada base
ca (iniciais de curto alcance): cadeias nascentes com restrição estérica de curto alcance
CCC (ou I): cúbica de corpo centrado
CFC (ou F): cúbica de face centrada
ci (iniciais de cadeia ideal): cadeias nascentes ideal
CS (ou P): cúbica simples
D_{PP}: Distância ponta a ponta média
DNA: ácido desoxirribonucléico
e.g.: do Latin *exempli gratia* (ou por exemplo)
et al.: abreviação da palavra Latina *et alii* (masculino), *et aliae* (feminino) e *et alia* (neutro) indicando genericamente "e outras pessoas"
F: face centrada, ponto adicional no centro de cada face
fA_{rm} : frequência das armadilhas topológicas
fD_{PP}: frequência da distância ponta a ponta
$\%fNA_{rm}$: percentagem da distribuição do número de armadilha normalizada
fR_G: frequência do raio de giração
G: nucleobase guanina
I: corpo centrado, ponto adicional no centro da célula
i.e.: abreviação da palavra Latina *id est*, que significa "isto é"
k: constante de proporcionalidade entre o raio de giração e o número de resíduos
la (iniciais de longo alcance): cadeias nascentes com restrição estérica de longo alcance
m: número de tentativas de crescimento de cadeias
M: números de conformações (ou cadeias) geradas
máx{NA_{rm}}: número máximo de armadilhas (NA_{rm}) com frequência não nula
MC (iniciais de Monte Carlo): método de simulação Monte Carlo
mRNA: RNA mensageiro
N: número de resíduos (ou comprimento) das cadeias nascentes
: expoente de escalonamento para o raio de giração em função do número de resíduos
NA_{rm}: iniciais de número de armadilha
nd : quantidade de distâncias não repetidas numa conformação
np: número de pares primeiro vizinhos
nr: resíduo armadilhado
: número de coordenação de uma rede cristalina
P: primitiva, célula unitária simples
PDB (iniciais em inglês de *Protein Data Bank*): Banco de Dado de Proteínas
$\mathbf{P}_i$ e $\mathbf{N}_i$: conjunto de oito vetores para rede tetraédrica (ou de diamante) nos sentidos *i*
PMC: Passos Monte Carlo
PP: pirofosfato
R: inicial de Romboédrico
R_G: Raio de giração médio
RNA: ácido ribonucléico
SAXS (iniciais em inglês de *Small angle X-ray scattering*): espalhamento de Raio-X em baixo ângulo

SAW (iniciais em inglês de *self-avoiding walk*): caminho de auto impedimento
SAWm1: SAW modificado do tipo 1
SAWm2: SAW modificado do tipo 2
T: nucleobase timina
tRNA: RNA transportador
U: nucleobase uracil (ou uracila)
fA_{rm}: número total de armadilhas
3D: tridimensional
∞ : símbolo para "infinito"

Capítulo 1: Crescimento de cadeia nascente, rede de diamante e interação de esfera rígida

1.1 Introdução

Sistemas biológicos (como proteínas) podem ser analisados por experimentos feitos diretamente nos organismos vivos (*in vivo*); num ambiente controlado, fora dos organismos vivos (*in vitro*); num computador ou via simulação computacional (*in silico*). Muitos estudos relativos ao processo de *folding* de proteínas realizado *in vitro* e *in silico* diz respeito ao *refolding* de proteínas prontas após o *unfolding*, demonstrando que toda informação necessária para se adquirir a estrutura tridimensional funcional (estrutura terciária) da proteína está contida em sua seqüência de aminoácido (estrutura primária).

Devido muitas proteínas terem seu *refolding* bem sucedido, sob condições adequadas (meio aquoso, pH, concentração salina, etc ...), para seus estados nativos na ausência da maquinaria celular, isso implica que nenhum fator celular é necessário para especificar o estado nativo final, justificando a concepção que o processo celular possa simplesmente ser inferido do que tem se aprendido dos estudos *in vitro* [C. B. Anfinsen et al., **1961**].

Os estudos do *refolding* proteico em geral se baseiam na compreensão do que se pode obter do *folding* de proteínas na célula, contudo é conhecido que fatores e condições celulares podem fazer o *folding* na célula significativamente diferente destes estudos convencionais [R. B. Freedman, **1992**]. Dessa forma, uma outra formulação do problema do *folding* de proteínas busca entender como o estado enovelado de proteínas é influenciado pela síntese proteica gerada *in vivo*. A convicção que a conformação funcional de uma proteína está implícita na sequência de aminoácido e portanto na informação do código genético inspira a expressão bem sucedida de proteínas em células hospedeiras, como a *Escherichia coli* numa demonstração dos fundamentos dos processos biológicos. O fato que proteínas possam administrar seu próprio *folding* sob condições apropriada *in vitro* não garante sua habilidade de fazê-lo na célula, neste caso deve-se atentar para as específicas condições e fatores

celulares necessários para o *folding* bem sucedido de proteínas *in vivo*.

No caso da simulação computacional do *folding* de modelos proteicos, o *refolding* normalmente inicia-se com a cadeia proteica numa configuração distendida ou aleatória; já a simulação de crescimento de cadeia, cuja base (e inspiração) é a síntese de proteínas *in vivo*, as cadeias polipeptídicas nascentes [H.-M. Lu & J. Liang, **2008**] são recriadas num número predefinido de vezes. Há semelhanças e diferenças entre estes dois métodos nas simulações tradicionais de Monte Carlo na rede.

Neste trabalho investigamos a contribuição somente da restrição ou impedimento estérico (interação de esfera rígida) em proteínas (ou polímeros) genéricas desnaturadas em rede, ou seja, sem considerarmos as demais interações que mantêm a proteína em sua forma funcional. Dessa forma separamos as contribuições geométricas e a restrição estérica dos demais ingredientes presentes nas proteínas. Para o crescimento e enovelamento[1] das cadeias são utilizados três itens distintos (Figura 1.1): o método de crescimento de cadeias; o posicionamento dos resíduos ligados covalentemente pela rede de diamante; a interação destes resíduos apenas pela força repulsiva não covalente de esfera rígida.

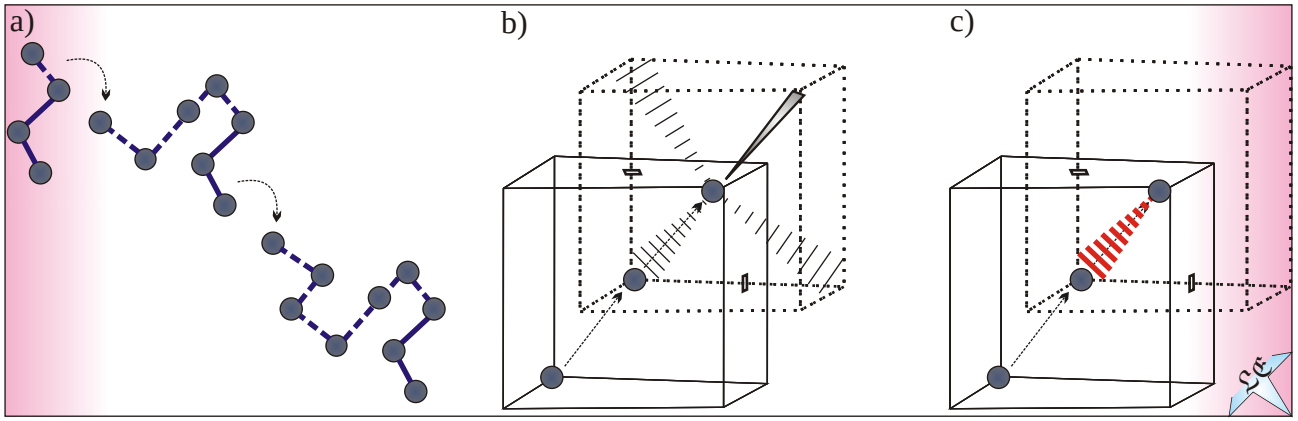


Figura 1.1 - Os três itens interdisciplinares utilizados para gerar configurações de proteínas genéricas desnaturadas são: crescimento de cadeia nascente (biologia molecular); rede de diamante (física de estado sólido); e interação de esfera rígida (física de polímeros). (a) O processo de crescimento de cadeia tem inspiração na incorporação dos aminoácidos nas proteínas; (b) o crescimento ocorre em sítios de uma rede de diamante, formada a partir de redes cúbicas de face centrada; (c) já as consequências da conectividade (interação covalente) e da restrição ou impedimento estérico (interação não covalente) entre resíduos será avaliada por todo este trabalho.

A seguir são analisadas suscintamente e sempre de aspectos gerais para específicos, como ocorre o crescimento de cadeias polipeptídicas nascentes de proteínas no ribossomo (seção 1.2); a geometria da rede de diamante, originada da rede de Bravais, cúbica de face centrada (seção 1.3); e por fim as forças covalentes e não covalentes em estruturas de biomoléculas (seção 1.4).

[1] Neste trabalho usamos significados diferentes para os termos *folding* (mantido do inglês) e enovelamento, pois preservamos *folding* para proteínas reais e enovelamento para cadeias genéricas desnaturadas, assim sem a relação restrita com as proteínas.

1.2 Processo de crescimento de cadeia

Nesta seção descrevemos os passos do primeiro código genético (H. S. Chan & K. A. Dill, **1993**), ou seja, como se dá a correspondência entre a sequência de bases na molécula de DNA e a sequência de aminoácidos da proteína, cuja síntese é controlada pelo DNA. A correspondência entre a sequência de resíduos de aminoácidos e a estrutura nativa da proteína, às vezes é referida como o segundo código genético, o que não é analisado neste trabalho.

1.2.1 Gene e cadeia polipeptídica

Na relação entre gene e síntese de proteína o ribossomo desempenha um papel fundamental. Aliás, os ribossomos são essenciais para a vida, já que todos os organismos vivos nas mais diferentes formas de vida na Terra tem ribossomos, em dois tamanhos básicos [D.L. Nelson & M.M.Cox, **2005**]. As bactérias e archaebactérias tem ribossomos menores, denominados ribossomos 70S. Já os ribossomos dos eucarióticos (animais, plantas e fungos) são maiores, denominados 80S.

Devido as diferenças entre nossos ribossomos e das bactérias, há muitos antibióticos que especificamente atacam os ribossomos 70S, como é o caso do *chloramphenicol* que bloqueia a reação que adiciona aminoácidos para as proteínas nascentes nas bactérias.

Em 2000, foi disponibilizada no Banco de Dados de Proteínas, PDB [H. M. Berman et al., **2000**] as primeiras estruturas das subunidades ribossomais, pelos biologistas estruturais V. Ramakrishnan, T. A. Steitz e A. E. Yonath, e por este trabalho eles receberam, em 2009, o prêmio Nobel em Química.

Ainda na relação gene e proteína [T. L. Blundell & K. Mizuguchi, **2000**], apenas recentemente foi possível fazer um mapa completo do funcionamento de um organismo e assim entender todos os processos bioquímicos que mantêm um ser vivo em funcionamento, o que abre caminho para uma meta acalentada: a criação de organismos sob medida em laboratório [C.Lartigue et al., **2009**]. Para esse fim foi escolhida a bactéria *Mycoplasma pneumoniae*, causadora de um tipo de pneumonia em seres humanos e que só sobrevive como parasita em outras células. Esta bactéria é uma das mais simples do planeta, possuindo meros 689 genes que codificam proteínas (com 816.395 “letras” de DNA, ou

nucleotídeos, enquanto que o humano tem 3 bilhões de nucleotídeos).

O mapeamento do genoma da *M. Pneumoniae* mostra que ela tem um genoma enxuto, mas com um metabolismo complexo (cujo alguns eventos são resumidos na Figura 1.2), usando um sistema flexível para transmitir os dados codificados em seu DNA para o mRNA, ou RNA mensageiro. O mRNA e o ribossomo iniciam o processo de produção de proteínas na célula (ver detalhes na seção 1.2.2), produzindo eficazmente apenas 689 proteínas, contra cerca de 3.000 diferentes proteínas na bactéria *E. Coli* e dezenas de milhares nos seres humanos.

A catalogação de todos os componentes e suas interações mostrou que as 689 proteínas servem como partes componentes de cerca de 200 máquinas diferentes, o chamado repertório proteico ou proteoma, cada um projetado para apoiar sua sobrevivência num ambiente distinto, de acordo com as condições externas.

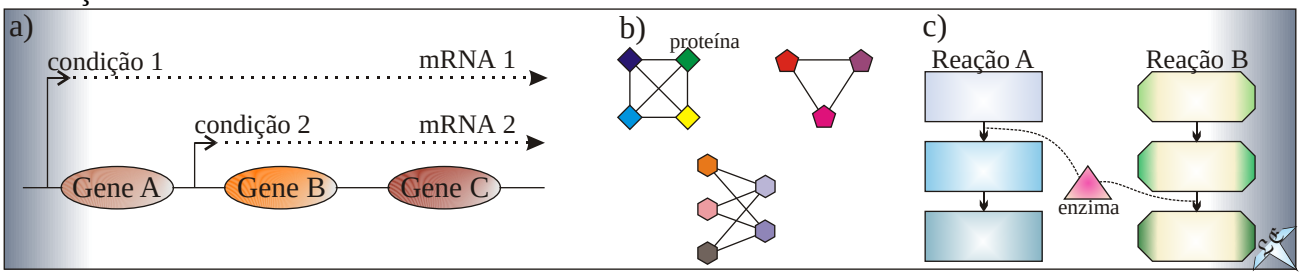


Figura 1.2 - O genoma da *M. Pneumoniae* é enxuto, mas seu metabolismo é complexo. a) Dependendo das necessidades, a bactéria usa um sistema flexível para transmitir dados do DNA para o mRNA, com genes (denominados operons) que normalmente seriam "ativados" juntos, mas nela ficam desacoplados; b) Proteínas formam redes dinâmicas de interação, facilitando o metabolismo da bactéria; c) Enzimas (ou proteínas que aceleram reações químicas na célula) podem participar de mais de uma reação.

Somente recentemente a partir do genoma sintético da bactéria *M. Mycoides* foi possível criar pela primeira vez em laboratório um organismo vivo [D. G. Gibson et al., **2010**].

1.2.2 O crescimento da cadeia polipeptídica

Os eventos que ocorrem para a fabricação de uma proteína são basicamente os seguintes: a partir da sequência do DNA no gene são transcritas as moléculas de mRNA, que por sua vez no ribossomo traduz a cadeia polipeptídica ou proteica [C. K. Matheus et al., **2000**]. Tais processos são descritos suscintamente, a seguir.

1.2.2.1 Transcrição: do DNA ao mRNA

No DNA há uma tira dupla com quatro tipos de nucleotídeos (com as bases adenina, A; citosina, C; guanina, G; e timina, T (substituída pela uracil, U, no RNA)). O mRNA é sintetizado a partir de um segmento de uma das tiras do DNA (Figura 1.3a). Os tripletes de nucleotídeos (ou códons) no mRNA, formam 4^3 ou 64 diferentes combinações possíveis. Cada uma destas combinações está associada a um dos 20 aminoácidos (com múltiplos tripletes, exceto na Trp (UGG) e na Met, AUG) e também aos sinais "INÍCIO" (códon AUG) para iniciar a construção da cadeia polipeptídica no amino-terminal, e "FIM" (códons UAA, UAG ou UGA) para encerrar a tradução no carboxil-terminal da cadeia (Figura 1.3b).

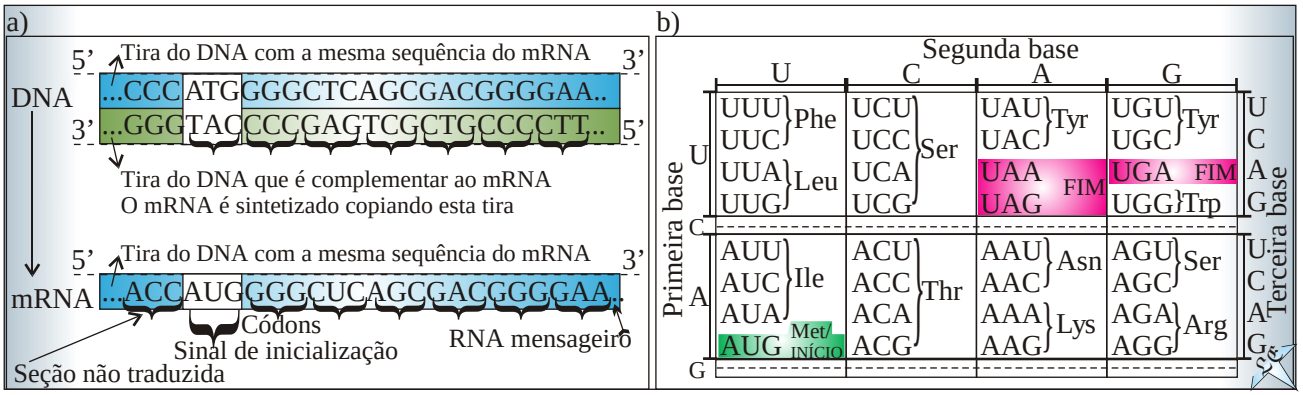


Figura 1.3 - (a) Relacionamento do DNA com o mRNA, e (b) tabela com códons. a) A tira de DNA transcrita (3' ---> 5'), é complementar ao mRNA; b) Dentre o total de 64, são mostradas apenas 32 combinações, já que de quatro linhas destacamos apenas as duas que contemplam os sinais de "INÍCIO" e "FIM". A partir das três bases dos códons do mRNA (escritos na direçaõ 5' ---> 3') é possível encontrar rapidamente quaisquer aminoácidos. Por exemplo, o aminoácido Ile correspondente ao códon 5'AUC 3' é obtido olhando sequencialmente primeiro a linha A, em seguida na coluna U, e por fim no espaço C.

A sequência particular de aminoácidos de uma cadeia polipeptídica ocorre pela tradução da informação codificada no mRNA, segundo as etapas de iniciação, alongamento e terminação, descritas a seguir.

1.2.2.2 Tradução: do mRNA à cadeia polipeptídica

Antes de ser adicionado a cadeia polipeptídica, cada aminoácido precisa ser ativado, o que é feito acoplando-o na terminação 3' de um apropriado tRNA (ou RNA transportador), e assim produzindo um aminoacyl-tRNA.

O acoplamento de cada particular aminoácido e seu apropriado tRNA é catalizado por enzimas específicas, denominada aminoacyl-tRNA sintetases, cuja fonte de energia livre para a reação é obtida da hidrólise do ATP para AMP e pirofosfato, PP (Figura 1.4a). Cada tRNA contêm uma sequência de

nucleotídeo (ou anti-códon), numa região conhecida como loop anti-códon, que é complementar ao apropriado códon na mensagem do mRNA.

Os detalhes mais importantes da tradução da mensagem genética, resumidos em quatro passos, são mostrados na Figura 1.4b. Estando o RNA mensageiro ligado a um ribossomo, o aminoacyl-tRNA também se liga a ele, um a um, associando seu anti-códon ao códon no mRNA (passo 1), e iniciando o processo de crescimento da cadeia polipeptídica.

Em seguida, passo 2, o aminoácido transportado por cada novo tRNA é transferido para a cadeia. No passo 3, o primeiro tRNA é liberado, e o ribossomo move para o próximo códon ao longo da mensagem, permitindo o próximo tRNA transferir seu aminoácido. O processo de crescimento da cadeia continua (passo 4) até o ribossomo encontrar o códon "FIM", neste ponto o fator de liberação provoca tanto a liberação da cadeia nascente, quanto o mRNA do ribossomo.

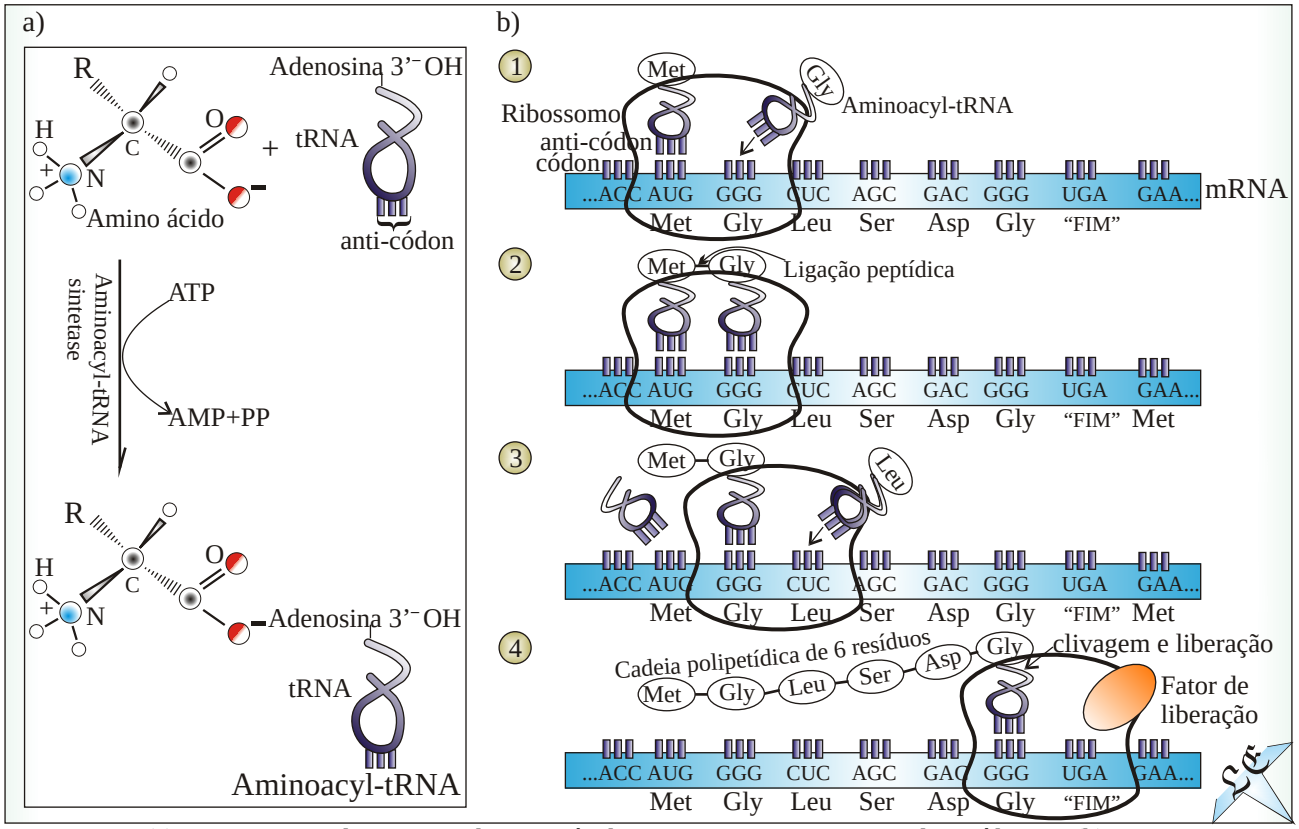


Figura 1.4 - (a) Os processos de ativação de aminoácidos para incorporação em polipeptídeos, e (b) as etapas para o crescimento da cadeia polipeptídica. a) A enzima Aminoacyl-tRNA sintetase cataliza a formação de um Aminoacyl-tRNA, com hidrólise de um ATP, e reconhece tanto o aminoácido específico, quanto o tRNA com seu anticódon. b) O ribossomo move-se ao longo da mensagem no mRNA, e emparelha o códon com o anti-códon (passo 1); um segundo aminoácido é transferido para a cadeia (passo 2); o ribossomo move para o próximo códon e adiciona o terceiro resíduo, enquanto libera o primeiro tRNA (passo 3); os passos anteriores são repetidos crescendo a cadeia até encontrar o sinal "FIM" (passo 4).

Em cada tipo de célula de todo organismo, esta notável maquinaria, mostrada nas Figuras 1.3 e

1.4, transcreve e traduz a informação codificada em milhares de diferentes genes, para milhares de diferentes cadeias polipeptídicas ou proteínas. Contudo, quando uma cadeia polipeptídica nascente é liberada de um ribossomo não necessariamente ela está enovelada em sua correta estrutura tridimensional para exercer sua atividade biológica, já que podem ocorrer várias modificações depois da tradução (ou modificações pós-traducionais) alterando sua estrutura tridimensional [M. Mann & O. N. Jensen, **2003**], e.g. com a formação de ponte cistína entre um par de resíduos de cisteínas.

Cadeias polipeptídicas nascentes também são sintetizadas em laboratórios, ou seja, numa segunda rota, diferente dos passos executados pela célula [S. Mukherji & A. van Oudenaarden, **2009**; S. Borman, **2007**]. O procedimento utilizado, no atual trabalho, segue uma terceira rota ao simular o crescimento da cadeia polipeptídica, de uma forma bastante simplificada, através de um algoritmo computacional, com cadeias crescidas numa rede de diamante, sendo tal rede uma particularidade de uma das redes de Bravais (especificamente a estrutura cristalina cúbica), descrita a seguir.

1.3 Estruturas cristalinas

A matéria no estado sólido apresenta-se sob dois estados fundamentais, de acordo com a ordenação atômica: o cristalino e o amorfo. No estado cristalino, os átomos encontram-se ordenados sobre longas distâncias atômicas formando uma rede cristalina tridimensional, construída a partir de uma simples replicação no espaço de uma estrutura elementar paralelepipédica mínima, denominada célula unitária.

No estado amorfo não existe ordem de longas distâncias na disposição dos átomos. Além disso, existem substâncias parcialmente cristalinas (como a maioria das rochas) e materiais aparentemente amorfos que podem ser cristalinos, como a areia de quartzo e as argilas. Além dos estados cristalino e amorfo, numa visão mais macroscópica há também a estrutura policristalina (Figura 1.5).

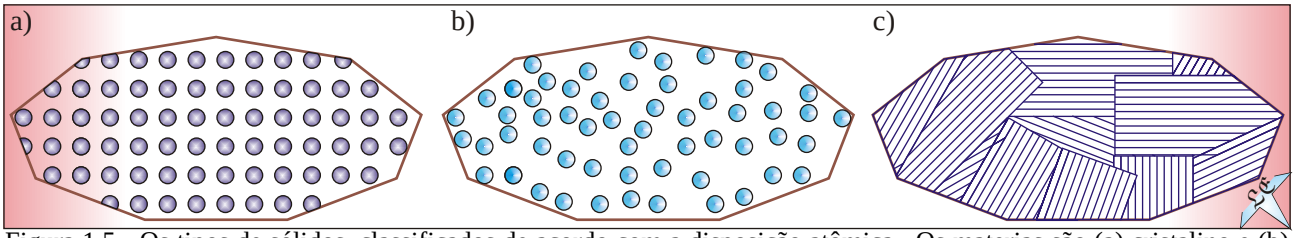


Figura 1.5 - Os tipos de sólidos, classificados de acordo com a disposição atômica. Os materias são (a) cristalino e (b) amorfo, vistos pela perspectiva microscópica dos átomos, enquanto que a estrutura (c) policristalina, é vista por uma perspectiva mais macroscópica de regiões simples cristalinas, tal qual (a).

1.3.1 Propriedades dos sólidos cristalinos e células unitárias paralelepipédicas

As propriedades dos materiais sólidos cristalinos dependem da maneira na qual suas partículas (átomos, moléculas ou íons) estão espacialmente dispostos na estrutura cristalina. Há um grande número destas estruturas cristalinas, desde estruturas simples (metais) até as mais complexas, como os cerâmicos e poliméricos.

A forma e tamanho da célula unitária de cada cristal depende das condições físicas em que o cristal se formou, das dimensões, valência química e estado de ionização das partículas que o compõem. Uma mesma substância, sob distintas condições de pressão e temperatura, pode formar cristais com células unitárias totalmente diversas, por exemplo, o carbono pode cristalizar sob centenas de formas, indo do diamante, passando pelos fulerenos e pelas inúmeras variantes da fibra de carbono até o grafite. Também as substâncias orgânicas, como os açúcares, proteínas e DNA, cristalizam em formas extremamente complexas devido seu elevado peso molecular e complexidade estrutural.

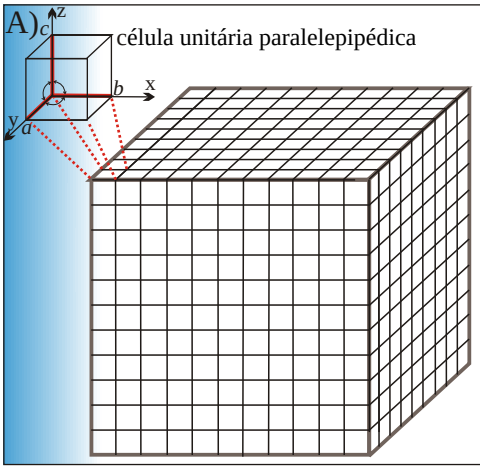
1.3.2 Redes de Bravais

A partir das células unitárias paralelepipédicas (Figura 1.6A) e levando em conta os eixos de simetria e a posição do centro geométrico de cada elemento do cristal, é possível descrever cristais pelas redes de Bravais, cuja denominação homenageia o físico Auguste Bravais (1811-1863), um pioneiro no estudo de cristalografia [M. I. Aroyo et al., **2006**]. As redes de Bravais são redes distintas constituídas de células unitárias e quando repetidas podem preencher todo o espaço. Cada rede pode, portanto, ser gerada por três vetores unitários, **a**, **b** e **c** e um conjunto de inteiros *f*, *g* e *h* de modo que cada ponto da

rede, pode ser identificado por um vetor **r**, onde **r** = f **a** + g **b** + h **c**.

Em duas dimensões (2D), há cinco distintas redes de Bravais: oblíqua, retangular, retangular centrada, hexagonal e quadrada. Em três dimensões (3D) existem catorze redes, ou configurações básicas (Figura 1.6B), visualizadas pelas magnitudes (módulos) dos vetores unitários (**a**, **b** e **c**) e pelos ângulos entre os vetores unitários (, e).

A)



B)

Sistema de cristalização	Número de redes no sistema	Símbolos das Redes	Restrições nos eixos das células e ângulos
Triclínico	1	P	$a \neq b \neq c$ $\neq \neq$
Monoclínico	2	P e C	$a \neq b \neq c$ $\neq \neq = 90^{o}$
Ortorrômbico (ou Rômbico)	4	P, C, F e I	$a \neq b \neq c$ $= = = 90^{o}$
Tetragonal	2	P e I	$a = b \neq c$ $= = = 90^{o}$
Cúbico	3	P ou CS; F ou CFC I ou CCC	$a = b = c$ $= = = 90^{o}$
Trigonal (ou Romboédrico)	1	R	$a = b = c$ $= =$ $^{o}, \neq 90^{o}$
Hexagonal	1	P	$a = b = c$ $= = 90^{o}, =$ ^{o}

Figura 1.6 - (A) Esquema de uma célula unitária paralelepipédica, e (B) os sete sistemas cristalinos, ambos com seus respectivos símbolos. A) A unidade básica repetitiva da estrutura cristalina tridimensional com suas restrições nos eixos, arestas (*a*, *b*, *c*) e ângulos (, ,) caracterizadores; B) diagramas dos sistemas cristalinos ou de cristalização ou de rede (primeira coluna), totalizando 14 redes de Bravais (soma do número de redes no sistema, segunda coluna), caracterizadas pelos símbolos (terceira coluna): Primitiva (P): célula unitária simples; Face centrada (F): ponto adicional no centro de cada face; Corpo centrado (I): ponto adicional no centro da célula; Base centrada (C): ponto adicional no centro de cada base; Romboédrico (R): Somente classe hexagonal; Cúbica simples (CS); Cúbica de corpo centrado (CCC); Cúbica de face centrada (CFC).

Os sistemas triclínico, monoclínico, tetragonal e hexagonal contêm redes também contempladas no sistema ortorrômbico. No sistema ortorrômbico há quatro tipos de redes, sendo o único com todos os tipos das redes de Bravais (Figura 1.7): uma rede tem uma célula unitária primitiva (P), com três eixos de comprimento e ângulo diferentes. A segunda rede é de base centrada (C), com pontos adicionais nos centros das bases das células. A terceira rede é de corpo centrado (I), com ponto adicional no centro da célula . A quarta rede é de face centrada (F), com ponto adicional no centro de cada face. No sistema cúbico, caracterizado por vetores unitários e ângulos iguais a 90^{o}, há três tipos de redes: cúbica simples (CS ou P), cúbica de corpo centrado (CCC ou I) e cúbica de face centrada (CFC ou F). No sistema trigonal um romboédrico (R) é escolhido como a célula primitiva (P).

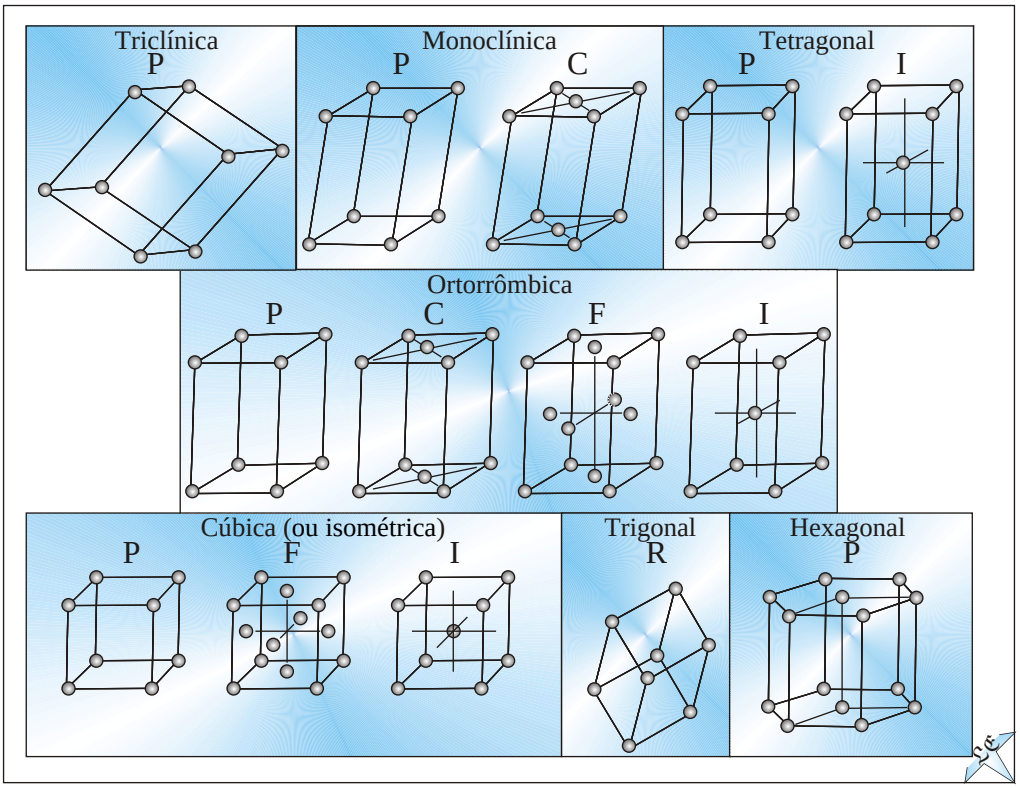


Figura 1.7 - As 14 redes de Bravais e os 7 sistemas de rede - A partir das restrições nos eixos e ângulos das células unitárias é possível construir os diversos sistemas de rede (ou cristalização) e as redes de Bravais. O volume dos sistemas de rede e das redes de Bravais são dados por: Triclínico (*abc* (1-cos^2 - cos^2 - cos^2 + 2 cos cos cos)½), Monoclínico (*abc* sen), Ortorrômbico (*abc*), Tetragonal (a^2c), Cúbico (a^3), Trigonal (a^3(1 - 3 cos^2 + 2 cos^3)½) e Hexagonal, 3 (3)½ a^2 c/2. Note que o sistema cúbico tem a mais simples expressão para seu volume.

A estrutura hexagonal pode formar um caso especial (ou variante), a estrutura hexagonal compacta (HCP), com três pontos no interior do hexágono, resultando uma célula unitária mais complexa, mas decomponível na estrutura hexagonal padrão.

Há diversas substâncias que cristalizam em cada um dos sistemas de rede, por exemplo: Triclínico cristalizam-se a $MnSiO_2$ (rondonita), $CuSO_4.5\ H_2O$, calcantita; Monoclínico (S_8 (enxofre), $CaSO_4.2H_2O$, gipsita); Ortorrômbico (S_8 (enxofre), $BeAl_2O_4$, crisoberilo); Tetragonal (MnO_2 (pirolusita), SnO_2 casserita); Cúbico (NaCl (sal de cozinha), C (diamante), FeS_2 (pirita), PbS, galena); Trigonal: Fe_2O_3 (hematita), SiO_2 (quartzo), $CaCO_3$ (calcita), HgS, cinábrio) e Hexagonal (C (grafite), $H_2O_{(s)}$, gelo).

Dentre os sistemas cristalinos tem sido extensivamente utilizado o sistema cúbico ou isométrico e seus casos especiais (ou variantes) em simulação de modelos simplificados de proteínas [P. E. Leopold et al.,**1992**; H. Taketomi et al., **1975**], devido sua simplicidade e diferentes propriedades,

dentre outros motivos, conforme mostrado abaixo.

1.3.3 Redes cúbicas

No sistema cristalino cúbico, há três redes de Bravais (P ou CS; F ou CFC; I ou CCC, Figuras 1.6 e 1.7), tendo seus pontos exatamente no grupo de simetria de um cubo, mas com diferentes propriedades (Tabela 1.1). Outras estruturas cristalinas cúbicas (como as redes de diamante e blenda), podem ser formadas adicionando uma base adequada em cada ponto dessas redes.

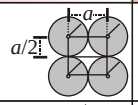
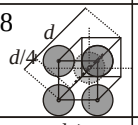
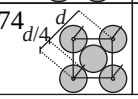
Tipos de Redes	Número de pontos da rede \ partículas por célula unitária	Distância mais próxima entre os pontos da rede	Densidade de empacotamento máximo	Exemplos
Cúbica Simples (CS)	1 \ 1	a	/6 = 0.52 (ou 52%) 	Fósforo
Cúbica de Corpo Centrado (CCC)	2 \ 2	$a(3)^{1/2}/2$	(3)½ /8 = 0.68 (ou 68%) 	Ba, Li, Na, Fe,K,Tl,W
Cúbica de Face Centrada (CFC)	4 \ 4	$a(2)^{1/2}/2$	/(3(2)½) = 0.74 (ou 74%) 	Al, Au, Cu, Pb, Pt, Ca, Ce, Ar

Tabela 1.1 - Algumas propriedades e exemplos das três redes cúbicas. A densidade de empacotamento máximo (4ª coluna) é dada pelo fator de empacotamento que mede a fração do espaço da célula unitária ocupada pelas partículas, supondo que estes são esferas sólidas, expresso por: (partículas por célula) x (volume da partículas) / (volume da célula), no caso da CS, (1) x ((4/3) $(a/2)^3$) / (a^3) = /6 = 52%. A célula unitária cúbica tem comprimento a, em todas as direções.

A densidade de empacotamento máximo para partículas idênticas, isto é, a maior fração de empacotamento possível para esferas sólidas e idênticas, varia de acordo com a rede cúbica.

A rede CS ocorre quando as partículas possuem um raio igual à metade do lado da célula unitária cúbica, correspondendo a densidade de 52%. Numa rede CCC o máximo empacotamento ocorre quando as partículas tem raio igual a um quarto da diagonal do corpo da célula unitária e densidade de 68%. Por fim, na rede CFC os átomos possuem raio igual a um quarto da diagonal de uma face da célula unitária, e densidade de empacotamento máxima e ao redor de 74%.

Na rede CFC, juntamente com a HCP (já definida anteriormente), ocorre a mais alta densidade de empacotamento possível [N. J. A. Sloane, **1994**], dentre quaisquer arranjos regulares ou irregulares, isto é, onde as partículas são empacotadas tão densamente quanto possível (Figura 1.8). Esta maior densidade média obtida num arranjo de rede regular de 74% (ou / [3(2)½]) foi provada pelo

matemático alemão C. F. Gauss (1777-1855). O interior de proteínas globulares pode ser tão densamente empacotado quanto o empacotamento máximo de um cristal [G. E. Schulz & R. H. Schirmer, **1990**]. Note porém que a proteína como um todo tem densidade muito inferior do que este empacotamento limite.

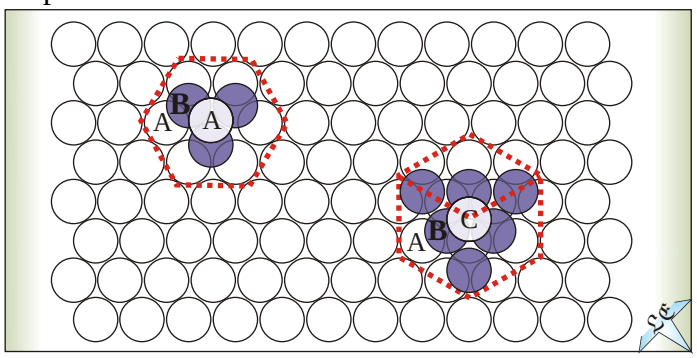


Figura 1.8 - Redes CFC e HCP, com máximas compactações - As letras (A, B, C) são referentes a camadas nos dois empilhamentos. No empilhamento HCP existem duas camadas "A" iguais (ABABABA), já no CFC as três camadas são diferentes (ABCABCA). O contorno das respectivas redes de Bravais são destacados em pontilhado vermelho. Note ainda que através de uma translação da terceira camada (A ou C), a rede HCP pode ser convertida em CFC, e vice-versa.

Das redes cúbicas é possível obter outras estruturas cristalinas, como casos especiais (ou variantes), por exemplo a rede de diamante. Este procedimento é similar ao feito para a estrutura hexagonal compacta (HCP), obtida da estrutura hexagonal padrão.

1.3.4 Rede de diamante

A rede de diamante consiste de um ponto da rede de Bravais CFC com duas partículas semelhantes, por ponto da rede, e pode ser considerada como duas redes CFC deslocadas, uma em relação à outra, e interprenetrando-se (Figura 1.9). A distância entre as duas partículas é igual a um quarto da diagonal do corpo do cubo.

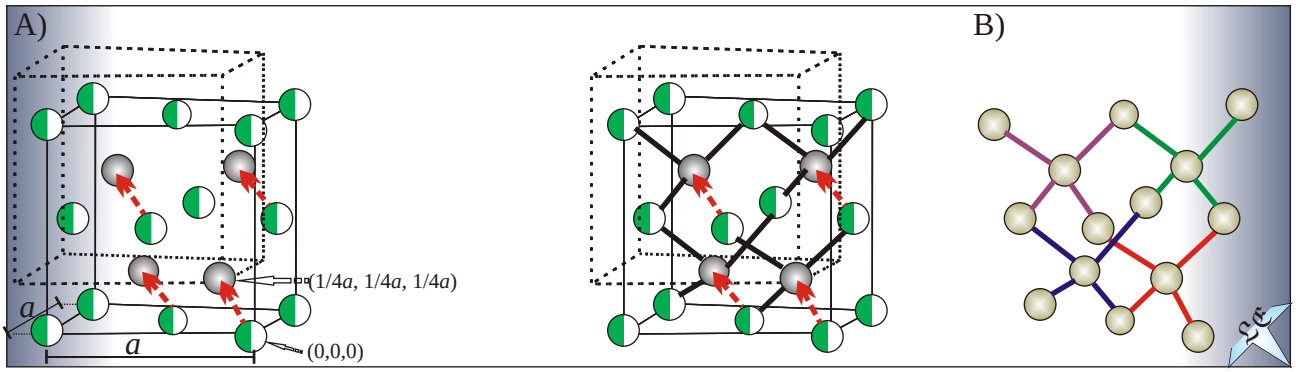


Figura 1.9 - Estrutura de uma rede de diamante: (A) a rede de diamante pode ser considerada como duas redes CFC interpenetrando-se, com base em (0,0,0) e (1/4*a*, 1/4*a*, 1/4*a*). Os círculos semipreenchidos indicam a rede CFC interprenetrada (em linhas sólidas) e os preenchidos indicam uma subrede CFC (linhas pontilhadas). (B) linhas sólidas coloridas conectando as partículas destacam a rede de diamante.

A rede de diamante contém quatro pontos da rede por célula unitária, e oito partículas por células unitárias, já a distância entre as partículas é um quarto da diagonal do corpo do cubo. A densidade de empacotamento máximo ocorre com partículas constituintes de raio igual a um oitavo da diagonal do corpo da célula unitária (*d*/8), correspondendo a 34% (Tabela 1.2).

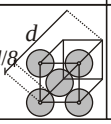
Tipo de Rede	Número de pontos da rede \ partículas por célula unitária	Distância mais próxima entre os pontos da rede	Densidade de empacotamento máximo	Exemplos
Diamante	4 \ 8	$a(3)^{1/2}/4$	$(3)^{1/2}/16=0,34$ 	diamante, germânio e silício

Tabela 1.2 - Mesmo que a Tabela 1.1, mas agora para a rede de diamante. São destacadas algumas características e exemplos para esta rede. Partículas de raio (d/8, com d = $(3)^{1/2}$ *a*) e 8 partículas por célula unitária, consequentemente a densidade de empacotamento é: $8[(4/3)\ (3)^{1/2}\ a/8)^3]/(a^3) = (3)^{1/2}/16 = 0,34$ ou 34%.

Além da rede de diamante, há várias outras estruturas cristalinas com base primitiva formada por dois sítios, como as redes: *honeycomb* (exemplo, grafite), hexagonal compacta, HCP (Be, Mg, Zn), sódio cloreto (NaCl, NaF, KCl).

Embora haja vários tipos de redes, como as descritas acima, a rede CS tem sido a mais amplamente utilizada em modelos simplificados de proteínas [P. E. Leopold et al.,**1992**], por diversos motivos. As redes CS são bastante versáteis, de fácil visualização e conveniente implementação computacional, além da vantagem do controle de variáveis. Como decorrência destas características, há duas informações relevantes e bastante utilizadas em simulação de modelo de proteínas globulares compactas, numa rede cúbica simples compacta com 27 partículas (Figura 1.10), a primeira que há 28 contatos totais, não ligados, e a segunda que existem 103.346 conformações maximamente compactas possíveis [L. Eller, **2003**].

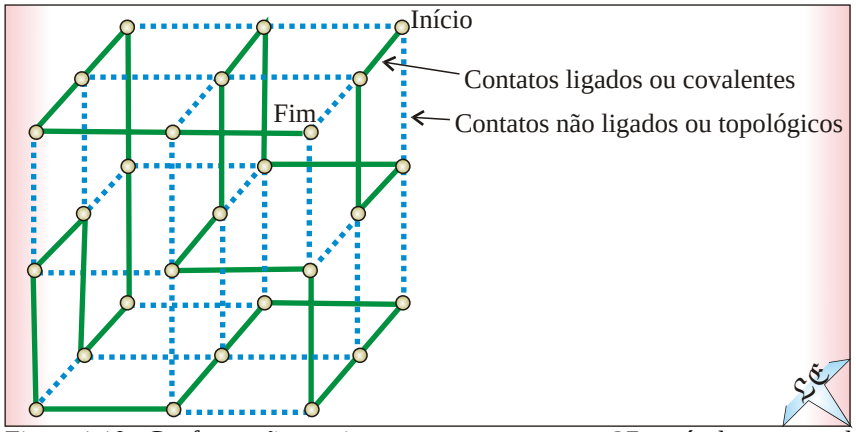


Figura 1.10 - Conformação maximamente compacta com 27 partículas numa rede CS. Nesta conformação (dentre 103.346 conformações) há 26 contatos ligados e 28 contatos não ligados, representados por retas continuas e pontilhadas, respectivamente. O número de contatos não ligados é dados pela expressão 2XYZ - XY - YZ - XZ + 1 =28, pois neste caso X=Y=Z=3 partículas.

Apesar da maior praticidade e uso da rede CS, a rede de diamante pode representar melhor as estruturas de proteínas e do solvente ao seu redor, pois as estruturas tetraédricas delineadas pelo carbono-alfa e pelas moléculas de água, ajustam bem os sítios da rede de diamante, cujos detalhes são fornecidos na seção abaixo e também no segundo capítulo. Consequentemente, a rede de diamante parece mais propícia para se analisar o processo de enovelamento e as estruturas de cadeias nascentes e daí a preferência por este tipo de rede no atual trabalho.

As interações físicas não covalentes (aquelas representados por retas pontilhadas na Figura 1.10) são de uma a três ordens de magnitude menores do que as covalentes (retas contínuas), porém ambas determinam eficaz e complementarmente as propriedades de proteínas [B. Nölting, **2006**]. Existem vários tipos de interações (também chamadas forças ou ligações) covalentes e não covalentes entre partículas (átomos, íons, moléculas, resíduos) atuando em diferentes situações, com intensidades variadas e finalidades específicas. A seguir, detalhamos estas interações, todas de natureza eletromagnética, com maior ênfase para o contexto específico de proteínas.

1.4 Forças em estruturas de biomoléculas (proteínas)

Nas estruturas de proteínas, há as interações (também chamadas forças ou ligações) covalentes e as não covalentes: eletrostáticas, Van der Waals e as hidrofóbicas. Estas duas interações possuem diferentes intensidades, tipos e estão presentes em diferentes níveis estruturais das biomoléculas, conduzindo desde seu enovelamento e estabilidade até seu reconhecimento molecular.

1.4.1 Interações covalentes

As interações que mantêm os átomos juntos como moléculas são denominadas covalentes e ocorrem quando átomos compartilham seus elétrons. No esqueleto peptídico linear as interações covalentes (como as ligações peptídicas) permitem a constituição do polipeptídeo. As energias típicas de uma ligação covalente são elevadas, de 300 a 400 kJ/mol, o que no conjunto das ligações contribui significativamente para a estabilidade, atráves do enrijecimento e proteção da cadeia.

Nos polipeptídeos e proteínas as interações covalentes estão presentes desde a conexão dos átomos num resíduo de aminoácido até a formação da estrutura primária com a ligação peptídica covalente, que ligam os resíduos no esqueleto polipeptídico (Figura 1.11a).

Um segundo tipo de interação covalente na estrutura primária ocorre entre cadeias laterais de pares de resíduos de cisteína (Figura 1.11b), produzindo pontes cistínas (ou dissulfetos), $-CH_2-S-S-CH_2-$ [T. Greighton, **1995**]. Tais interações são essenciais para aspectos funcionais de várias proteínas; além disso muitas proteínas pequenas não podem formar estruturas estáveis sem tais interações.

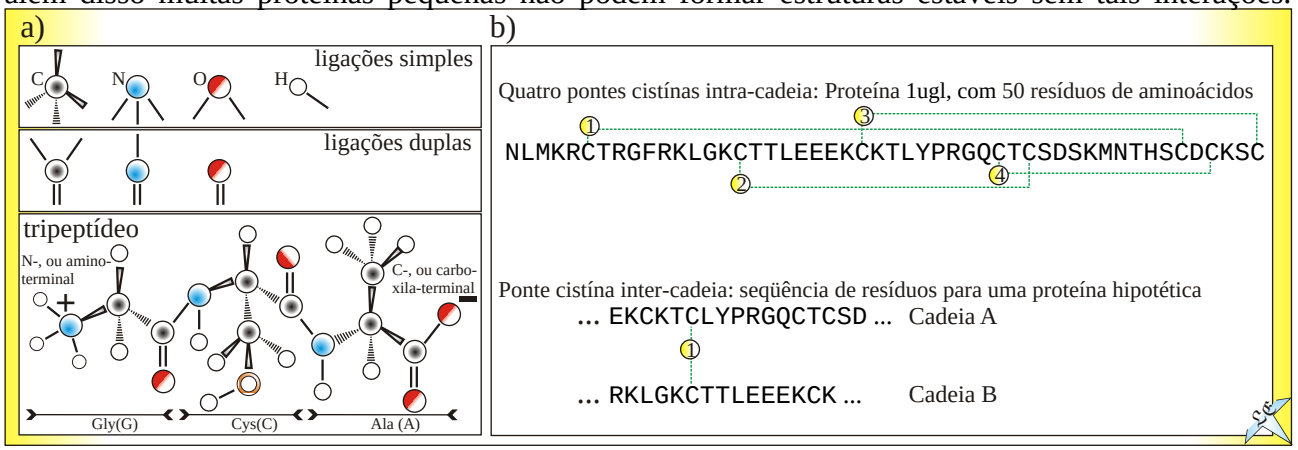


Figura 1.11 - Forças covalentes na estrutura primária de polipeptídeos, representadas por ligações covalentes e pontes cistínas. a) Todas as ligações nos átomos e no tripeptídeo, Gly CysAla, são covalentes; b) Exemplo de pontes cistínas intra- e inter-cadeia, com cada resíduo representado pelo código de uma letra. Cada átomo forma um número fixo de ligações covalentes, pelo compartilhamento de seus elétrons, em específicos arranjos espaciais (diferentes nas ligações simples e duplas), por exemplo, os átomos de carbono, nitrogênio e oxigênio, formam quatro, três e duas ligações, respectivamente. As ligações são representadas por setas pontilhadas e preenchidas, entrando e saindo no plano do papel, respectivamente.

Apesar da indiscutível importância das pontes cistínas, nem todas as proteínas possuem resíduos de cisteínas formando pontes. Além disso, quando tais pontes existem é difícil predizer suas exatas localizações e quantidades somente a partir da sequência de aminoácidos, o que se torna um grande desafio atual enfrentado por vários métodos de predição [Z. Dosztányl et al., **2003**].

1.4.2 Interações não covalentes

As interações não covalentes são interações entre átomos, íons e moléculas, de uma a três ordens de magnitude menores do que as covalentes [G. E. Schulz & R. H. Schirmer, **1990**]. As interações não covalentes (constituída por contatos não ligados ou topológicos) são suficientemente fortes para manter agrupados íons, moléculas e grupos individuais de moléculas, mas por outro lado

também são suficientemente fracas para serem continuamente quebradas e novamente formadas contribuindo para a natureza dinâmica característica dos sistemas biológicos, como proteínas.

1.4.2.1 Eletrostática

A intensidade da força **F** entre um par de cargas (q e q') separadas por uma distância *r* num meio uniforme de constante dielétrica , é dada pela lei de Coulomb (equação 1.1):

$$|\mathbf{F}| = k(qq')/(r^2), \quad (1.1)$$

onde k é uma constante, e =1 (no vácuo); 1,89 (em 20 °C) para o hexano; 2,21 (25 °C) para o dioxano; 32,6 (25 °C) para o metanol; e 87,8 (25 °C) e 74,2 (37 °C) para a água. A constante dielétrica do meio ambiente é alta para um solvente polar e baixa para solventes orgânicos apolares.

A lei de Coulomb explica como o meio celular efetivamente protege as cargas umas das outras, pois quanto maior o valor de , menor a força entre elas, ou seja, a intensidade das interações eletrostáticas é dependente do meio. Em sistemas biológicos, as cargas estão separadas por água, outras moléculas ou partes de moléculas. Nestes casos a escolha do valor de é difícil e complexa, por exemplo, o interior de uma proteína pode ser similar ao dioxano, já o exterior da proteína é aquoso e a presença de sal também pode alterar o valor de [N. A. Baker et al., **2001**] Além disso, quando o ambiente consiste de regiões com diferentes, considerações especiais devem ser feitas, com o uso da equação de Poisson-Boltzmann [W. Rocchia et al., **2001**].

A energia de interação *U* (equação 1.2) entre duas cargas (ou dipolos) é o trabalho requerido para trazer tais cargas de separação infinita para a distância desejada *r*, sendo calculada integrando a força sobre a distância, e obtendo-se

$$U = k(qq')/(r), \quad (1.2)$$

A energia de interação *U* é uma boa estimativa da entalpia de interação das cargas e além disso pode ser estimada para outros casos. De interesse em sistemas biológicos são as interações entre cargas fixas, dipolos e dipolos induzidos, que varia com a distância de separação (Tabela 1.3). Ao contrário de

outras interações, em proteínas, as eletrostáticas são efetivas em relativamente longas distâncias, sobretudo em meio com baixa constante dielétrica.

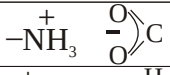
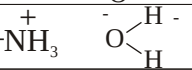
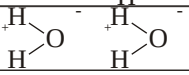
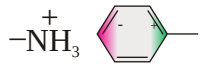
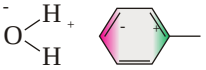
Tipo de interação	Dependência da Energia com a distância	Exemplo
Carga-Carga Forças de mais longo alcance, não direcional	$1/r$	
Carga-Dipolo Depende da orientação do dipolo	$1/r^2$	
Dipolo-Dipolo Depende da orientação mútua dos dipolos	$1/r^3$	
Carga-Dipolo induzido Depende da polarizabilidade das moléculas em que o dipolo é induzido	$1/r^4$	
Dipolo-Dipolo induzido Depende da polarizabilidade das moléculas em que o dipolo é induzido	$1/r^5$	
Dispersão Envolve mútua sincronização de flutuação de carga	$1/r^6$	

Tabela 1.3 - Tipos de interações não covalentes, com suas respectivas energias e exemplos. Aqui os dipolos são assumidos orientados para interação máxima e depois fixados. Nas interações acima foi mostrada apenas a parte atrativa (ou força de dispersão, na última linha da tabela) da interação de Van der Waals e omitida propositalmente a parte repulsiva, que devido sua importância para o presente trabalho será comentada a parte ainda neste capítulo.

Note que no caso da interação entre dipolos a energia de interação U deve ser tratada com um pouco mais de sutileza, devido os aspectos direcionais dos dipolos. No caso de dipolos com (vetor) momento de dipolo $_1$ e $_2$, a energia de interação *U* (equação 1.3) pode assumir os dipolos como pontuais. Portanto, a distância *r* deve ser maior do que o comprimento dos dipolos.

$$U = 332\{[(\ _1 .\ _2)/r^3_{1,2}] - [3(\ _1.\mathbf{r}_{1,2}).(\ _2.\mathbf{r}_{1,2})/r^5_{1,2}]\}/ \qquad (1.3)$$

A energia de interação *U*, entre dipolos, depende da orientação entre eles. Há dois casos que ilustram bem a variação da energia *U* (Figura 1.12). No caso de dipolos orientados aleatoriamente a energia *U* é zero. Já no caso de dipolos com mesma/ou oposta direção e sentido a energia depende destas orientações.

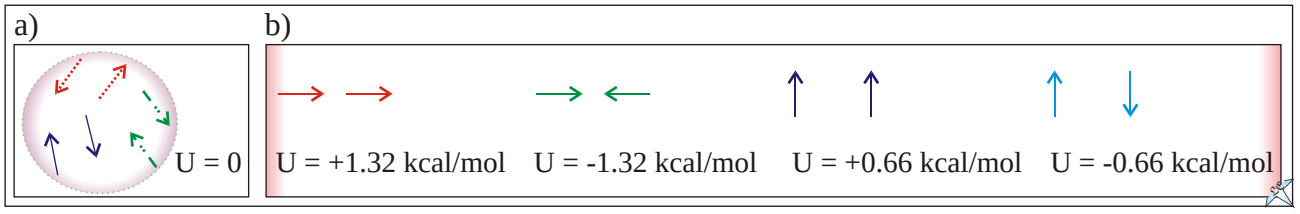


Figura 1.12 - Casos particulares da interação entre dipolos. (a) Para sistema com dipolos orientados aleatoriamente, a energia de interação total *U* é igual a zero; (b) Para os valores acima foram assumidos dois dipolos de 1 D (debye) separados por 5 Å num meio com = 4.
Lembrando que o produto escalar entre dois vetores $\mathbf{a}.\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos$, sendo o ângulo entre **a** e **b**. Para o interior da proteína, normalmente é estimado a constante dielétrica de 2 a 4, por isso aqui foi adotado = 4.

Dentre as forças eletrostáticas de longo alcance, presentes em proteínas, a seguir destacamos as ligações iônicas (ou pontes salinas), ligação de hidrogênio e as interações de Van der Waals [G. Faure

et al., **2008**].

1.4.2.1.1 Ligações Iônicas

Nas ligações iônicas há atrações eletrostáticas fortes entre grupos carregados opostamente (Figura 1.13a), por exemplo, entre os grupos amino ($-NH3^{+}$) nas cadeias laterais de aminoácidos básicos carregados positivamente (Arg, His, Lys) e os grupos carboxila ($-COO^{-}$) carregados negativamente nos aminoácidos ácidos, Asp, Glu [J. B. O. Mitchell et al., **1992**]. As forças iônicas são responsáveis pela dureza de minerais, como o mármore, ágata e o cristal de cloreto de sódio (Seção 1.3). Em alguns sistemas biológicos estas ligações podem ser muito importantes, por exemplo, um aminoácido ácido no lugar apropriado de uma enzima pode se ligar a outro aminoácido carregado positivamente num substrato, formando o complexo enzima-substrato.

Em geral, as ligações iônicas no interior proteico são raras já que a maioria dos aminoácidos carregados encontram-se, em geral, na superfície da proteína; além disso muitas proteínas mantêm sua conformação nativa sem a presença destas ligações.

1.4.2.1.2 Ligações de Hidrogênios

A ligação de hidrogênio é uma interação atrativa fraca entre um átomo de hidrogênio, que tem uma carga parcial positiva numa molécula (doadora de ligação de hidrogênio), com um átomo eletronegativo próximo, como o nitrogênio, oxigênio, com uma carga parcial negativa, numa outra molécula, aceptora de ligação de hidrogênio [M. S. Weiss et al., **2001**]. Em geral, essas cargas são partes de dipolos (Figura 1.13b). As ligações de hidrogênio são importantes e estão presentes em inúmeros sistemas biológicos, por exemplo, elas mantêm unidas duas bases (e.g., C com G) no DNA e RNA (Seção 1.2).

Grandes quantidades de ligações de hidrogênios são encontradas não somente no estado nativo, mas também no estado desnaturado da proteína. No estado nativo (e às vezes no estado intermediário, globulo fundido), em geral, estas interações se mostram efetivas na estabilização de estruturas de

proteínas, como na manutenção de estruturas secundárias regulares, e.g., helices e folhas .

As estruturas tridimensionais das proteínas são frequentemente dependente de uma intricada rede de ligações de hidrogênio, contudo algumas moléculas, como oligomeros, mostram que é possível biopolímeros adquirem uma conformação estável mesmo sem formação de pontes de hidrogênio intra-cadeia [J. C. Nelson et al., **1997**].

1.4.2.1.3 Forças de Van der Waals

A força de Van der Waals (VDW) corresponde a uma interação atrativa (força de dispersão) de muito curta distância entre quaisquer dois átomos, e com intensidade fraca, devido suas flutuações de carga elétrica (dipolos elétricos transitórios) na nuvem eletrônica que cerca cada átomo. Assim, a distribuição de elétrons, ao redor de um átomo, varia a cada instante, ocasionando uma separação momentânea de carga maior de um lado do átomo. De fato, cada átomo comporta-se como um dipolo oscilante, gerado pelas distribuição assimétrica dos elétrons movendo em relação ao núcleo, o que num par de átomos, polariza o átomo adjacente.

A força de VDW [H. C. Hamaker, **1937**], além do termo referente à força de dispersão, também possui um termo referente a interação eletrostática e quando dois átomos estão muito próximos eles repelem-se mutuamente através de uma força repulsiva por meio de suas camadas de elétrons (comentada separadamente abaixo). A atração entre os dois átomos ocorrem até uma distância de contato, igual a soma de seus raios atômicos, denominados raios de VDW, específico para cada átomo (Figura 1.13c).

Em geral uma série de simplificações são feitas no cálculo da força de VDW, por exemplo, considerar a interação somente entre átomos muito próximos. Já a interação eletrostática, ente as cargas efetivas q e q', é simplificada quanto à geometria das interações. Como consequência a força de VDW (*E*, equação 1.4) pode ser considerada isotrópica (igual em todas as direções) e pontual, ou seja, não depende da orientação, mas somente da distância (*r*) entre as entidades interagentes,daí:

$$E = \text{termo repulsivo} + \text{termo de dispersão} + \text{interação eletrostática} = Ar^{-12} - Br^{-6} + (qq')r^{-1} \qquad (1.4)$$

Embora a força de dispersão seja uma interação atrativa fraca, quando em grande quantidade, ela pode ser suficientemente forte para proporcionar uma ligação firme, por exemplo, entre duas superfícies complementares muito próximas (Figura 1.13c), como no caso dos reconhecimentos proteína-proteína e anticorpo-antígeno.

Em alguns casos, a interação eletrostática pode ser incorporada nos parâmetros A e B (equação 1.4). Uma descrição completa para a força de dispersão e de repulsão pode ser dada pelo potencial de Lennard-Jonnes.

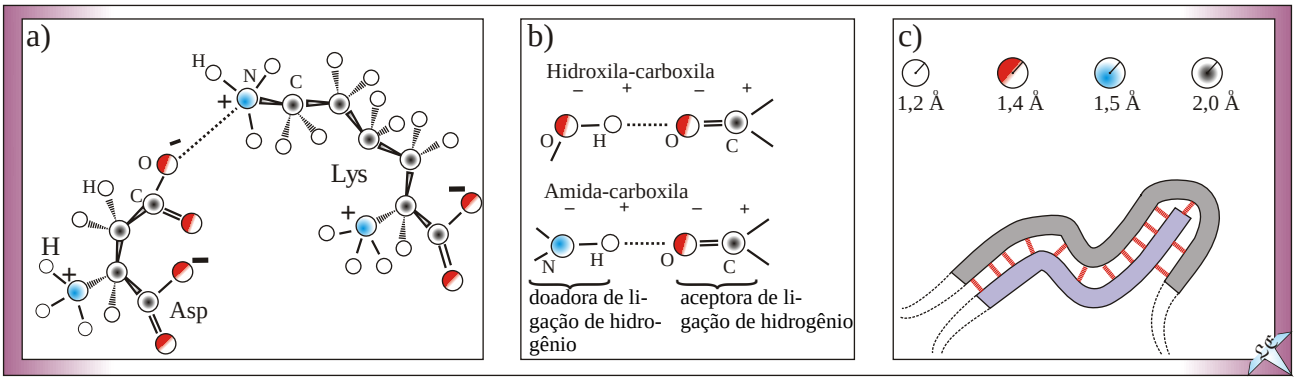


Figura 1.13 - Exemplos de interações não covalentes: ligações iônicas, hidrogênio e forças de Van der Waals. (a) Exemplo de interações atrativas iônicas entre aminoácidos carregados com cargas opostas; (b) Dois casos de ligação de hidrogênio, aqui é destacado o aspecto direcional destas ligações, que neste caso são mais fortes devido os dois átomos (ambos eletronegativos, como O e N) estarem numa linha reta; (c) Acima são fornecidos os raios de VDW para os átomos H, O, N, C, e abaixo é mostrado duas superfícies moleculares complementares ligadas por muitas forças fracas (no caso pela de força de dispersão) e assim resistindo a energia térmica que tende a afastá-las.

1.4.2.2 Interações Hidrofóbicas

A interação hidrofóbica pode ser qualitativamente considerada a força responsável por grupos hidrofóbicos (ou apolares) se agregarem ou agruparem do meio polar a sua volta, geralmente a água (ou solvente aquoso). Ao se agregarem os grupos hidrofóbicos minimizam a exposição de superfície ao meio, resultando no efeito hidrofóbico.

Da maneira como descrita aqui, a interação hidrofóbica não é classificada como uma ligação (ou força) verdadeira, já que ela não resulta de uma atração intrínsica entre grupos hidrofóbicos. Mais propriamente, tal interação resulta do sistema como um todo obter maior estabilidade termodinâmica, através da minimização do número de moléculas de água ao redor do grupo hidrofóbico (Figura 1.14).

Outra abordagem das interações hidrofóbicas sugerem que elas são forças fracas [G. I.

Makhatadze, **1995**], o que fica evidente na atração de superfícies e grupos hidrofóbicos [E. E. Meyer et al., **2006**]. Além disso, os grupos hidrofóbicos podem formar ligação fraca entre si baseado em atração eletrostática, como a de VDW.

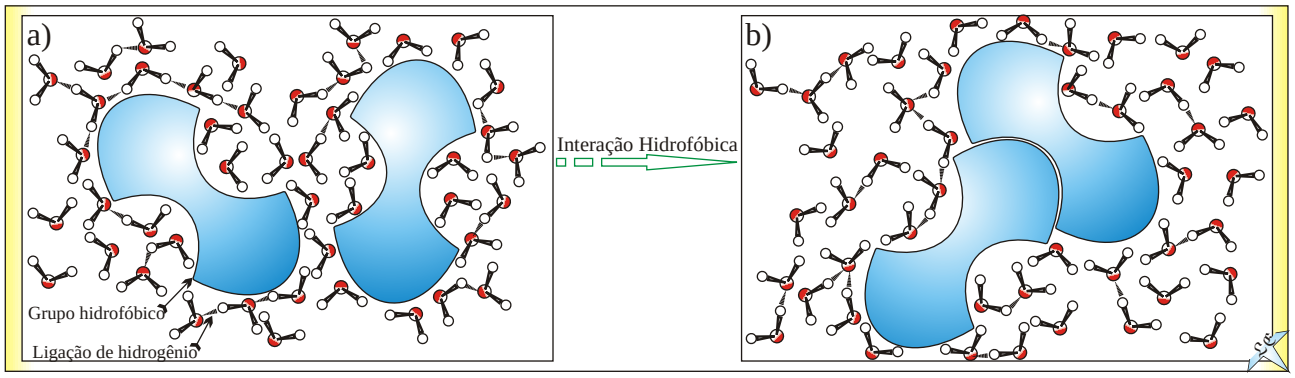


Figura 1.14 - Uma das abordagens da interação hidrofóbica - os grupos hidrofóbicos inicialmente separados (a) são forçados a se agregarem (b), e assim minimizando suas exposições as moléculas de água ao seu redor.

Numa biomolécula (como uma proteína) além dos grupos hidrofóbicos, há os hidrofílicos (ou polares) que ficam mais expostos ao meio polar, inclusive interagindo com ele e simultaneamente protegendo os hidrofóbicos internamente da exposição ao solvente.

Uma decorrência direta das diferentes abordagens das interações hidrofóbicas e do efeito hidrofóbico pode ser observado nas diversificadas implementações computacionais de tal interação e efeito. Às vezes o meio polar é utilizado explicitamente na interação com o biomolécula de interesse [L. F. O. Rocha, et al. **2009**], outras vezes o meio é tratado implicitamente e a interação hidrofóbica é representada por funções energias dependente dos tipos de resíduos (hidrofóbicos ou hidrofílicos) interagentes [L. Eller, **2003**]. Estas diferentes implementações da interação hidrofóbica nem sempre levam a resultados equivalentes ou complementares.

Apesar das divergências quanto a implementação computacional, o efeito hidrofóbico tem sido considerado a maior fonte de variação de energia livre que guia o enovelamento de proteína e também o maior fator estabilizante de estruturas biopoliméricas. E também parece desempenhar outros papéis chaves, como na formação de membrana, ligação proteína-proteína e ligante-proteína.

1.4.2.3 Forças repulsivas de Van der Waals

A força repulsiva de VDW é a interação de repulsão entre dois átomos muito próximos, com

separação inferior a uma distância mínima (ou distância de contato) entre eles, tal que suas nuvens eletrônicas ao redor do núcleo atômico repelem-se fortemente (Tabela 1.4). A distância de contato refere-se ao menor valor da distância interatômica cujas forças de VDW atrativas de dispersão e eletrostática são contrabalanceadas pela força de repulsão.

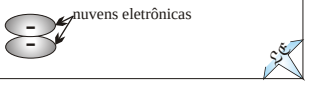
Tipo de interação	Dependência da Energia com a distância	Exemplo
Repulsão de Van de Waals Ocorre quando as nuvens atômicos tentam se sobrepor.	$1/r^{12}$	

Tabela 1.4 - Força repulsiva de VDW - esta força é onipresente e origina as interações estéricas, ingrediente essencial na determinação estrutural de moléculas biológicas.

Uma enorme simplificação para a distância de contato é considerá-la como um contato de esferas rígidas, cujos tamanhos são fornecidos pela soma dos raios de VDW dos átomos interagentes (Figura 1.13c). Um tipo elementar de repulsão de esferas rígidas é conhecido como interações de volume excluído [D. Ben-Avraham & S. Havlin, **2000**], ou seja, uma partícula (ou esfera rígida) não pode ocupar espaço que já está ocupado por outra partícula. No caso de simulações em rede dizemos que partículas não podem ocupar o mesmo sítio simultaneamente, o que conduz ao caminho de auto impedimento (do inglês *self-avoiding walk* , SAW). Um SAW é um caminho aleatório que não cruza sobre si mesmo [C. M. Guttman, **1986**].

Frequentemente grupos que interagem via ligação de hidrogênio violam a distância de contato de VDW. Isso ocorre quando os átomos doadores e aceptores na ligação de hidrogênio tem grandes cargas parciais (positivas e negativas), então há uma deslocamento da nuvens de elétrons ao redor do átomo de hidrogênio e assim reduzindo significativamente a camada de repulsão. Como resultado, o átomo mais eletronegativo pode se aproximar muito mais do hidrogênio, antes de ocorrer a repulsão. Tais distâncias próximas resultam em maiores atrações colombianas e de dispersão.

Conforme exposto acima, na contextualização do atual trabalho, todas as interações não covalentes, inclusive a força repulsiva de VDW, possuem limitações e aplicabilidades ainda em estudo. Contudo a força repulsiva de VDW reconhecidamente desempenha um papel principal na limitação da imensa quantidade de conformações possíveis de um biopolímero, como uma proteína, além de ser uma interação utilizada na maioria dos modelos que analisam proteínas atualmente, daí sua utilização preferencial neste trabalho.

Apesar da argumentação, no último parágrafo, a favor de uma análise mais detalhada utilizando apenas a força repulsiva de VDW (esfera rígida), a seguir apresentamos argumentos adicionais a fim de justificar a escolha somente desta força para o presente trabalho.

1.4.3 Argumentos adicionais a favor da utilização apenas da força repulsiva de VDW

No estudo de proteínas, apesar de avanços em casos particulares, ainda é uma questão aberta como ocorre seu processo de *folding*, ou seja, como é o relacionamento entre sua sequência de resíduos de aminoácidos e sua conformação nativa funcional. Tal relacionamento é basicamente uma questão com três aspectos sob estudos intensos [V. S. Pande, et al., **2000**], a saber, como a proteína atinge seu estado nativo?; como são selecionadas suas sequências de resíduos?; e como predizer seu estado enovelado dado sua sequência de resíduos?.

O problema do folding de proteínas é um assunto multidisciplinar combinando conhecimentos de áreas como a biologia molecular, física de polímeros e física de sistemas desordenados (e.g. Figura 1.1). Atualmente com o aumento significativo de estruturas de proteínas determinadas experimentalmente e suas disponibilidades no Banco de Dados de Proteínas, PDB [H. M. Berman et al., **2000**], juntamente com simulação computacional e ferramentas teóricas especialmente planejadas para uma análise cuidadosa e delicada destas estruturas, podemos obter novas abordagens de temas aparentemente consolidados, como a não aleatoriedade das sequências e das conformações nativas das proteínas.

Um exemplo da combinação de experimentos com simulação computacional (método Monte Carlo, em ou fora de rede cristalina) e ferramentas teóricas é dada pela função discrepância configuracional, D (equação 1.5), ver L. F. O. Rocha, **2005**. Para cada par de configuração no modelo (A) e na proteína (B), ambas com N resíduos, são calculadas as distâncias intracadeias normalizadas $d_{i,j}$ entre os resíduos i e j, que em seguida são subtraídas. Excluíndos 2 pares primeiros vizinhos ao longo da cadeia (de 35 resíduos), totalizamos 528 distâncias intracadeia, cujo valor normaliza a função D. Quanto menor D, maior é a semelhança entre as configurações A e B, tal que $D = 0$ corresponde a

duas configurações idênticas.

$$D_{A,B} = \frac{1}{528} \sum_{i=1}^{N} \sum_{j=3}^{N} |{}^{A}d_{i,j} - {}^{B}d_{i,j}| \qquad (1.5)$$

Utilizando um modelo fora da rede para uma cadeia desnaturada com 35 resíduos imersa num solvente com 12.565 unidades (cada uma de mesmo tamanho dos resíduos) e somente com interação de esfera rígida, são geradas configurações do modelo em 100.000 Passos Monte Carlo e cada uma delas comparada com a estrutura nativa de cada uma de 12 proteínas globulares compactas, também com 35 resíduos [L. F. O. Rocha, **2005**]. Para tais comparações é empregada a função D e calculada sua frequência $f(D < 0{,}17)$, para cada proteína (Figura 1.15), no transcorrer do tempo de simulação, em quatro simulações independentes.

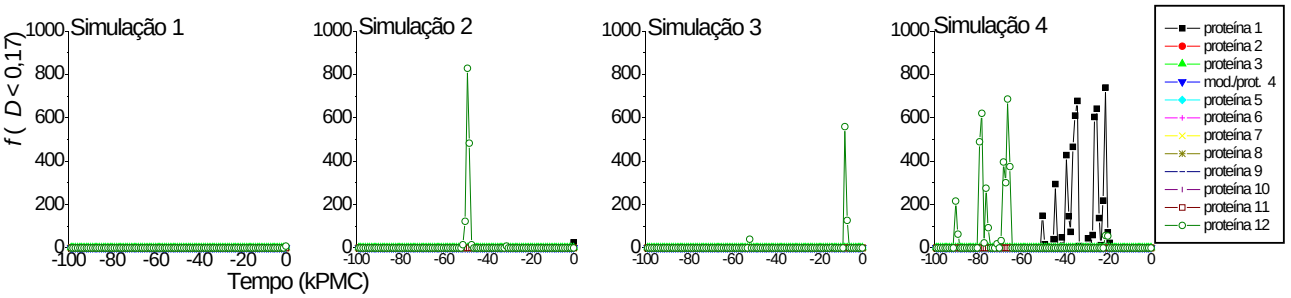


Figura 1.15 - Freqüência da discrepância configuracional D, ao longo de quatro simulações independentes, comparando configurações do modelo, somente com interação de esfera rígida, e conformações nativas de 12 proteínas globulares, a nível de resíduos. O valor pequeno de 0,17 serve para compararmos configurações modelo-proteínas, que neste caso são muito semelhantes. A amplitude de $f(D < 0{,}17)$ mede o número total de vezes com que a condição $D < 0{,}17$ é satisfeita, e está relacionada com a freqüência com que a cadeia no m odelo se aproxima da estrutura nativa de cada uma das proteínas.
Obs. - Tempo negativo se refere ao período de termalização (em $k_B T \to \infty$), após o qual a simulação passa a utilizar outras interações juntamente com a de esfera rígida, no processo de *refolding.*
- As proteínas são organizadas sequencialmente da mais para a menos compacta, 1c2u (nº 1), 1roo, 1txm, 1tsk, 1bei, 2pta, 1c49, 1quz, 1c4e, 1mb6, 1lqr, 1d1h (nº 12).

Observando a evolução da frequência de discrepância configuracional, notamos que a interação de esfera rígida não proporciona estabilidade configuracional para o glóbulo, já que as grandes semelhanças (dada por $f(D < 0{,}17)$) não são persistentes no transcorrer do tempo de simulação, mas tal interação é capaz de induzir a cadeia através de conformações próximas ao estado nativo das proteínas nºs 1 e 12.

1.5 Comentários Finais

Os resultados na Figura 1.15, embora obtidos de outro modelo/método (com outra abordagem e Finalidade), mostram que a restrição estérica juntamente com a conectividade da cadeia sozinhas não estabilizam a cadeia próximo da conformação nativa de uma dada proteína, cujo papel cabe a outras interações não covalentes, mas revela que esta interação pode ter um papel mais importante do que tradicionalmente tem-se atribuído a ela. Portanto parece adequado atualmente revisitar e reavaliar alguns conceitos já estabelecidos, como a tradicional diferenciação polímero-proteína. Aliás tais questionamentos tem sido feitos em outros tópicos (ou temas), e em diversas áreas da ciência, como em cosmologia [M. Gleiser, **2010**]**.**

Nos próximos dois Capítulos geramos configurações utilizando o modelo de crescimento de cadeia nascente numa rede tetraédrica (ou de diamante) infinita e observamos a influência das restrições estéricas (interação de esfera rígida) e das armadilhas topológicas no processo de crescimento e enovelamento das cadeias geradas.

No Capítulo 2 avaliamos os efeitos das restrições estéricas em diferentes intensidades, com relação a algumas quantidades estruturais ou conformacionais padrões (tais como o raio de giração e a distância ponta a ponta), número de sobreposições e armadilhas topológicas, nos métodos de cadeia ideal (com sobreposição permissível entre resíduos), cadeia com restrição estérica de curto alcance (impedimento de sobreposição apenas entre resíduos primeiros vizinhos, ao longo da cadeia, *i* com *i*-2) e cadeia com restrição estérica de longo alcance (ou total), ou seja, sem quaisquer sobreposições entre resíduos ao longo da cadeia. Nossa análise dos resultados mostram, entre outros aspectos, que a restrição estérica de curto alcance estende-se além dos primeiros vizinhos ao longo da cadeia, já que é equivalente a restrição estérica de longo alcance para cadeias curtas ($N \lesssim 100$ resíduos) e pelo menos duas vezes maior, mesmo para cadeias com N = 1000. Tais resultados mostram os papéis correspondentes das interações de curto e longo alcance no enovelamento de proteínas desnaturadas, um assunto atual e ainda bastante discutido.

No Capítulo 3 utilizamos dois métodos independentes que crescem diferentemente cadeias armadilhadas (ou sem alternativas de crescimento), o que implica em quantidade e frequência de

armadilhas muito diferentes para finalizar o processo de crescimento, principalmente para cadeias longas ($N \geqslant 1000$). Nossos resultados mostram que apesar da maneira muito diferente de sair das armadilhas os dois métodos levam a cadeias com compactação média (medidas pelo raio de giração e expoente de escalonamento) muito semelhantes, porém em tempos de enovelamento muito diferentes. Tais resultados corroboram para um melhor entendimento do processo de construção de cadeias longas, inclusive servindo de base para construção de outros modelos mais sofisticados, como aqueles que usam um sistema de pesos a fim de que todas as possíveis cadeias permitidas (geradas) sejam contadas ou estimadas igualmente [M. N. Rosenbluth & A. W. Rosenbluth, **1955**], o que não é o caso dos dois atuais métodos empregados aqui.

Capítulo 2: Implicações das restrições estéricas para o crescimento e enovelamento de cadeias peptídicas nascentes

2.1 Introdução

Um importante elemento em estruturas de biomoléculas é dado pelos vínculos geométricos e pelas dimensões de ligações e ângulos [G. E. Schulz & R. H. Schirmer, **1990**], dentre os quais destacamos alguns exemplos na Figura 2.1.

Nas moléculas de água e no carbono-alfa (de resíduos de aminoácidos) os vínculos da geometria tetraédrica é um ingrediente bastante comum (Figura 2.1a). Por outro lado todas as ligações peptídicas (exceto para o resíduo Prolina) estão na configuração trans (Figura 2.1b), pois é geometricamente preferível por não ser tão restritiva quanto a configuração cis (não mostrada na Figura 2.1). Com uma ligação peptídica rígida (inibida por ressonância nas ligações C - O e C - N) e comprimentos e ângulos de ligação um tanto quanto rígidos, a conformação da cadeia polipeptídica é essencialmente descrita pelos ângulos diedrais, e , no carbono-alfa (Figura 2.1c), o que contribui significativamente para a estabilidade do polipeptídeo ou proteína.

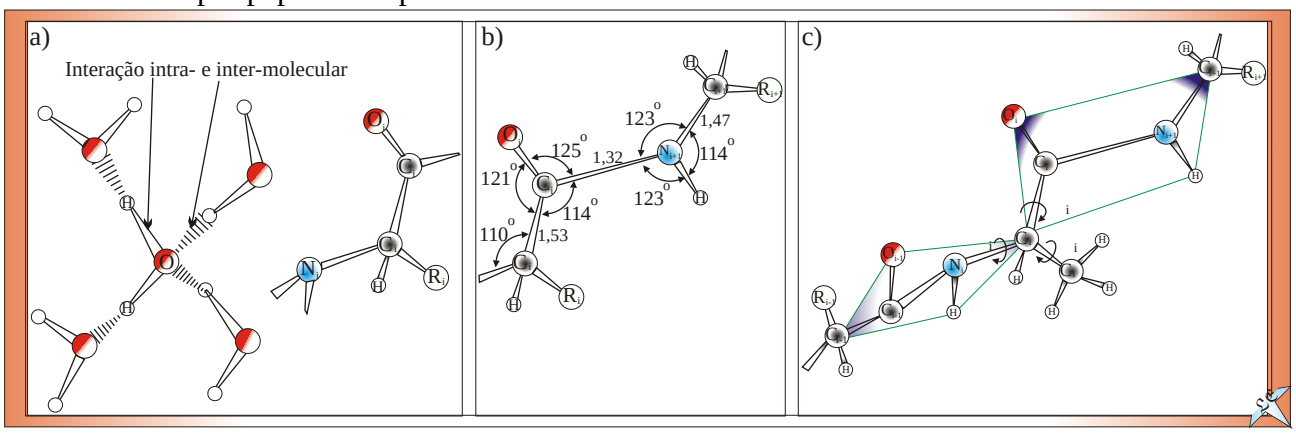


Figura 2.1 - Exemplos de vínculos geométricos e dimensões em estruturas de biomoléculas. a) Geometria tetraédrica em moléculas de gelo (na água líquida a temperatura ambiente, cerca de 20% das moléculas ainda preservam a estrutura tetraédrica) e no carbono-alfa da cadeia principal; b) distâncias inter-atômicas e ângulos de ligações para uma ligação peptídica trans usual [L. Pauling et al., **1951**]; c) Ângulos diedrais (,) na cadeia principal e na cadeia lateral (aqui para o resíduo *i*, estamos utilizando como exemplo a Alanina).

Em estrutura de proteína a geometria tetraédrica assume um papel importante por ser a geometria adotada pelo carbono-alfa (Figura 2.1). Além do que, em geral, a conformação da cadeia

principal de uma proteína costuma ser representada apenas pela sequência de posições dos carbonos-alfas.

2.2 Modelo e métodos

Aqui analisaremos os resultados decorrentes das restrições estéricas em associação com imposições geométricas sobre as conformações de cadeias nascentes de polipeptídeos. Para emular o crescimento e enovelamento de cadeias nascentes reais (seção 1.2, Capítulo 1) usaremos um modelo alternativo [A. Caliri & M. A. A. da Silva, **1997**], que descreve o comportamento de polímeros lineares puros em rede. Aqui simulamos cadeias de comprimento (ou número de resíduos) máximo de 1000 resíduos.

Os resíduos na estrutura cristalina tetraédrica são bem ajustados pelos sítios numa rede de diamante, que por sua vez é originada de uma rede de Bravais CFC (cúbica de face centrada), com dois pontos semelhantes em sua base primitiva. Replicando-se a rede CFC, com um átomo por base, ficam duas redes cúbicas deslocadas uma em relação à outra, sendo que o vértice da célula unitária de uma delas é sempre o centro da outra (seção 1.3, Capítulo 1).

2.2.1 Descrição do modelo

O modelo consiste em crescer cadeias com N resíduos numa rede de diamante (ou tetraédrica) utilizando um conjunto de oito vetores (equação 2.1), separados em dois grupos (Figura 2.2) a qual chamamos de $\mathbf{P}_k$ e $\mathbf{N}_k$, onde $1 < k < 4$:

$$\mathbf{P}_1 = (+1,+1,+1)\ ;\ \mathbf{P}_2 = (+1,-1,-1)\ ;\ \mathbf{P}_3 = (-1,+1,-1)\ ;\ \mathbf{P}_4 = (-1,-1,+1)$$

e (2.1)

$$\mathbf{N}_1 = -\mathbf{P}_1\ ;\ \mathbf{N}_2 = -\mathbf{P}_2\ ;\ \mathbf{N}_3 = -\mathbf{P}_3\ ;\ \mathbf{N}_4 = -\mathbf{P}_4$$

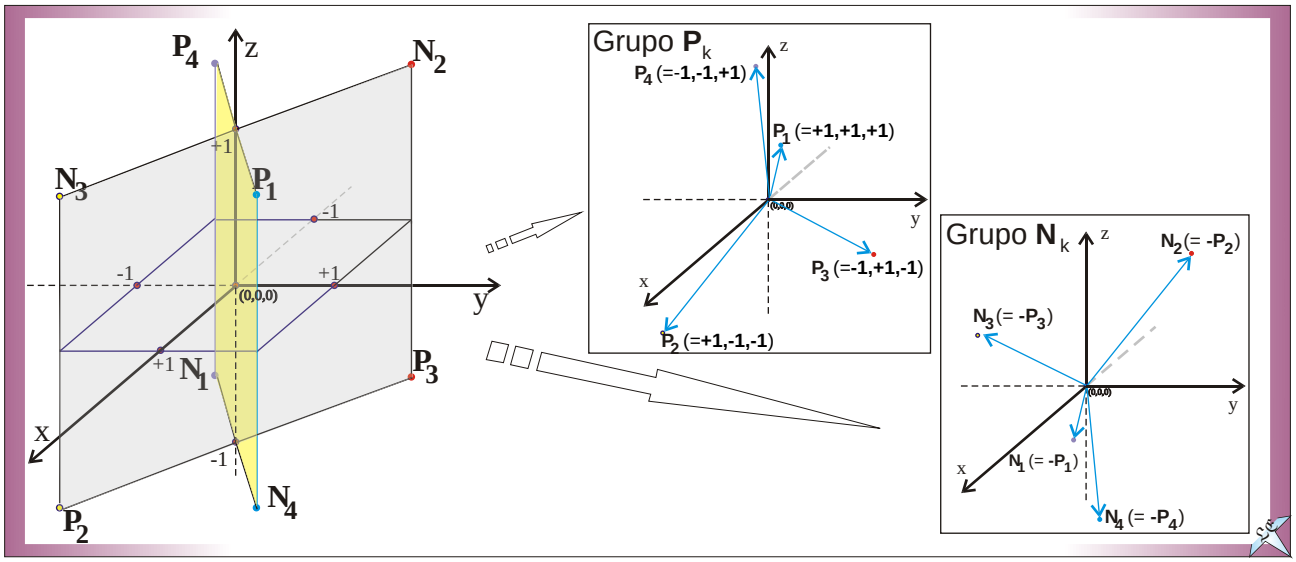


Figura 2.2- Disposição dos oito vetores com mesmo comprimento, nos grupos $\mathbf{P}_k$ e $\mathbf{N}_k$. Estes vetores formam a base do processo de crescimento de cadeia na rede tetraédrica e note que cada correspondente par *k* para $\mathbf{P}_k$ e $\mathbf{N}_k$ são diagonalmente opostos. Dentre os oito vetores, nenhum tem as coordenadas (0,0,0), assim nunca haverá sobreposição entre resíduos consecutivos(4 com 3, 5 com 4, e assim por diante). Os sítios em $\mathbf{P}_k$ e $\mathbf{N}_k$ na rede de diamante ajustam bem resíduos numa estrutura tetraédrica, daí a utilização do conjunto de vetores $\{\mathbf{P}_k\}$ e $\{\mathbf{N}_k\}$ para o atual método de crescimento de cadeia.

Os vetores $\mathbf{P}_k$ e $\mathbf{N}_k$, num total de 8 (quatro vetores sendo a imagem especular dos outros quatro, equação 2.1) são representados por ternos de números da base ortogonal.

Na construção da cadeia carbônica adicionamos consecutivamente a sequência de vetores $\mathbf{P}_i$ e $\mathbf{N}_j$ (ou $\mathbf{N}_i$ e $\mathbf{P}_j$) alternados e obedecendo as seguintes relações (equação 2.2) dos produtos escalares

$$\mathbf{P}_i.\mathbf{N}_j = 1, \text{com } i \neq j;\ i, j = 1, 2, 3 \text{ e } 4 \qquad 2.2$$

A distância entre quaisquer par de resíduos consecutivos na estrutura tetraédrica é constante e tem como valor, ou seja, $|\mathbf{P}_k| = |\mathbf{N}_k| = (3)^{1/2}$.

A seguir detalhamos o modelo de rede tetraédrica para o início e continuação do processo de crescimento de cadeia.

2.2.1.1 Início do crescimento da cadeia

A cadeia nascente inicia-se a partir de um ponto arbitrário no espaço, que neste caso foi escolhido como sendo o ponto (0,0,0), onde é fixado o primeiro resíduo (no N-terminal), sem índice e

portanto sem nenhuma sequência entre $\mathbf{P}_k$ e $\mathbf{N}_k$. O segundo resíduo a ser colocado na cadeia tem coordenada fornecida pelo vetor $\mathbf{P}_k$, no caso foi escolhido arbitrariamente o índice "1". Já o terceiro resíduo, depende do vetor anterior (ou seja, $\mathbf{P}_1$), apenas podendo ser $\mathbf{N}_2$, $\mathbf{N}_3$ ou $\mathbf{N}_4$; não podendo ser o vetor $\mathbf{N}_1$ pois ocorrerá sobreposição reversa (entre os resíduos 3 e 1), neste caso foi escolhido arbitrariamente o índice "2".

Quando não é admitida sobreposição entre resíduos primeiros vizinhos (*i* com *i* - 2), o mesmo procedimento realizado do segundo para o terceiro resíduo é repetido para o restante da cadeia, ou seja, todos os vetores consecutivos também deverão ter sequência $\mathbf{P}_k$ e $\mathbf{N}_k$ alternada e índices distintos, pois índices iguais representam sobreposição reversa de resíduos primeiros vizinhos, como pode ser visto na inserção do quarto resíduo (Figura 2.3). Esta alternância de vetores consecutivos $\mathbf{P}_k$ e $\mathbf{N}_k$ com índices diferentes para resíduos primeiros vizinhos constitui a base do processo de crescimento de cadeia numa rede tetraédrica e evita gasto em tempo de simulação para checagem de sobreposição entre vetores consecutivos.

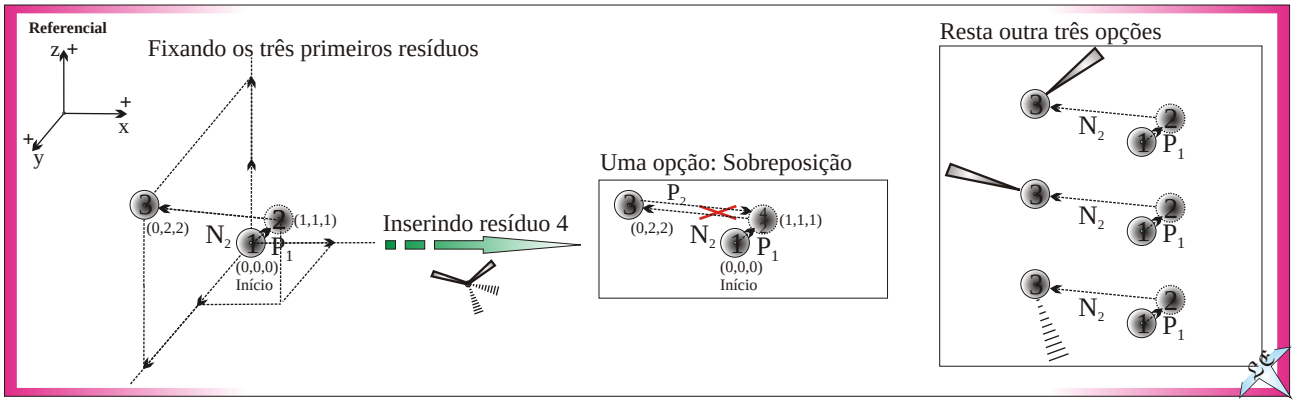


Figura 2.3 - Ilustração para modelo sem sobreposição entre resíduos primeiros vizinhos para cadeia com 4 resíduos, numa rede tetraédrica. Cada novo resíduo adicionado a cadeia deverá ter vetores consecutivos $\mathbf{P}_k$ e $\mathbf{N}_k$ alternados e com índices diferentes. No caso como o resíduo "3" foi originado do vetor $\mathbf{N}_2$, o resíduo quatro deve ser originado de um vetor $\mathbf{P}_k$ com índice diferente de "2", ou seja, dentre 4 configurações possíveis , resta 3 alternativas de crescimento, ou -1.

2.2.1.2 Continuando o crescimento da cadeia

Fixado os três primeiros resíduos, para todas as cadeias a serem geradas, continuamos crescendo a cadeia colocando sequencialmente vetores escolhidos numa sequência alternada de $\mathbf{P}_k$ e $\mathbf{N}_k$ e com índices consecutivos diferentes, agora sempre escolhidos aleatoriamente. Dessa forma uma

configuração específica para uma cadeia com *N* resíduos é construída usando *N*-1 vetores, por exemplo, gerando um fragmento de seis resíduos, com a seguinte sequência: ...$\mathbf{P}_1\mathbf{N}_2\mathbf{P}_4\mathbf{N}_1\mathbf{P}_2\mathbf{N}_3$...

Portanto as conformações são construídas, independentemente umas das outras, adicionando resíduo a resíduo até a *N*-ésima unidade, em cada comprimento, com *N* variando de quatro até 1000 resíduos.

Devido a utilização apenas de interações de curtíssima distância (esferas rígidas), não é necessário realizar o processo de termalização, e se forem geradas muitas cadeias (ou Passos Monte Carlo, ver abaixo), em cada comprimento com *N* resíduos, então as cadeias nascentes são estatisticamente independentes umas das outras. O que também leva a uma média estatisticamente significativa para as quantidades estruturais (como raio de giração e distância ponta a ponta) que medem o comportamento de cadeias poliméricas lineares na rede tetraédrica, e portanto uma maior confiabilidade nos resultados obtidos.

Ainda devido a utilização somente das interações de esferas rígidas, cada uma das três configurações permitidas (cujos vetores consecutivos possuem índices diferentes) tem o mesmo peso de Boltzmann (com energia configuracional E = 0), e a configuração com sobreposição reversa, resta apenas evitá-la ($E \rightarrow \infty$). Mais especificamente, no início as quatro configurações tem a mesma possibilidade de ocorrer, porém se for escolhida aquela com sobreposição reversa (i.e. vetores consecutivos com mesmos índices), então escolhe-se uma das outras três configurações de forma aleatória, e com a mesma probabilidade. Tal procedimento é repetido apenas para resíduos primeiros vizinhos, até o último resíduo (no C-terminal), e as cadeias nascentes originadas deste procedimento são denominadas cadeias com restrição estérica de curto alcance.

Resumindo o processo de crescimento de cadeia, temos o esqueleto da cadeia carbônica iniciando-se a partir de um átomo de carbono-alfa (C) na base de uma célula cúbica. Os outros dois átomos de carbono também são fixados de forma a determinar a sequência inicial de vetores $\mathbf{P}_i$ e $\mathbf{N}_j$ (Figura 2.4a). Do quarto átomo em diante é repetido o processo feito para o segundo e terceiro átomo, porém com escolha aleatória para os índices dos vetores alternados $\mathbf{P}_i$ e $\mathbf{N}_j$. Para cadeias sem

sobreposição entre resíduos primeiros vizinhos, necessariamente os índices dos vetores consecutivos devem ser diferentes (Figura 2.4b). Para cadeias com restrição estérica de longo alcance além das restrições impostas pela sequência de vetores consecutivos $\mathbf{P}_i$ e $\mathbf{N}_j$ (Figura 2.3) são calculadas todas as distâncias entre resíduos não ligados covalentemente e imposta a não sobreposição entre quaisquer resíduos.

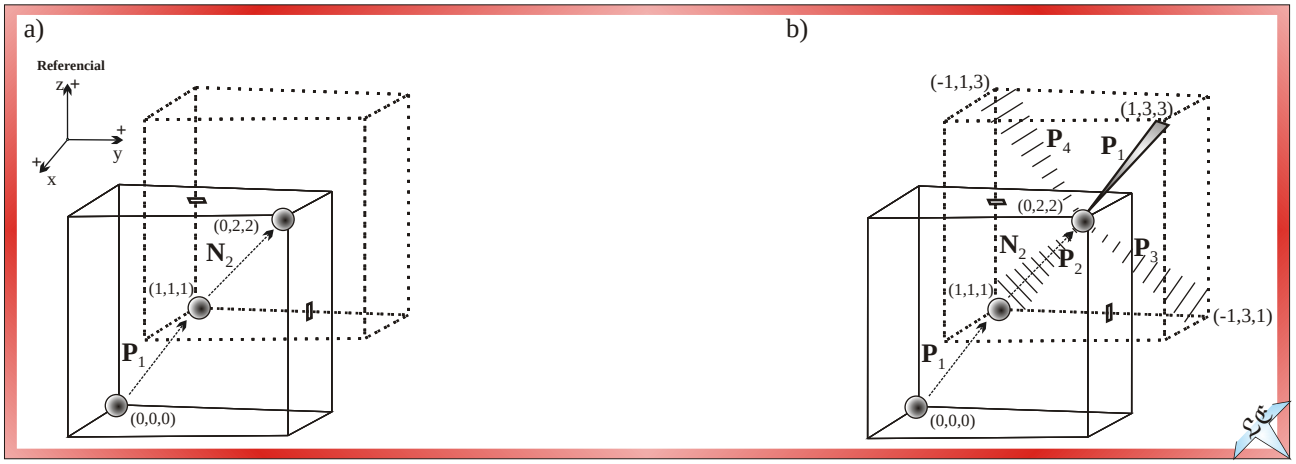


Figura 2.4 - Montagem da estrutura tetraédrica para ligação de quatro átomos de carbonos. a) Os três primeiros átomos são fixos, com o primeiro na origem (0,0,0), para o segundo é escolhido o vetor **P** com índice "1", somado as coordenadas do átomo anterior (1º átomo); já no terceiro é escolhido o vetor **N**, com índice "2", somado as coordenadas do átomo anterior (2º átomo). b) A partir do quarto átomo, na simulação, a cada novo átomo inserido na cadeia, repete-se o processo realizado para o 2º e 3º átomos, ou seja, alternar **P** e **N**, mas agora com índices escolhidos aleatoriamente, desde que diferente do vetor anterior, caso não seja permitido sobreposição entre resíduos *i* e *i*-2. Por exemplo, o quarto átomo deve ser originado do vetor **P**, pois o anterior foi **N**, com índice assumindo qualquer valor inteiro entre 1 e 4, exceto 2, já utilizado para o 3º átomo, pois dessa forma ocorreria sobreposição com o 2º átomo. O mesmo procedimento feito para o quarto átomo é realizado os demais resíduos subsequentes.

Note que para a construção de um polímero, sem sobreposição entre resíduos, escolhemos apenas um dos vetores, dentre três possibilidades (no máximo), o que para um modelo mais simples de cadeia carbônica resulta em cadeias poliméricas lineares, os alcanos. Este modelo além de simples, é eficiente em termos de tempo computacional porque tem baixo número de coordenação ( ), e.g., enquanto o cúbico simples tem = 6, o de diamante tem = 4.

A seguir são fornecidas outras informações sobre o processo de crescimento dos três tipos de cadeias utilizadas neste Capítulo, a saber: cadeia ideal, com restrição estérica de curto e longo alcance.

2.2.2 Cadeias nascentes com diferentes níveis de restrições estéricas

Através dos três tipos de cadeias é possível analisar de forma simples as implicações somente das interações de volume excluído em três níveis (ou intensidades ou alcances) de restrições para a

seleção do espaço conformacional.

2.2.2.1 Cadeia Ideal (**ci**)

A cadeia ideal (**ci**, cujas iniciais são de cadeia ideal) é o modelo mais simples para descrever um polímero, já que ele apenas assume o polímero como um passeio (ou caminhada) aleatório (a) (ou *random walk*, em inglês), e negligenciando quaisquer tipos de interação entre os resíduos. Assim todas as configurações disponíveis (ou número de coordenação) podem ser aceitas com a mesma probabilidade, ou seja, é violada deliberadamente inclusive a interação de volume excluído (esfera rígida), com dois ou mais resíduos podendo co-existir no mesmo sítio.

Embora idealizada, a generalidade da cadeia **ci** pode nos fornecer algumas informações sobre aspectos físicos básicos de polímeros, inclusive quando comparada com outros modelos mais realísticos. Note que o atual modelo de cadeia **ci** (Figura 2.5a) mantem intacto a conectividade da cadeia, expresso pela ligação covalente e geometria tetraédrica entre os resíduos que formam as cadeias nascentes, também presente nos outros dois modelos, descritos abaixo .

2.2.2.2 Cadeia com restrições estéricas de curto alcance (**ca**)

No modelo anterior (cadeia **ci**) foi assumido vetores de ligações entre resíduos independente da posição do vetor anterior, o que torna possível o vetor atual, para o resíduo *i*, se sobrepor a posição do vetor anterior, para o resíduo *i*-2. Para a cadeia com restrição estérica de curto alcance (**ca**, iniciais de curto alcance) fica suprimida a possibilidade de sobreposição entre o vetor atual e o anterior (Figuras 2.3 e 2.4). Aqui o volume é parcialmente excluído, já que o volume excluído é de curto alcance (apenas entre primeiros vizinhos, ou entre resíduos *i* e *i*-2) e não evita o vetor atual se sobrepor com outros vetores de ligação distantes ao longo da cadeia nascente, ou seja, não evita sobreposições entre resíduos não primeiros vizinhos (Figura 2.5b)

Mais especificamente, o vetor atual além de ser alternado em relação ao anterior ($\mathbf{P}_k$ e $\mathbf{N}_k$, ou $\mathbf{N}_k$ e $\mathbf{P}_k$) deve ter índices *k* diferentes, assim de configurações disponíveis, resta -1 aceitas

aleatoriamente com a mesma probabilidade. Consequentemente a cadeia nascente tem resíduos *i* não sobrepostos simultaneamente com *i*-2 e *i*+2, exceto os resíduos no N-terminal apenas com *i*+2, e no C-terminal com *i*-2.

2.2.2.3 Cadeia com restrições estéricas de longo alcance (**la**)

Um terceiro tipo de cadeia utilizada neste trabalho é decorrente da restrição estérica de longo alcance (**la**, iniciais de longo alcance). Numa primeira etapa da simulação são suprimidas as sobreposições entre vetores de ligação consecutiva (como na cadeia **ca**), e em seguida (Segunda etapa) também são suprimidas quaisquer outras sobreposições entre vetores de ligação, decorrente da interação de volume excluído de longo alcance, o que implica na ausência de sobreposição entre quaisquer resíduos. O método de gerar cadeias nascentes **la**, segundo estas duas etapas passa a ser chamado no próximo Capítulo por SAW modificado do tipo 1, ou SAWm1, neste Capítulo ainda manteremos o termo **la**.

Uma outra característica do método SAWm1 é a maneira de sair de configurações proibidas, ou armadilhas topológicas (descritas ainda neste Capítulo), onde a cadeia tem todos seus quatro sítios de crescimento ( ) já ocupados, e neste caso o crescimento da cadeia recomeça a partir do primeiro resíduo (ou seja, aquele com menor valor) que faz parte da armadilha. Aqui utilizamos o termo SAW modificado do tipo 1, pois o SAW original recomeça o crescimento da cadeia assim que ocorrer a primeira sobreposição, ao longo da cadeia, o que não é o caso aqui.

A cadeia **la** (Figura 2.5c), representa o modelo adequado para cadeias reais, pelo menos do ponto de vista estérico, já que cadeias reais além da interação de volume excluído, tem várias outras interações não covalentes, também de origem eletrostástica, conforme discutido anteriormente (seção 1.4, Capítulo 1). Embora mais realista, a utilização de cadeias **la** levam a drástico aumento do tempo de simulação, principalmente em cadeias nascentes de comprimentos acima de 300 resíduos.

Devido a importância das cadeias nascentes **la**, alguns resultados serão analisados separadamente no próximo capítulo, segundo dois métodos distintos, a saber, os métodos SAWm1 e

SAWm2, ou SAW modificado do tipo 2 (mais detalhes de ambos os método no Capítulo 3).

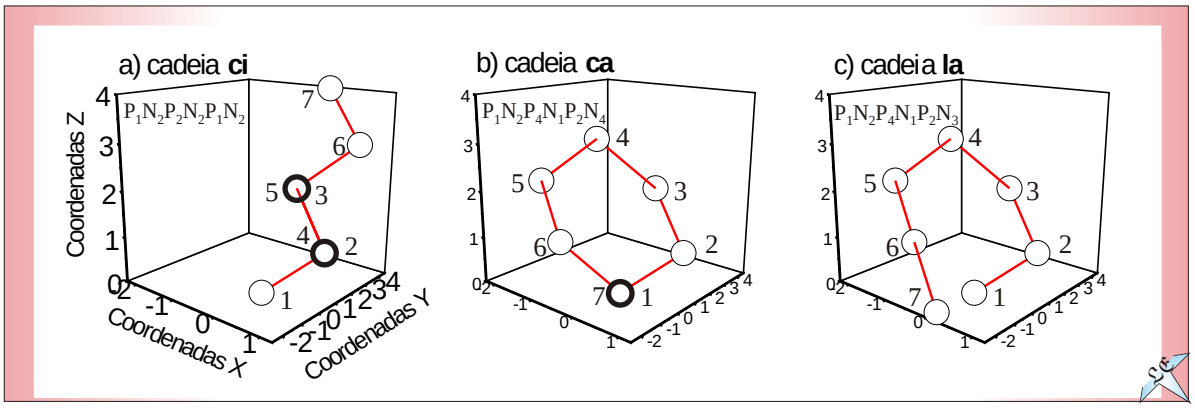


Figura 2.5 - Exemplo de cadeias (heptapeptídeos) **ci**, **ca** e **la**, com suas respectivas sequencias de vetores $\mathbf{P}_k$ e $\mathbf{N}_k$ - a) Na cadeia **ci**, há duas sobreposições consecutivas (entre resíduos 4 com 2 (primeira sobreposição possível), e 5 com 3). b) Na cadeia **ca** note que as restrições estéricas de curto alcance se estendem além dos vizinhos mais próximos ao longo da cadeia, pois ocorre sobreposições somente a partir do resíduos 7 com 1. c) Na cadeia **la** nunca é permitido a existência de sobreposições entre resíduos.

Determinado os três tipos de cadeias nascentes a serem utilizadas nas simulações Monte Carlo (MC), a seguir passamos aos cálculos de grandezas físicas que medem vários aspectos conformacionais destas cadeias, e também a atribuição de valores iniciais necessários as simulações.

2.2.3 Medidas de aspectos conformacionais

Com o intuito de observar a influência somente das restrições estéricas (esferas rígidas) nas conformações dos três tipos de cadeias nascentes, descritas acima, são utilizadas várias grandezas físicas, cujos resultados revelam diferentes aspectos conformacionais das cadeias geradas, assim como se complementam fornecendo o comportamento predominante do enovelamento de tais cadeias em cada comprimento (ou número de resíduos constituintes), *N*.

2.2.3.1 Grandezas empregadas nas simulações MC

As grandezas físicas medidas durante a simulação são: raios de giração médios; distâncias ponta a ponta médias; freqüências dos raios de giração; frequências das distâncias ponta a ponta; número de sobreposições e número de armadilhas.

Para cada conformação da cadeia nascente com *N* resíduos, o quadrado do raio de giração (R_G^2) é definido como sendo a distância ao quadrado média entre os resíduos da cadeia e seu centro de massa.

Para cadeia com resíduos *i* e *j* de massas iguais (equação 2.3) é possível mostrar que

$$R_G^2 = \left\{ \frac{1}{\frac{N(N-1)}{2}} \left[\sum_{j=i+1}^{N} \sum_{i=1}^{N-1} d_{i,j}^2 \right] \right\} \qquad (2.3)$$

onde *i* é relativo a um resíduo de referência e *j* os demais resíduos da cadeia, e $d_{i,j}$ é a distância entre o resíduo *i* e *j*.

Para *m* cadeias com *N* resíduos são geradas *M* conformações ou Passos Monte Carlo (PMC) e o raio de giração médio, $\langle R_G \rangle$ (equação 2.4) é tal que

$$\langle R_G \rangle = \left\{ \frac{1}{M} \left[\sum_{m=1}^{M} (R_{G,m}^2) \right] \right\}^{½} \qquad (2.4)$$

Ao contrário de R_G que utiliza todas as distâncias intra cadeia, entre resíduos, a distância ponta a ponta (D_{PP}) para cada conformação da cadeia nascente é simplesmente tomada como a distância entre o resíduo *i* no N-terminal e o resíduo *j* no C- terminal. Similarmente ao $\langle R_G \rangle$ (equação 2.4) também é calculada a distância ponta a ponta média, $\langle D_{PP} \rangle$.

A terceira grandeza a ser avaliada é o número de sobreposições que um novo resíduo a ser adicionado a cadeia nascente faz com outros resíduos já pertencentes a cadeia, ou seja, quantas vezes a distância entre dois resíduos quaisquer foi inferior a distância mínima de (3)½, já que nos modelos **ci** e **ca** são permitidas sobreposições entre resíduos.

A última grandeza mensurada durante a simulação é o número de armadilhas topológicas, ou simplesmente armadilhas [M. N. Rosenbluth & A. W. Rosenbluth, **1955**] para cada cadeia com *N* resíduos. Armadilha pode ser considerada como uma completa falta de alternativa para o crescimento da cadeia nascente, pois cada vez que a cadeia tenta adicionar mais um resíduo, não consegue fazê-lo, assim ficando presa num labirinto sem ter alternativa de crescimento, já que todos os sítios da rede, que poderiam ser ocupados, já estão ocupados por outros resíduos, o que viola a condição de volume excluído.

2.2.3.2 Inicialização de grandezas

Polipeptídeos pequenos podem enovelar num intervalo de 10^{-2}s, o que é muito mais rápido do que uma procura aleatória dentre todas suas configurações possíveis, mas muito mais lento do que a interconversão conformacional no estado desenovelado de 10^{-9}s. Portanto cada cadeia pode acessar até 10^{7} conformações (ou 10^{-2}s / 10^{-9}s) durante seu processo de enovelamento rápido [T. Greighton, **1995**]. Tomando como referência estes valores (embora hajam proteínas que enovelam mais rapidamente e outras mais lentamente que 10^{-2}s), no atual processo de crescimento de cadeia numa rede tetraédrica são crescidas 10^{5} cadeias para cada comprimento, ou seja, numa cadeia típica de 10^{2} resíduos são geradas no total 10^{7} conformações.

Há um expressivo custo computacional, expresso em aumento no tempo de simulação, em exceder 10^{5} cadeias (i.e., M = 100.000), principalmente quando estas se tornam mais extensas, com mais resíduos (acima de 300 resíduos), por isso fixamos o número máximo de 10^{3} resíduos (i.e., N = 1000) para as cadeias geradas e utilizamos o método SAWm1 descrito anteriormente (seção 2.2.2.3).

2.3 Resultados e discussões

A influência das restrições estéricas na compactação das cadeias nascentes **ci**, **ca** e **la**, pode ser fornecida pelos raios de giração médios ($<R_G>$ ou apenas R_G) em relação as quantidades de resíduos, *n*, destas cadeias (Figura 2.6). Aqui salientamos a diferença entre os termos configurações e conformações, já que até a conformação da cadeia com N = 1000 resíduos ter finalizado seu crescimento, são geradas configurações de comprimentos dados pela variável *n*.

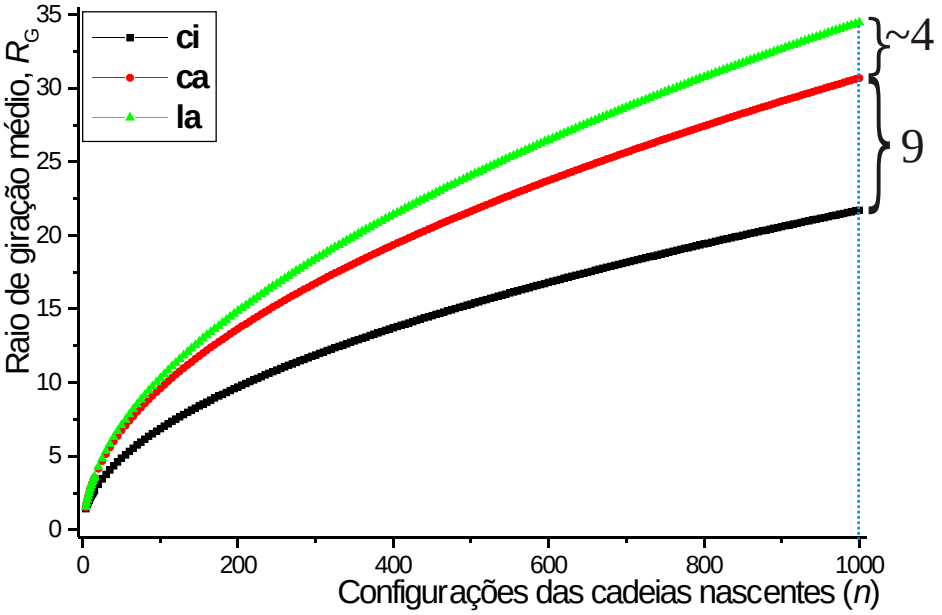


Figura 2.6 - Raio de giração médio, R_G, em função do comprimento (ou número de resíduos) *n* de configurações das cadeias nascentes, durante 10^5 Passos Monte Carlo (PMC), note que mesmo em n = N = 1000, $<R_G>$ de **ca** em relação a **ci**, ainda é cerca de duas vezes maior que **la** em relação a **ca**. A maior diferença conformacional ocorre entre cadeias nascentes dos tipos **ci** e **ca**, expresso pela maior separação entre suas curvas, desde cadeias curtas, ao redor de 30 resíduos. Por outro lado a separação entre as curvas de **ca** e **la**, começam a ficar mais nítida somente nas cadeias ao redor de 100 resíduos.
Neste gráfico foi utilizado intervalos unitários para *n* de 4 a 15, e intervalos de 5 resíduos para *n* de 15 a 1000 (i.e. 15, 20, 25...).

Com o aumento dos comprimentos (*n*) das cadeias geradas fica nítido que a maior diferença conformacional, quanto a compactação, ocorre entre cadeias nascentes dos tipos **ci** e **ca**, reforçando o papel chave desempenhado pelas restrições estéricas de curto alcance durante o enovelamento e na seleção da conformação final das cadeias. A medida que as cadeias são mais longas (a partir de100 resíduos) as restrições estéricas começam gradativamente aumentar e diferenciar sua importância, com o tipo **la** separando-se do **ca**, mas ainda assim as restrições de curto alcance continuam a desempenhar um papel predominante na determinação das conformações das cadeias nascentes, conforme expresso pela maior separação entre **ca** e **ci** do que **ca** e **la**, mesmo em N = 1000.

A variação do R_G em função do comprimento *n* (a partir de 30 resíduos) mostra que nas cadeias nascentes **ci** há conformações muito compactas (i.e. com menor R_G), devido o grande número de sobreposições acumuladas intra-cadeias. Já nas cadeias **ca**, apenas evitando uma sobreposição entre os resíduos primeiros vizinhos, *i* com *i*-2, as cadeias ficam menos compactas (com maiores R_G), e além disso há um efeito acumulativo tornando cada conformação próxima do caso real **la**, mesmo para cadeias com n > 100 resíduos.

Nas cadeias **la**, onde nunca é permitido sobreposição entre resíduos, as conformações são um pouco mais distendidas, com maiores valores de R_G, devido a presença da interação repulsiva de esfera rígida por toda a cadeia.

Similarmente ao cálculo do R_G, as mesmas conformações geradas também foram usadas para calcular a distância ponta a ponta média (D_{PP}) em função da quantidade de resíduos *n* (Figura 2.7), das cadeias nascentes.

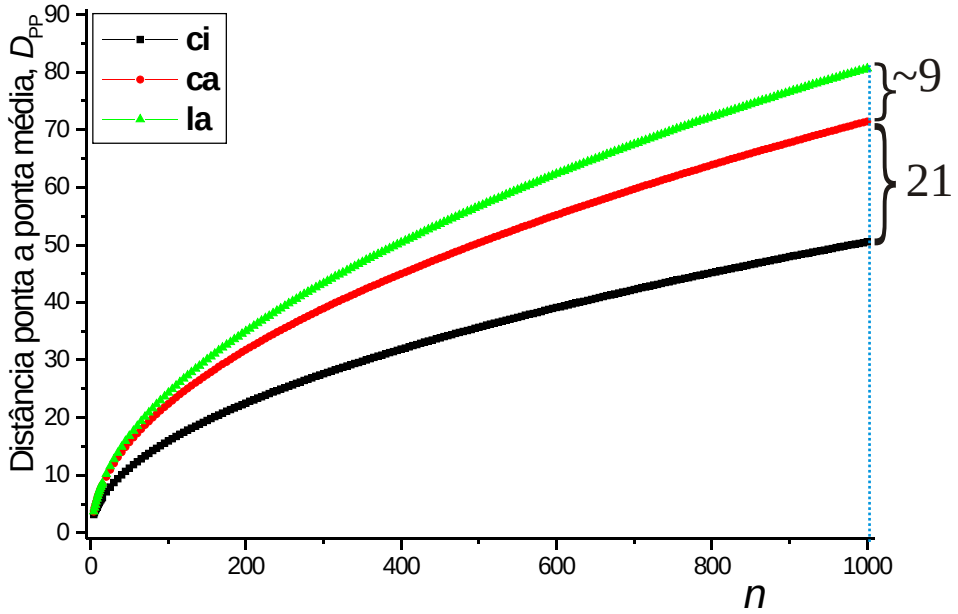


Figura 2.7 - O mesmo que a Figura 2.6, para a distância ponta a ponta média, D_{PP}, notamos mesmo em n = N = 1000, D_{PP} de **ca** em relação a **ci,** ainda é cerca de duas vezes maior que **la** em relação a **ca**. Os comportamentos da D_{PP} é similar ao R_G para os três tipos de cadeias nascentes.

O comportamento da D_{PP} é similar ao R_G, confirmando que as interações estéricas tem um efeito importante em restringir o tamanho do espaço de fase configuracional, devido principalmente as componentes de curto alcance que desempenham um papel predominante, mesmo sobre as de longo alcance conforme expresso pela maior separação entre as curvas de **ca** e **ci** do que **ca** e **la**, mesmo em N = 1000. Além disso fica confirmado os peculiares comportamentos conformacionais nos comprimentos ao redor de 30 (separação das curvas de **ca** e **ci**) e 100 resíduos (separação de **ca** e **la**), também identificado anteriormente no R_G.

Os valores dos raios de giração e das distâncias ponta a ponta mostrados até aqui (Figuras 2.6 e 2.7) são médias tomadas num tempo total de simulação de 10^5 PMC, contudo a visualização individual, em alguns comprimentos de cadeias específicos, destas duas grandezas podem nos fornecer

informações adicionais e complementares dos comportamentos das conformações geradas para os tipos **ci**, **ca** e **la**. Inicialmente mostramos a frequência (fR_G) com que cada valor de R_G aparece durante o tempo de simulação, para cadeias nascentes com 50 resíduos (Figura 2.8). Excepcionalmente para as frequências do raio de giração e distância ponta a ponta (próximos dois gráficos) estamos crescendo cadeias com N = 50, nos demais caso continuaremos com N = 1000.

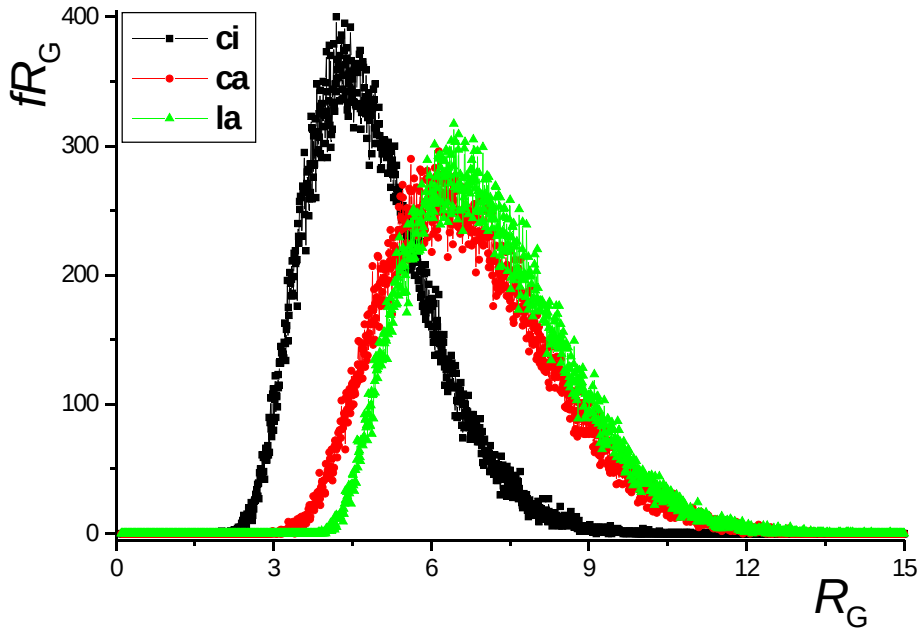


Figura 2.8 - Distribuição do raio de giração (fR_G) para cadeias nascentes com N = 50 resíduos, durante 10^5 PMC (com 10^5 conformações geradas). Nas três curvas há um deslocamento para a direita, numa estatística gaussiana (curvas log Normal). As três curvas são caracterizadas por diferentes comportamentos assimétricos e também por serem bem definidas de um ponto a outro. Para o cálculo de fR_G foi utilizado R_G no intervalo de 0,01, e.g., cadeias com R_G no intervalo entre 5,000 e 5,009; 5,010 e 5, 019; 5,020 e 5,029 são contados como 5,00; 5,01 e 5,02, respectivamente.

Observando as três curvas acima notamos que todas são deslocadas para a direita, segundo uma estatística gaussiana, conforme obtido por outros trabalhos [J. R. Banavar et. al., **2005**], mas com maior separação entre **ca** e **ci** do que **ca** e **la** (que são muito próximas entre si). As cadeias **ci** são mais compactas do que **ca** e **la**, devido a existência de muitas sobreposições entre seus resíduos, daí menores valores de R_G. As curvas **ca** e **la** são muito próximas entre si, devido as interações estéricas repulsivas de longo alcance terem eficácia comparável as de curto alcance, em cadeias curtas, até ao redor de $n = 100$ resíduos (Figura 2.6), e aqui foi utilizado cadeias com N = 50 resíduos.

Outra característica dos valores de fR_G é a assimetria das três curvas, com deslocamentos de fR_G para a direita, ou seja, os picos das três curvas ficam à esquerda de seus correspondentes valores médios R_G, cujos valores nas cadeias com 50 resíduos (Figura 2.6), para as cadeias **ci**, **ca** e **la**, são 4,86; 6,73 e

7,09, respectivamente.

Os deslocamentos para a direita, a separação e os diferentes comportamentos assimétricos das três curvas acima, é uma forma direta de contemplar a ação das interações estéricas repulsivas de esferas rígidas em diferentes níveis (ou intensidades).

As curvas de fR_G para os três tipos de cadeias nascentes são razoavelmente suaves, devido ao grande número de conformações ($M = 10^5$), e em cada conformação há muitas distâncias entre pares de resíduos, envolvidas no cálculo de R_G (equação 2.3). Para uma conformação com N resíduos, se forem excluídos os np pares primeiros vizinhos, de cada resíduo ao longo da cadeia, háverá nd distâncias não repetidas entre pares de resíduos i e j (equação 2.5), numa combinação simples, dada por

$$nd = \binom{N-np}{2} = (N-np)![(N-np-2)!2!] \qquad (2.5)$$

Note que para $N = 50$ resíduos, se não for excluído nenhum vizinho ($np = 0$), como é o caso para o cálculo do R_G, então há no total 1225 distâncias (ou $50!/(48!2!) = 50_x49/2$).

Os comportamentos observados em fR_G, Figura 2.8, pode ser complementado, nas mesmas 10^5 conformações, pela frequência (fD_{PP}) de ocorrência de cada distância ponta a ponta (D_{PP}), para cadeias com 50 resíduos (Figura 2.9).

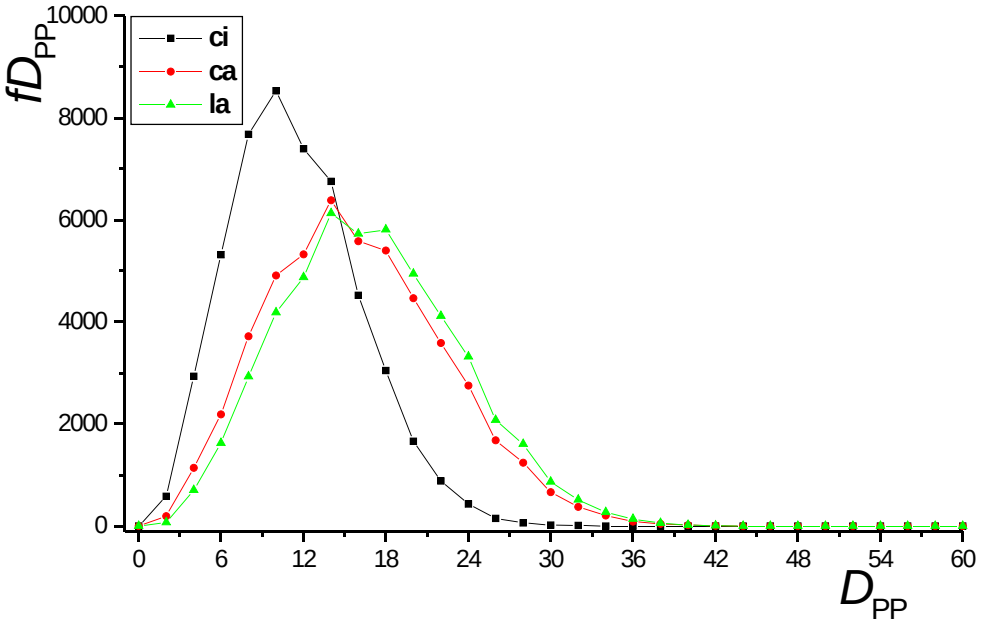


Figura 2.9 - Mesmo que Figura 2.8, mas para a distribuição da distância ponta a ponta (fD_{PP}). Embora com menos pontos, é possível verificar a separação e assimetria de fD_{PP} similar ao de fR_G para os três tipos de cadeias. Para o cálculo de fD_{PP} foram utilizados intervalos de duas unidades, por exemplo, cadeias com D_{PP} no intervalo entre 2,00 e 3,99; 4,00 e 5,99; 6,00 e 7,99 são contadas como 2; 4 e 6, respectivamente.

O comportamento das curvas da fD_{PP} são muito similares a fR_G (Figura 2.8), quanto à nítida separação entre as cadeias **ci** e aquelas de **ca** e **la**, e também quanto à assimetria das curvas, com seus picos não coincidindo com seus valores médios, D_{PP} (Figura 2.7) que são 11,25; 15,75; 16,80 para as cadeias **ci**, **ca** e **la**, respectivamente.

Há poucos pontos no histograma das curvas de fD_{PP}, mesmo para uma amostra de 10^5 cadeias, quando comparado com as curvas de fR_G, basicamente devido, em cada conformação, a D_{PP} ser calculada de apenas uma distância entre o primeiro e o último resíduo, ao contrário do R_G (equação 2.3) que é obtido a partir de todas as distância intra-cadeia, e portanto uma maior quantidade e diversificação de valores (Figura 2.8). Por exemplo para cadeias de 50 resíduos há 1225 distância entre pares de resíduo para R_G e apenas uma distância (entre os resíduos 1 e 50) para D_{PP}.

Um outro motivo para a pequena quantidade de pontos em fD_{PP}, além de apenas um valor por conformação, é a escassa diversificação deste valor, pois as cadeias estão em rede e assim os resíduos das pontas não assumem quaisquer posições no espaço tri-dimensional, consequentemente D_{PP} é naturalmente menos diversificada do que o R_G e este fato é observado no truncamento dos intervalos

em faixa duplas em fD_{PP}.

Devido a menor quantidade de pontos e os intervalos duplos, o cruzamento da curva de fD_{PP} para **ci** se dá um pouco mais deslocado para a direita levando a um cruzamento nos picos de **ca** e **la**, ao contrário de fR_G para **ci** que cruza **ca** e **la** um pouco antes de seus picos.

O papel das interações estéricas também pode ser visualizado acompanhando a quantidade de sobreposições normalizadas (soma das sobreposições divididas por 10^5), para os três tipos de cadeias, em função de seus comprimentos variáveis *n* (Figura 2.10).

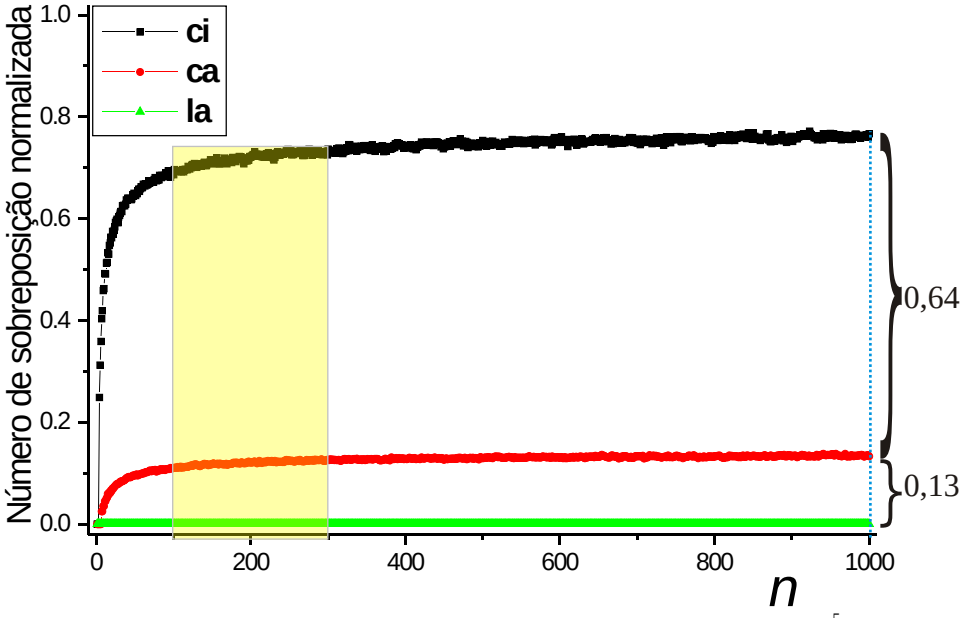


Figura 2.10 - Quantidades totais de sobreposições normalizadas (por 10^5) em função dos comprimentos (*n*) das configurações das cadeias nascentes. O número de sobreposições é reduzido em cinco vezes de **ca** e **ci** em relação a **ca** e **la**, em n ≳ 100. Nas cadeias **ci** e **ca**, podemos notar três fase de comportamentos quanto as quantidades totais de sobreposições normalizadas, até 100, de 100 a 300 (destacadas por um retângulo translúcido) e a partir de 300 resíduos. Para cadeias **ci** a primeira sobreposição ocorre no quarto resíduo, com um quarto de chance de ocorrer sobreposição, daí seu primeiro ponto em 0,25. Nas cadeias **ca** a primeira sobreposição ocorre apenas a partir do sétimo resíduo, mas agora com possibilidade de sobreposição média de apenas 0,025, devido o aumento do espaço disponível para o crescimento das cadeias.

Para cadeias nascentes **ci**, como são permitidas livremente sobreposição entre resíduos, ocorrem grandes quantidades de sobreposições. No caso das cadeias **ca**, restringindo-se apenas sobreposições entre resíduos primeiros vizinhos *i* com $i \pm 2$, há uma redução expressiva acumulada (em aproximadamente cinco vezes para N = 1000, mas tal fato começa em n ~ 100) no número de sobreposições de **ca** para **ci** em relação à **ca** para **la**. Para cadeias **la** não são permitidas sobreposições entre resíduos, então o número de sobreposições é constante e igual a zero, o que é representado por uma linha reta horizontal em zero.

As quantidades totais de sobreposições normalizadas parecem terem estabilizadas em torno de 0,80 (ou 77% para N =1000) e 0,15 (ou 13% para N =1000) para cadeia **ci** e **ca**, respectivamente. Este resultado salienta que o aumento do comprimento das cadeias, em $n \geqslant 300$ alteram pouco a chance de sobreposição, já que há muito mais espaço de crescimento disponível para a cadeia. As maiores quantidades de sobreposições de **ci** para **ca** são justificáveis, pois enquanto em **ci** a probabilidade de sobreposição pode ocorrer entre resíduos primeiros vizinhos, em **ca** existe possibilidade de sobreposição apenas entre resíduos não primeiros vizinhos, ou seja, distantes ao longo da cadeia.

Quanto a variação das quantidades totais de sobreposições normalizadas, podemos notar três fases, sendo a primeira até ao redor de 100 resíduos com grande variação nas sobreposições intra-cadeia tanto em **ci** quanto em **ca**; a seguir (segunda fase), de 100 a 300 resíduos, há uma grande redução relativa na quantidade das sobreposições, o que leva a terceira fase (após 300 resíduos) com valores de sobreposições quase estabilizados ao redor de 0,8 (ou 80%) e 0,15 (ou 15%) para cadeia **ci** e **ca**, respectivamente. Embora na terceira fase os números de sobreposições pareça constante, de fato continua crescente, mas num ritmo bastante lento.

As variações das quantidades de sopreposições em três fases refletem diretamente nas compactações das cadeias, pois a partir de 100 e 300 resíduos se inicia e se consolida a separação da compactação entre **ca** e **la**, respectivamente.

Qualquer cadeia nascente **la**, com *N* resíduos, antes de ter concluído seu processo de crescimento até o C-terminal, pode acessar configurações proibidas, denominadas armadilhas topológicas, ficando assim sem alternativas de crescer (Figura 2.11a) e portanto parando seu processo de crescimento, já que fica apenas com *n* resíduos (sendo n < N = 1000), devido a condição de volume excluído. Para cada cadeia com *n* resíduos pode-se quantificar o número de armadilhas topológicas acessadas durante a simulação de 10^5 PMC (Figura 2.11b).

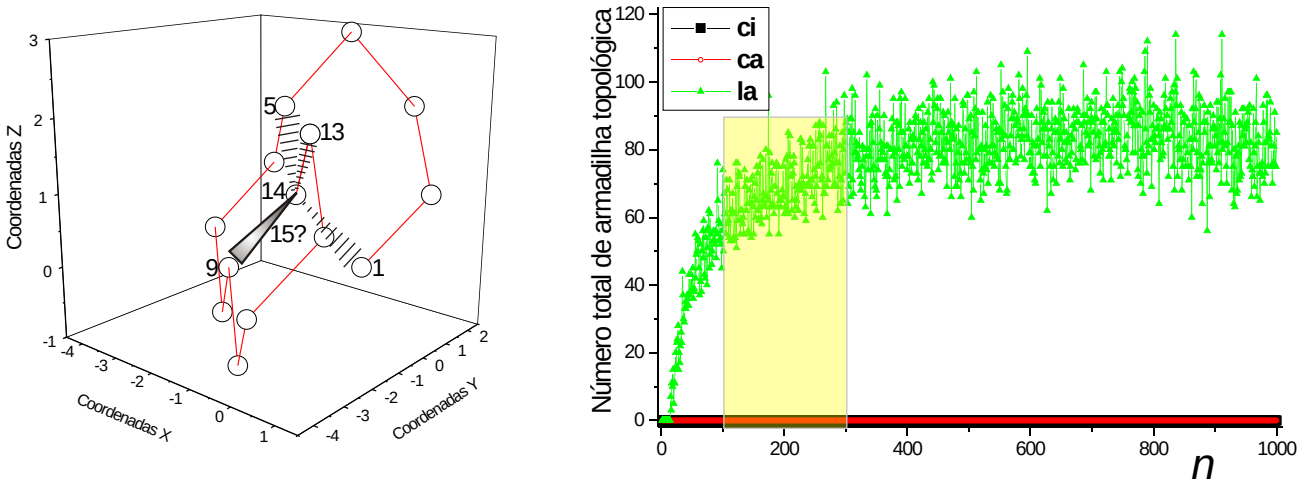


Figura 2.11 - (a) Configuração tridimensional da cadeia **la** com 15 resíduos, em armadilha topológica, e (b) Quantidades totais de armadilhas topológicas em função dos comprimentos *n*, durante 10^5 PMC. (a) Nas cadeias **la** com 15 resíduos ocorrem as primeiras armadilhas, totalizando 7 armadilhas durante 10^5 PMC, a configuração mostrada acima é referente a primeira destas armadilhas ao se tentar inserir o resíduo 15, que se sobrepõe com os resíduos 1;9;5;13. (b) Podemos identificar três comportamentos para os números de armadilhas, até 100 , de 100 a 300 (destacadas por um retângulo translúcido) e a partir de 300 resíduos. Apenas cadeias **la** possuem armadilhas topológicas e sempre que uma cadeia acessar uma configuração com armadilha topológica num determinado PMC uma nova cadeia inicia novamente seu crescimento (pelo método SAWm1), no mesmo PMC.

As cadeias **ci** e **ca** permitem sobreposições entre resíduos e portanto não possuem armadilhas topológicas, daí suas quantidades de armadilhas dadas por linhas retas horizontais em zero. No caso das cadeias **la**, há muitas armadilhas pois não são permitidas sobreposições entre resíduos.

As quantidades totais de armadilhas (Figura 2.11b) nas cadeias **la** tem curvas similares (logarítmicas) as quantidades de sobreposições nas cadeias **ci** e **ca** (Figura 2.10), pois em ambas podemos identificar três comportamentos distintos: até 100; de 100 a 300 e a partir de 300 resíduos.

Até 100 resíduos, há grandes variações nos valores das armadilhas, de 100 a aproximadamente 300 resíduos há uma drástica redução na tendência crescente das armadilhas. A partir de 300 resíduos o número de armadilhas continua aumentando muito lentamente, mas ficando, em média, quase na horizontal.

O comportamento logarítmico é basicamente devido a maneira como o método SAWm1 utilizado aqui para gerar cadeias **la** (seção 2.2.2.3), sai do armadilhamento para concluir o processo de crescimento da cadeia, em N = 1000. Uma das principais características do método SAWm1 é sair do armadilhamento a partir do primeiro resíduo que faz parte da armadilha, no caso da Figura 2.11a o crescimento reinicia em n=5 (e não em n =1, pois os três primeiros resíduos são fixos). Ocorre que em cadeia curtas não faz muita diferença como sair da armadilha topológica, mas em cadeias longas (e.g.

com 1000 resíduos) isto muda muito a forma da curva do número de armadilhas, conforme mostraremos mais detalhadamente no próximo capítulo.

As quatro grandezas (R_G, D_{PP}, número de sobreposições e armadilhas) utilizadas para observar o comportamento das conformações das cadeias nascentes **ci**, **ca** e **la** fornecem um panorama global sobre a influência das restrições estéricas em três níveis de intensidades (nenhum, curto e longo alcance) no enovelamento destas cadeias.

As quatro grandezas individualmente medem aspectos particulares da formação e consolidação das conformações das cadeias nascentes, por outro lado, em conjunto estas grandezas fornecem informações úteis ao entendimento de proteínas reais, como as mudanças configuracionais que ocorrem nos comprimentos *n* ao redor de 30 (Figuras 2.6 e 2.7), 100 e 300 resíduos. Por acaso ou não, em torno de 30 resíduos as estruturas começam a se tornar mais estáveis, tal que neste comprimento alguns autores já começam a se referir a estas estruturas como proteínas e não mais como peptídeos. Em 100 resíduos já é praticamente consolidado o termo proteína. Em 300 resíduos temos o comprimento médio das proteínas atualmente disponibilizadas no PDB.

2.4 Conclusões

A rede tetraédrica, quando comparada com a tradicional rede cúbica simples, além de mais realista (por representar a geometria do C), também fornece menos possibilidade de ocupar o espaço disponível para as configuracões com 4 e 6 possibilidades (ou número de coordenação), respectivamente. Isso provavelmente torna as sobreposições relativamente maiores na rede tetraédrica, e portanto os efeitos das restrições estéricas de curto e longo alcance podem ser notados mais nitidamente nas cadeias nascentes.

Mostramos que os vínculos geométricos impostos pela conectividade da cadeia (ligação covalente e rede tetraédrica), juntamente com as restrições estéricas [J. R. Banavar et al., **2002**] podem parcialmente levar a uma seleção no espaço de estruturas [C. Levinthal, **1968**], e influenciar

significativamente a compactação da cadeia (curvas de R_G e D_{PP} com comportamentos distintos em **ci**, **ca** e **la**), independente de detalhes químicos.

Apesar da nítida diferença em significado físico, as quantidades de sobreposições nas cadeias **ci** e **ca** são similares as quantidades de armadilhas nas cadeias **la**, basicamente com as mesmas formas das curvas (logarítmicas), em três fases: (i) Até 100 resíduos, ambas são não nulas a partir de um determinado comprimento, e.g. no caso das armadilhas isso inicia em 15 resíduos, e além disso esta fase é marcada por grandes variações de valores, nas duas grandezas; (ii) De 100 a 300 resíduos, há uma drástica redução na tendência de aumento das quantidades de sobreposições e armadilhas, o que leva a terceira fase (iii), a partir de aproximadamente 300 resíduos, a curva muda seu formato, ficando, em média, quase na horizontal, mas ainda aumentando muito lentamente seus valores, já que surgem muitas novas possibilidade de ocupar sítios da rede ainda desocupados.

As restrições estéricas locais (nas cadeias **ca**) mostraram se estender além dos vizinhos mais próximos ao longo da cadeia e assim restringindo significativamente o tamanho do espaço conformacional disponível para a cadeia [R.V.Pappu et al., **2000**], independente dos comprimentos das cadeias nascentes. Já as restrições estéricas de longo alcance [C. R. Cantor & P. R. Schimmel, **1980**], por sua vez impõem restrições adicionais ao espaço conformacional, além dos resíduos adjacentes, principalmente em cadeias com $n \gtrsim 100$.

As restrições estéricas repulsivas de curto e longo alcance ao longo da cadeia, quando comparadas com a ausência destas restrições na cadeia ideal, parecem desempenhar um papel relevante em influenciar significativamente o crescimento, o enovelamento e a compactação das cadeias nascentes, e em criar as condições para a ação de outras interações não covalentes (não tratadas aqui) e que dão o caráter específico das conformações nativas, e que por fim levam a atividade funcional da proteína.

Capítulo 3: Influências das armadilhas topológicas na compactação e tempo de enovelamento de cadeias nascentes longas em rede tetraédrica

3.1 Introdução

Neste Capítulo analisaremos como as restrições estéricas (ou interações de esferas rígidas) de longo alcance **la**, juntamente com duas diferentes abordagens (que diferem no tratamento das armadilhas topológicas) de gerar cadeias nascentes, influenciam os processos de crescimento e enovelamento destas cadeias. Para tal finalidade utilizaremos algumas grandezas físicas que medem diferentes aspectos das cadeias geradas, são elas as quantidades e frequências das armadilhas topológicas, o raio de giração, o expoente de escalonamento e o tempo de processamento.

O raio de giração (R_G) tem sido amplamente utilizado para monitorar a compactação e a estabilidade estrutural de polímeros randômicos à proteínas, já que é razoavelmente fácil de ser medido, pois tem significado físico e expressão matemática simples, além de ser obtido experimentalmente por espalhamento de Raio-X em baixo ângulo (ou SAXS, do inglês).

Significativas variações no R_G são decorrentes de mudanças estruturais associadas ao tamanho e compactação da estrutura analisada. A média configuracional obtida pelo raio de giração médio ($<R_G>$, ou apenas R_G) segue uma lei de potência (equação 3.1) com o comprimento ou número de resíduos, N [M. A. Moret et al., **2009**], através de

$$R_G \sim kN \qquad 3.1$$

onde k é uma constante e o expoente de escalonamento , assume valores diferentes dependente do tipo de cadeia (ideal, compacta e distendida ou inchada) e condições do solvente em torno da cadeia [L. Hong & J. Lei, **2009**].

O expoente costuma variar entre dois casos extremos [G. A. Arteca, **1994**; G. A. Arteca, **1995**; G. A. Arteca, **1996**; J. Liang & K. A. Dill, **2001**], em ~ 1/3 (ou 0,333) para homopolímero (se todos os

resíduos forem hidrofóbicos) em condições de mau solvente, e ~ 3/5 (ou 0,60) para homopolímero (somente com resíduos hidrofílicos) em condições de bom solvente. Para proteínas compactas gerais, em condições fisiológicas, ~ 2/5 (ou 0,4); para cadeia ideal em condições de mau solvente, assume valor 0,5; e polímeros lineares em soluções diluídas de bom solvente, ~ 0,588.

O processo de crescimento de cadeia nascente numa estrutura tetraédrica com restrição estérica de longo alcance (**la**), pelo método SAWm1 já foi comentado no Capítulo 2, porém na seção Métodos abaixo fornecemos mais algumas informações adicionais deste método e também do método SAWm2.

3.2 Métodos

Nós utilizamos o Método Monte Carlo para crescer cadeias nascentes de comprimento *N* até 2250 resíduos numa rede tetraédrica [G. Zifferer, **1990**] e com estes resíduos interagindo somente via restrições estéricas, pelos métodos SAWm1 e SAWm2.

Os métodos SAWm1 e SAWm2 utilizados neste Capítulo fazem uso das restrições estéricas de curto e longo alcance (**ca** e **la**, Capitulo 2), com a de longo contendo a de curto alcance, e assim excluindo propositalmente as demais interações existente no polímero [N. Giovambattista et al., **2008**], ou seja, é analisado somente a influência das restrições estéricas no processo de crescimento de cadeia na rede tetraédrica. Ambos os métodos se diferenciam apenas na maneira como são corrigidas as armadilhas topológicas, para concluir o crescimento da cadeia (Figura 3.1).

No método SAWm1 (ou SAW modificado do tipo 1, Capítulo 2) quando a cadeia no resíduo *n* fica sem alternativa de crescimento, pois ocorre sobreposição deste com outros quatro resíduos (nos quatro sítios da coordenação), então é verificado qual destes quatro resíduos foi inserido inicialmente, ou primeiro, na cadeia e recomeçamos o crescimento a partir deste resíduo.

No método SAWm2 (ou SAW modificado do tipo 2) se a cadeia ficar armadilhada, recomeçamos a cresce-lá do início, ou seja, a partir do quarto resíduo, já que os três primeiros resíduos são fixos.

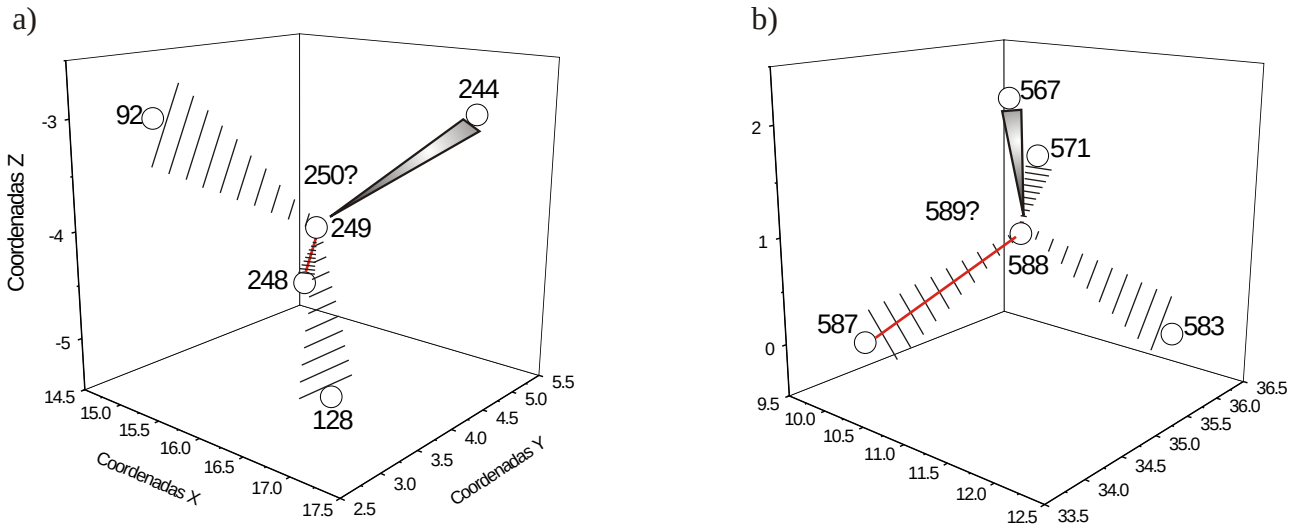


Figura 3.1 - Exemplo de duas configurações (obtidas durante a simulação Monte Carlo) que não conseguem concluir seus crescimentos devido ficarem armadilhadas e que são tratadas diferentemente pelos métodos SAWm1 e SAWm2. No método SAWm1 recomeçamos o crescimento a partir do resíduo inserido inicialmente na armadilha, no caso da configurações acima em 92 (a) e 567 (b), respectivamente; já em SAWm2 recomeçamos a crescer ambas do resíduo 4.

Observando as duas configurações acima, notamos que na primeira (Figura 3.1a) os dois métodos, parecem não se diferenciarem muito já que entre começar dos resíduos 92 (SAWm1) e 4 (SAWm2) há pouca diferença, com 88 unidades; já na segunda configuração (Figura 3.1b), tal diferença se eleva ficando em 563 unidades.

Nos resultados anteriores (Capítulo 2), utilizamos apenas o método SAWm1 para fazer o crescimento de cadeias com restrição estérica de longo alcance, **la**. Neste capítulo utilizaremos os métodos SAWm1 e SAWm2 para crescer cadeias **la**, com o objetivo de medirmos a influência das armadilhas topológicas na compactação e no tempo de enovelamento de cadeias nascentes em rede tetraédrica. Consequentemente acabamos fazendo um estudo comparativo entre os dois métodos, já que ambos diferem apenas nas abordagens das armadilhas topológicas ou geométricas [M. N. Rosenbluth & A. W. Rosenbluth, **1955**]. Até a primeira armadilha (num determinado instante ou PMC, que ocorre bem no início das simulações) ambos os métodos geram as mesmas configurações, mas a partir deste instante, as configurações passam a ser completamente diferentes e independentes entre si, já que são determinadas segundo um processo estocástico.

3.3 Resultados e discussões

A maneira como se realiza o crescimento das cadeias nascentes, com diferente enfoque nas armadilhas topológicas, como feito independentemente nos métodos SAWm1 e SAWm2, pode ou não implicar em grandes mudanças conformacionais, conforme avaliado pela compactação e tempo de enovelamento.

As medidas para verificar as armadilhas são feitas durante o crescimento da cadeia, em cada configuração *n* (com *n* resíduos), já as medidas na compactação são feitas após concluído o crescimento da cadeia nascente, após cada PMC.

3.3.1 Armadilhas topológicas

Para analisarmos os efeitos das armadilhas geramos cadeias nascentes desde curtas, com dezenas de resíduos, até muito longas, com milhares de resíduos, simuladas em *m* tentativa de crescimento até *M* igual a 10^5 conformações (ou PMC) para cada comprimento total *N*. Devido a riqueza de informações que podemos obter das armadilhas topológicas, elas são avaliadas por diferentes aspectos (pelo total de quatro variáveis), mostrados separadamente a seguir, a começar pela frequência de suas ocorrências.

3.3.1.1 Frequência das Armadilhas topológicas

Para a medida das armadilhas escolhemos os comprimentos (máximos) de referência *N* de 50, 300, 500, 1000, 1500 e 2250 resíduos. Em cada *N* e para cada resíduo armadilhado *nr*, é medido a quantidade total de armadilhas topológicas, $fA_{rm}(nr)$ ou apenas fA_{rm} (equação 3.2), durante o crescimento da cadeia, tal que quando ocorrer uma armadilha (A_{rm}) numa configuração *nr*, temos $A_{rm} = 1$, e do caso contrário $A_{rm} = 0$.

$$fA_{rm} = \sum_{m=1}^{M} \sum_{nr=1}^{N} A_{rm} \qquad 3.2$$

Devido as somatórias em *m* e *nr* note que fA_{rm} para todos os resíduos armadilhados *nr* em cada

comprimento *N* pode ser superior a 10^5, o número de conformações geradas. A influência das armadilhas topológicas nas conformações das cadeias nascentes pode ser avaliada por suas ocorrências, em função de cada comprimento *nr* num total de *M* conformações, nos métodos SAWm1 e SAWm2 (Figura 3.2).

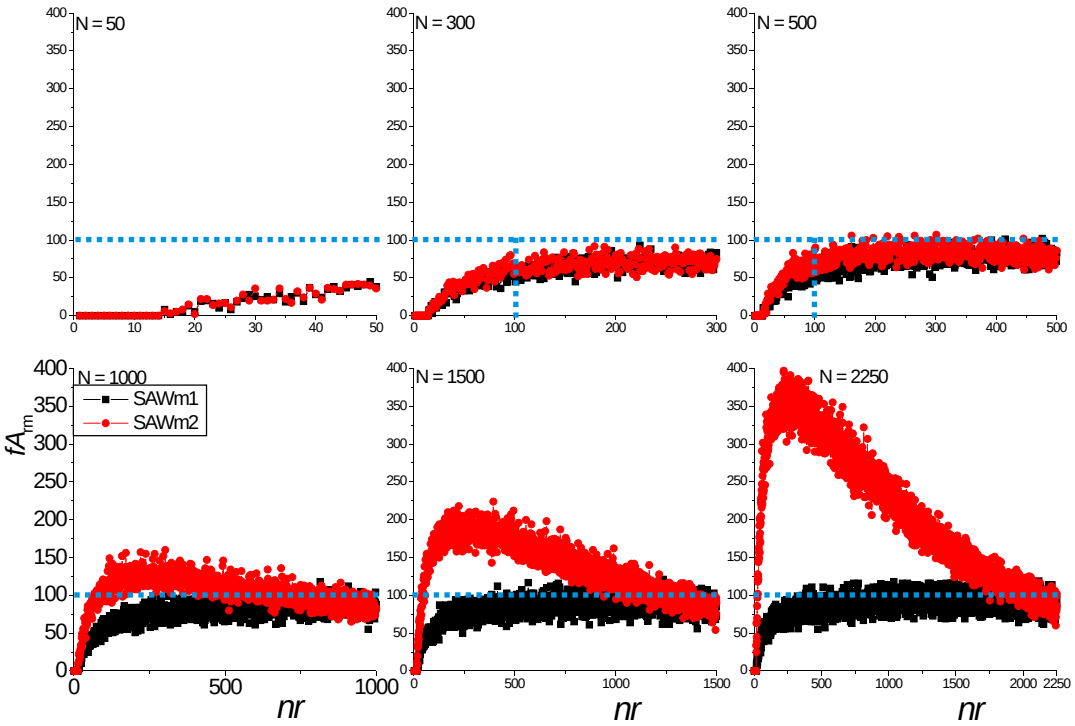


Figura3.2 - Frequência das armadilhas (fA_{rm}) para cada resíduo armadilhado no comprimento *nr* para cadeias nascentes em seis comprimentos totais *N*, segundo os métodos SAWm1 e SAWm2 - É possível identificar três comportamentos para as ocorrências das armadilhas fA_{rm} em função de *nr*, em cadeias curtas (N = 50), intermediárias (N = 300 e 500) e longas (N $\gtrsim$ 1000). O valor de fA_{rm} igual a 100 (destacado por linhas horizontais pontilhadas) muda pouco a partir de nr =100 (linhas verticais pontilhadas, N=300 e 500) para *N* de 500 até 2250, pelo método SAWm1. Note que as escalas na abcissa variam de acordo com o comprimento *N*.

A ocorrência das armadilhas (fA_{rm}) em função do comprimento *nr* revelam três comportamentos distintos para estas configurações proibidas, de acordo com cadeias curtas (N=50), intermediárias (N = 300 e 500) e longas (N $\gtrsim$ 1000).

Em cadeias curtas (N=50) há relativamente poucas armadilhas se comparadas com outros comprimentos, por exemplo, no método SAWm1 há no total 846 e 14913 armadilhas para *N* igual a 50 e 300, respectivamente. A pequena e crescente quantidade de armadilhas fA_{rm}, é decorrente do pequeno, mas também crescente espaço configuracional disponível para o crescimento das cadeias. O surgimento das armadilhas começam a partir de 15 resíduos (Figura 2.11). O aumento de fA_{rm} tende a reduzir seu ritmo crescente e ser mais estável em nr $\gtrsim$ 100, o que pode ser observado nos outros cinco

comprimentos totais *N*.

Para cadeias com comprimentos intermediários (N = 300 e 500) há um aumento no espaço configuracional para as cadeias geradas, daí fA_{rm} aumenta apenas 25 unidades de N = 50 para 300 e deste para 500 resíduos. Aliás as frequências fA_{rm} passam a ter dois comportamentos distintos de acordo com o comprimento *nr*, tanto em SAWm1 quanto SAWm2. No primeiro, até cerca de *nr* igual a 100, há uma crescente ocorrência de armadilhas (como em N = 50), a partir de então, num segundo comportamento fA_{rm} reduz um pouco seu ritmo de crescimento, atingindo um patamar de 75 e 100 para N igual a 300 e 500, respectivamente. A estabilização de fA_{rm}, em 75 e 100, implica em muitas armadilhas acumuladas o que determina o comportamento de fA_{rm} para cadeias longas.

Em cadeias nascentes longas ($N \geqslant 1000$), há uma nítida diferença entre os resultados fornecidos pelos dois métodos, tornando-se equivalentes somente próximos ao término do crescimento das cadeias, com resultados praticamente inalterados para SAWm1. Em SAWm2 ocorre muitas armadilhas principalmente no intervalo entre *nr* de 100 e 500, com pico em torno de 150, 200 e 250 para *N* de 1000, 1500 e 2250, respectivamente. A lacuna existente entre as duas curvas é devido o aumento significativo no número de armadilhas, até ser concluído o crescimento da cadeia em cada PMC, devido o procedimento utilizado em SAWm2, pois ao recomeçar do início o crescimento da cadeia são encontradas múltiplas armadilhas até que seja finalmente concluído seu processo de crescimento em *N*.

Tais resultados destacam a eficácia no procedimento de gerar cadeias nascentes longas do método SAWm1 sobre SAWm2, já que o número reduzido de armadilhas é um importante ingrediente na geração de cadeias, no ambiente celular, que em geral costuma ocorrer num intervalo de tempo curto.

A seguir avaliamos as causas para a diferença de comportamento para fA_{rm} em cadeias longas entre SAWm1 e SAWm2. Para isso utilizaremos grandezas diferentes, a começar pela distribuição do número de armadilhas. Após concluída a análise das armadilhas, passamos a verificar as implicações das diferenças entre SAWm1 e SAWm2 para a compactação e o tempo de simulação.

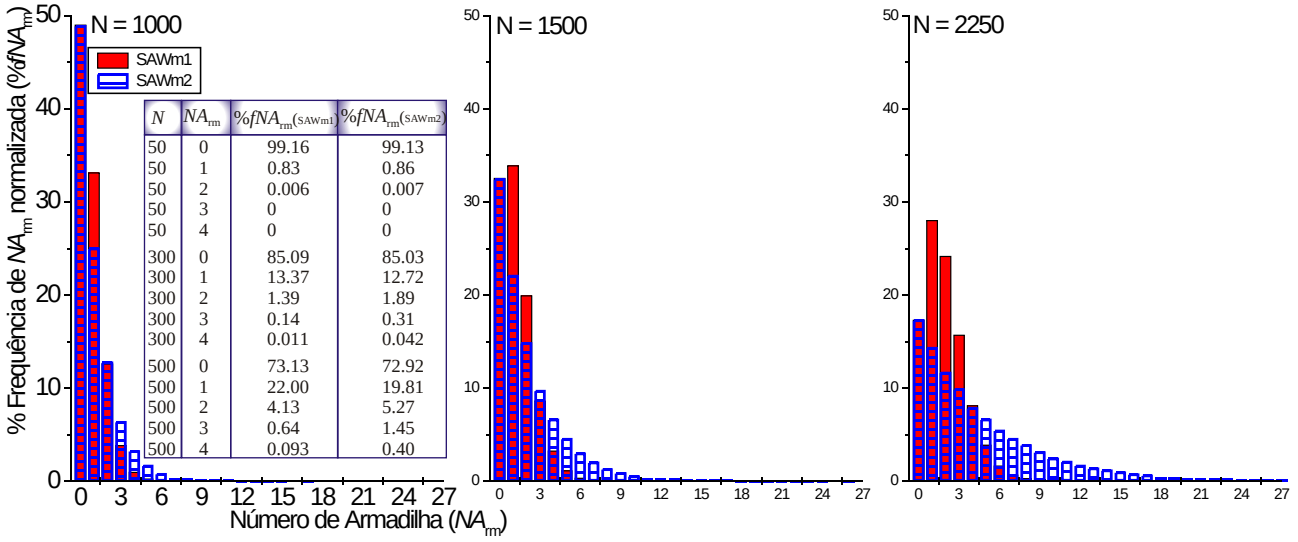
3.3.1.2 Distribuição do número de armadilhas

A distribuição do número de armadilhas (NA_{rm}) que ocorre durante o crescimento das cadeias de comprimento *N* em cada *m* (ou PMC) pode revelar, pelo menos parte das diferenças entre fA_{rm} para SAWm1 e SAWm2. Por exemplo, para N = 50 em SAWm1, há no total 846 armadilhas (dado pela soma total de fA_{rm}, em todos os comprimentos *nr*, Figura 3.2) distribuídas em 834m (com $NA_{rm} = 1$) e 6m (com $NA_{rm} = 2$); já para SAWm2 há 870 armadilhas dispostas em 856m ($NA_{rm} = 1$) e 7m ($Na_{rm} = 2$), ou seja, os valores de NA_{rm} para N = 50 são próximos, daí também a equivalência entre os dois métodos (Figura 3.2).

Para cadeias com *N* resíduos, cada valor de NA_{rm} (= 0, 1, 2, 3, 4 ...) mede quantas armadilhas foram encontradas em cada PMC (num total de 10^5 PMC) até que seja concluído o crescimento até o C-terminal. Já $\%fNA_{rm}$ (NA_{rm}) ou apenas $\%fNA_{rm}$, mede a distribuição de NA_{rm}, i.e., quantas vezes cada valor de NA_{rm} ocorreu, dividido por 1000, pois ela é feita em percentagem e em relação a 10^5 PMC (i.e., $1/10^3 = 10^2/10^5$, equação 3.3), nos métodos SAWm1 e SAWm2.

$$\%fNA_{rm} = \frac{1}{10^3} \sum_{m=1}^{M} NA_{rm} \qquad 3.3$$

Note que para cadeias nascentes com 50 resíduos, cujos valores de NA_{rm} foram mostrados acima e utilizando a equação 3.3, em SAWm1 para NA_{rm} igual 1 e 2 temos $\%fNA_{rm}$ de 0,83 (ou $834/10^3$) e 0,006 (ou $6/10^3$), respectivamente. Utilizando procedimento semelhante ao de *N* igual a 50 em outros comprimentos de cadeia para SAWm1 e SAWm2, temos a distribuição de NA_{rm} expressa por $\%fNA_{rm}$ (Figura 3.3).



N	NA_{rm}	$\%fNA_{rm(SAWm1)}$	$\%fNA_{rm(SAWm2)}$
50	0	99.16	99.13
50	1	0.83	0.86
50	2	0.006	0.007
50	3	0	0
50	4	0	0
300	0	85.09	85.03
300	1	13.37	12.72
300	2	1.39	1.89
300	3	0.14	0.31
300	4	0.011	0.042
500	0	73.13	72.92
500	1	22.00	19.81
500	2	4.13	5.27
500	3	0.64	1.45
500	4	0.093	0.40

Figura 3.3 - Histogramas com a frequência percentual normalizada (%fNA_{rm}) do número de armadilha (NA_{rm}) para 10^5 PMC nos métodos SAWm1 e SAWm2 em conformações com N de 1000, 1500 e 2250, e na Tabela ao lado do 1º histograma para N de 50, 300 e 500. Note que para cadeias com N pequeno ou intermediário (Tabela) há valores muito próximos de %fNA_{rm} para os dois métodos o que justifica número de armadilhas próximos para fA_{rm} (Figura 3.2), contudo para cadeias com N grande (Histogramas) fNA_{rm} para SAWm1 e SAWm2 são muito diferentes, daí consequentemente as grandes diferenças para fA_{rm} nos dois métodos.

Note que para cadeias de comprimentos pequeno ou intermediário (Tabela na Figura 3.3) há valores muito próximos de %fNA_{rm} para SAWm1 e SAWm2, daí a utilização apenas de NA_{rm} pequenos (até 4), o que justifica valores equivalentes para fA_{rm} nos dois métodos (Figura 3.2). Para cadeias com comprimentos grandes (Histogramas na Figura 3.3), em geral, os valores de %fNA_{rm} para muitos NA_{rm} nos dois métodos são muito diferentes, daí também as diferenças para fA_{rm}.

Em NA_{rm} nulo (ou seja, sem armadilhas), tanto SAWm1 quanto SAWm1, fornecem fNA_{rm} próximas, saindo de 99% para cerca de 17%, em N de 50 e 2250, respectivamente. Note que para cadeias de comprimentos longos ($N \gtrsim 1000$) a equivalência entre os dois métodos, para NA_{rm} diferente de zero, é deslocada para a direita, à medida que o comprimento da cadeia aumenta, em 2, 3 e 4 armadilhas para N de 1000, 1500 e 2250, respectivamente.

A distribuição de NA_{rm} nos dois métodos para cadeias longas ($N \gtrsim 1000$) também é diferente, pois em SAWm2 os valores de %fNA_{rm} para NA_{rm} nulo é sempre maior que NA_{rm} não nulo, embora tais diferenças vão sendo gradativamente reduzidas, além disso %fNA_{rm} são deslocadas para a direita, para grandes NA_{rm}. Já em SAWm1 para $N \gtrsim 1500$ observamos NA_{rm} não nulo ultrapassando NA_{rm} nulo, além

disso $\%fNA_{rm}$ de SAWm1 são deslocadas para a esquerda, ou seja, com predomínio de poucas NA_{rm}, em relação a SAWm2. Tais resultados salientam as grandes diferenças para fA_{rm} nos dois métodos para cadeias longas.

Os resultados de $\%fNA_{rm}$ em função de NA_{rm} podem ser confirmados acompanhando o último, ou máximo NA_{rm}, representado por máx$\{NA_{rm}\}$, para $\%fNA_{rm}$ não nula nos seis comprimentos N de referência analisados até aqui.

3.3.1.3 Número máximo de armadilhas com frequência não nula

O máx$\{NA_{rm}\}$ com $\%fNA_{rm}$ não nula pode novamente salientar as grandes diferenças encontradas para os métodos SAWm1 e SAWm2 (Tabela 3.1).

N	50	300	500	1000	1500	2250
máx$\{NA_{rm}$ (SAWm1)$\}$	2	5	6	9	12	15
máx$\{NA_{rm}$ (SAWm2)$\}$	2	5	10	17	31	73

Tabela 3.1 - Variação de máx$\{NA_{rm}\}$ com $\%fNA_{rm}$ diferente de zero para seis comprimentos N, nos métodos SAWm1 e SAWm2. É importante destacar que para cadeias com N $\geqslant$ 500 e principalmente em SAWm2 existem alguns NA_{rm} com $\%fNA_{rm}$ nulos antes dos valores mostrados nesta Tabela, por exemplo, para SAW em N = 500 temos NA_{rm} = 10, mas antes em 9 há $\%fNA_{rm}$ = 0. Note porém que depois dos valores mostrados nesta tabela, sempre temos $\%fNA_{rm}$ nulo.

Pelos valores de máx$\{NA_{rm}\}$ com $\%fNA_{rm}$ não nulos, também é possível justificar as grandes diferenças observadas em fA_{rm} nos métodos SAWm1 e SAWm2 (Figura 3.2), principalmente em cadeias longas (N $\geqslant$ 1000), pois quanto mais armadilhas são encontradas até se concluir o crescimento da cadeia, conforme medido por $\%fNA_{rm}$ não nulos para NA_{rm} grandes, também maiores as frequências fA_{rm}.

Além dos valores máximos de máx$\{NA_{rm}\}$ com $\%fNA_{rm}$ não nulos (Tabela 3.1) podemos também acompanhar a variação do número total de armadilhas ocorridas durante o crescimento das cadeias para os métodos SAWm1 e SAWm2.

3.3.1.4 Número total de armadilhas

O número total de armadilhas, fA_{rm}, ou seja, soma total de fA_{rm} (Figura 3.2) em todos os comprimentos *nr*, dividido por 10^5 PMC, para cada comprimento *N* (Figura 3.4) pode revelar novos aspectos das diferenças entre os métodos SAWm1 e SAWm2.

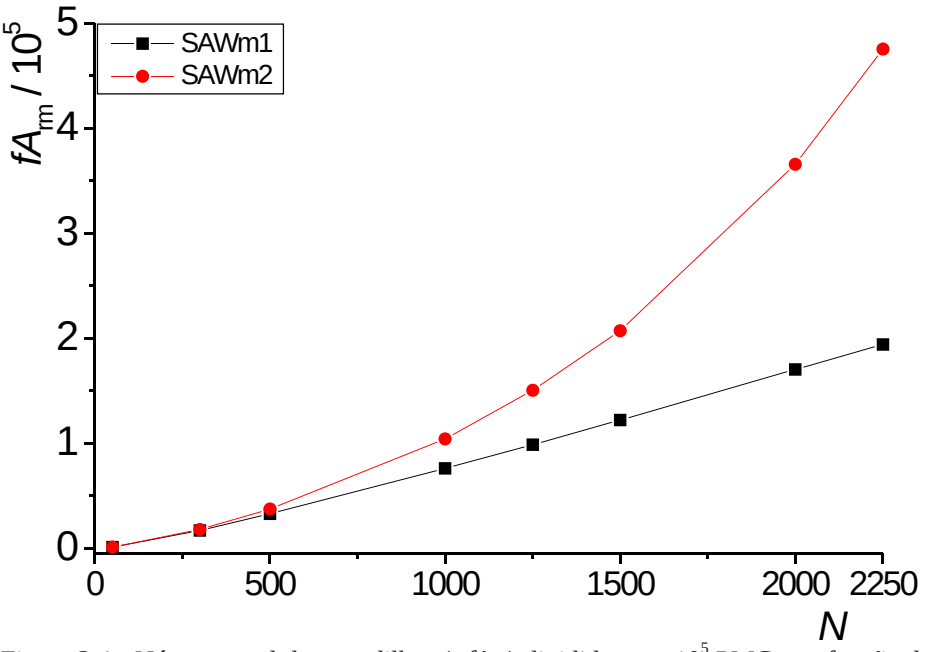


Figura 3.4 - Número total de armadilhas (fA_{rm}) divididos por 10^5 PMC, em função de oito comprimentos *N* de cadeias nascentes para os métodos SAWm1 e SAWm2. A variação de fA_{rm} em função de *N* segue comportamentos lineares e quadráticos para os métodos SAWm1 e SAWm2, respectivamente. Foram inseridos valores de fA_{rm} para *N* de 1250 e 2000, apenas para dar maior precisão e confirmar aos valores de fA_{rm} obtidos para os seis comprimentos de N, já utilizados até aqui.

Somente em cadeias longas (N $\gtrsim$ 1000) há separação significativa entre fA_{rm} para SAWm1 e SAWm2, revelando dois comportamentos distintos, sendo um linear e o outro polinomial de ordem 2, respectivamente. Note ainda que em N $\gtrsim$ 1000, principalmente para SAWm2, notamos um aumento muito grande de fA_{rm} em relação 10^5 PMC, tal que em *N* de 2250 são geradas cerca de duas e cinco vezes mais armadilhas do que o total de PMC inicialmente previsto para a simulação em SAWm1 e SAWm2, respectivamente.

As armadilhas ocorridas durante a simulação dos dois métodos empregados aqui, medidas pela frequências, distribuições e total de armadilhas (Figuras 3.2 a 3.4), juntamente com as armadilhas máximas (Tabela 3.1), em relação à diversos comprimentos *N* de cadeias nascentes, revelam comportamentos distintos destes dois métodos.

A seguir passamos a analisar as implicações dos resultados encontrados para as armadilhas,

quanto a compactação e o tempo de simulação.

3.3.2 Influência das armadilhas topológicas nas compactações das cadeias

As duas diferentes abordagens dadas para as armadilhas topológicas pode implicar em diferentes rotas para o enovelamento das cadeias, o que pode refletir em diferentes compactações das cadeias nascentes de comprimento *N*. Para verificarmos tais implicações fazemos uso do raio de giração e do expoente de escalonamento.

3.3.2.1 Variações dos raios de giração médios, nos métodos SAWm1 e SAWm2

O raio de giração médio, R_G, em 10^5 PMC, nos métodos SAWm1 e SAWm2, são analisados de duas maneiras independentes (Figura 3.5), pois assim é possível além de comparar os resultados fornecidos pelos dois métodos, também averiguar se estes são independentes da maneira como são medidos.

Numa primeira medida observamos a compactação das cadeias nascentes, após concluído o crescimento, ao final de cada PMC, em oito comprimentos N específicos, 50, 300, 500, 1000, 1250, 1500, 2000, 2250 resíduos. Na outra medida a compactação das cadeias são medidas de forma estratificada somente para *N* de 2250, ao final de cada PMC, em segmentos de cadeia acumulados de 100 resíduos, i.e., de 1 para 100, para 200, e assim sucessivamente até o segmento de 1 à 2200 resíduos, totalizando 22 pontos.

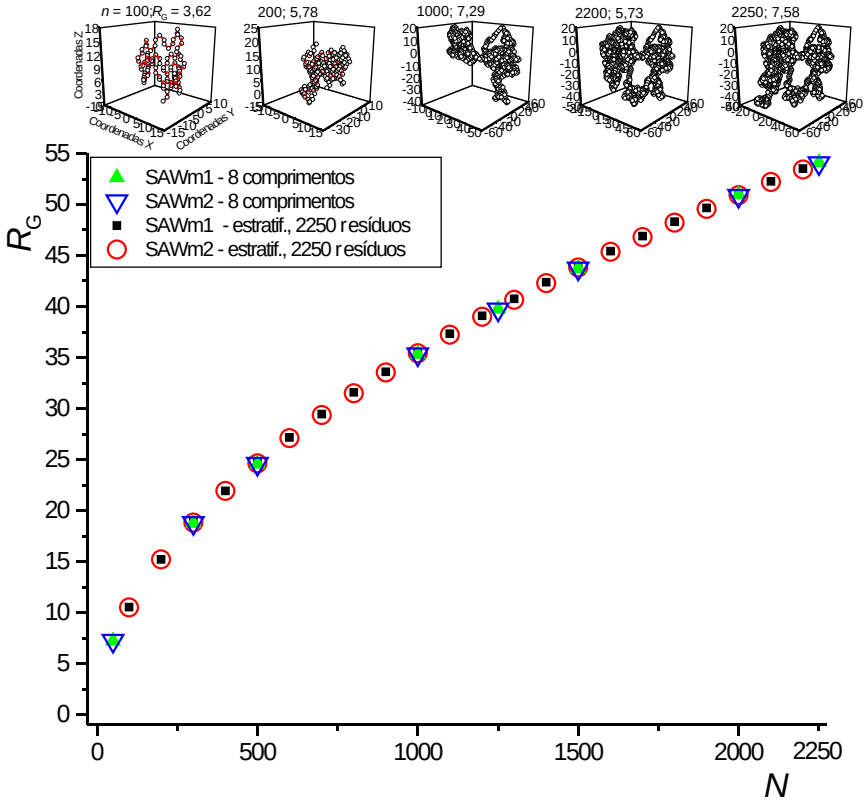


Figura 3.5 - Gráfico com variação do raio de giração médio, R_G, em 10^5 PMC, em função do comprimento N das cadeias nascentes para os métodos SAWm1 e SAWm2, segundo duas medidas diferentes e independentes; acima do gráfico são mostrados um exemplo de configurações estratificadas (em diferentes escalas) tomadas de SAWm1 no primeiro PMC (m=1), e acima de cada configuração são fornecidos seus valores de comprimento n e raio de giração. Inicialmente o R_G é calculado em oito comprimentos N de cadeias entre 50 e 2250 resíduos, e em seguida para N igual a 2250, em intervalos acumulados de 100 resíduos. O valor de R_G nas quatro distribuições, ou seja, para dois métodos e duas medidas, em função de N, são muito semelhantes.

Observando a variação de R_G em função do comprimento N para SAWm1 (símbolos (triângulos e quadrados) sólidos) e SAWm2 (triângulos e círculos vazados), nas duas maneiras independentes de medida é possível verificar a equivalência entre os métodos, com símbolos sólidos inscritos nos vazados, ou seja, praticamente com os mesmos R_G.

Apesar da quantidade e ocorrência das armadilhas topológicas (Figuras 3.2 a 3.4 e Tabela 3.1) evidenciarem diferentes dificuldades para crescer cadeias nos métodos SAWm1 e SAWm2, em termos configuracionais, pelo menos em relação a compactação, não é possível identificar quaisquer diferenças significativas nas conformações geradas pelos dois métodos.

Sabemos que a variação de R_G em função do comprimento N, segue uma lei de potência (equação 3.1) e se tomarmos os logaritmos neperianos destas duas variáveis (equação 3.4), temos

$$\ln(R_G) = \ln(N) + \ln(k) \qquad 3.4$$

A partir da variação de R_G em função de N é possível determinar o expoente de escalonamento (equação 3.4) e então compará-lo com outros tipos de cadeias.

3.3.2.2 Determinação do expoente de escalonamento para SAWm1 e SAWm2

Podemos determinar o expoente de escalonamento acompanhando a variação de ln(R_G) em função de ln (N) para o método SAWm1, em oito comprimentos de cadeias nascentes (Figura 3.6), cujos pontos são bem ajustados por uma linha reta.

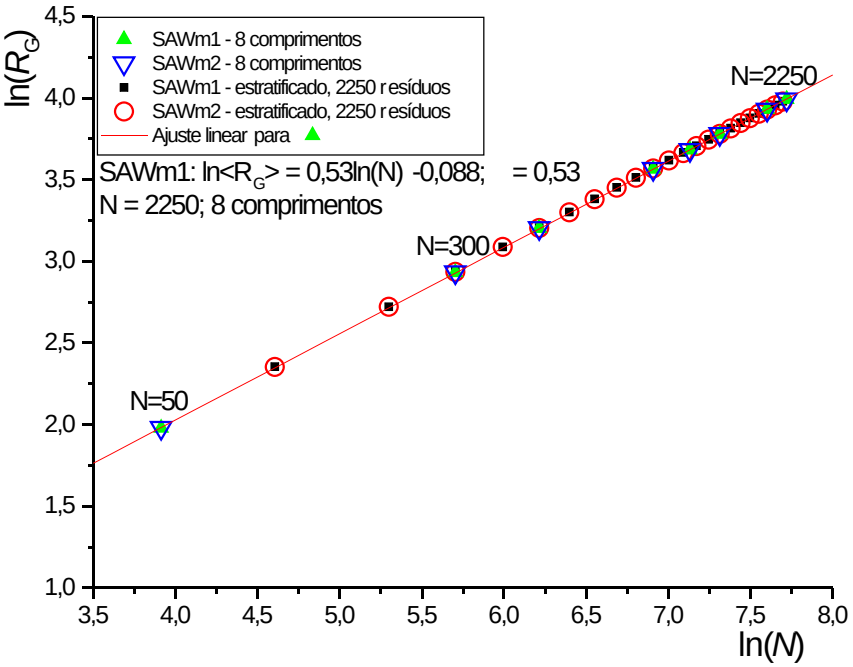


Figura 3.6 - Logaritmo neperiano do raio de giração médio, ln(R_G), em função de ln(N), das cadeias nascentes para os métodos SAWm1 e SAWm2,em duas medidas distintas, são idênticos; e com ajuste linear para SAWm1 em 8 comprimentos N. A reta contínua representa o melhor ajuste linear, dado por ln (R_G) = ln(N) + ln (k), com = 0,53 e erro estatístico de 6,72E-4; ln (k) =-0,088, e coeficiente de correlação linear de 0,99.

Para nosso caso, com cadeias nascentes representandas por polímeros desnaturados (homopolímeros), nós encontramos uma lei de escalonamento (equação 3.1) de $R_G \sim N^{0,53}$ que é independente de qual das quatro medidas (quatro símbolos na Figuura 3.6) tomamos para obter o valor de e além disso parece bem situada em relação a outros diferentes tipos de cadeias (ideal, compacta e inchada) e em diferentes condições do solvente ao redor da cadeia [G. A. Arteca, **1994**; G. A. Arteca, **1995**; G. A. Arteca, **1996**; J. Liang & K. A. Dill, **2001**].

O valor que encontramos fica entre de proteínas compactas gerais, em condições fisiológicas,

e de polímeros lineares em soluções diluídas de bom solvente, com valores de 0,4 e 0,588, respectivamente. O valor do expoente que obtivemos é maior que o de proteínas e cadeia ideal em mau solvente, já que temos apenas força repulsiva estérica, e portanto faltando as forças atrativas que tornam a proteína mais compacta, e também não temos as sobreposições entre resíduos da cadeia ideal (com = 0,5), daí nosso maior valor de e consequentemente de R_G.

Não encontrada diferenças significativas entre as compactações das conformações das cadeias nascentes (Figuras 3.5 e 3.6) passamos a analisar a eficiência dos métodos SAWm1 e SAWm2 atráves da variação do tempo real de CPU.

3.3.3 Tempo de processamento para SAWm1 e SAWm2

As diferenças observadas nos números de armadilhas topológicas para cadeias longas, entre os métodos de SAWm1 e SAWm2 pode ou não ter consequência direta no tempo de processamento, aqui representado pelo tempo real de CPU, em horas (Figura 3.7), para enovelar 10^5 cadeias nascentes em oito comprimentos.

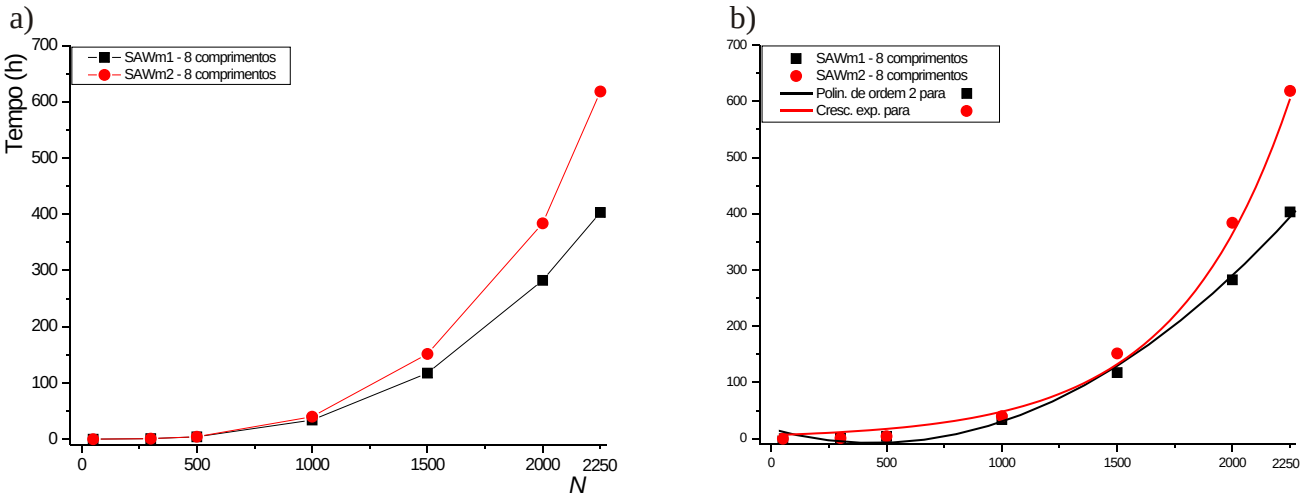


Figura 3.7 - Variação do tempo de processamento, em horas, em função de oito comprimentos de cadeia N paraSAWm1 e SAWm2, representado(a) por pontos conectados; e (b) com seus respectivos ajustes polinomial e exponencial.Aqui o tempo de simulção é medido pelo tempo real de CPU obtido num micro pentium 4 IBM, monocore, ano 2002, com sistema operacioanal Red Hat. Todas as simulações foram executadas nas mesmas condições, ou seja, no mesmo micro, em dedicação exclusiva, com execução de um programa por vez.
b) Os tempos de CPU, para SAWm1 e SAWm2, são dados por um polinômio de ordem 2 e um crescimento exponencial, respectivamente.
Para Sawm1: Tempo $=A + B_1N + B_2N^2$, com A, B_1 e B_2, igual a 16,89; -0,11; 1,23E-4, respectivamente.
Para SAWm2: Tempo = $Tempo_0 + Ae^{(N/b)}$, com $Tempo_0$=0; A=6,33; b=493,53

Observando a variação do tempo de CPU para SAWm1 e SAWm2, sob condições controladas (i.e., na mesma máquina e apenas um programa executado por vez) é possível verificar que em SAWm1 o tempo de simulação varia com *N* segundo um polinômio de ordem 2, ou seja, muito mais lentamente do que a exponencial crescente de SAWm2. A medida que as cadeias se tornam mais longas ($N \geqslant 1500$) fica nítida a diferença entre os dois métodos, e.g., em 2250 resíduos temos para SAWm1 um tempo de ~400 h (ou ~17 dias), enquanto para SAWm2 ~ 600 h (~26 dias), i.e., um aumento de 65% no tempo de simulação de SAWm2 em relação a SAWm1.

Note que as duas diferentes abordagens das armadilhas topológicas por SAWm1 e SAWm2 trazem significativas diferenças nos tempos de simulação, principalmente em cadeias longas (Figura 3.7b), já que o tempo da primeira cresce por um polinônio de ordem 2, enquanto o segundo é por uma exponencial crescente. Tais diferenças de tempo podem ser cruciais para a constituição de proteínas reais, com alguns milhares à dezenas de milhares de resíduos.

3. 4 Conclusões

As abordagens diferentes, nos métodos SAWm1 e SAWm2, para retificar configurações proibidas, as denominadas armadilhas topológicas, mostram comportamento dependente do comprimento *N* da cadeia nascente, e com diferentes implicações para a compactação e o tempo de enovelamento.

As armadilhas comportam-se diferentemente, em SAWm1 e SAWm2, principalmente em cadeias longas ($N \geqslant 1000$ resíduos), conforme medido por quatro grandezas diferentes, que mostram diferentes aspecto do armadilhamento topológico e cujas principais diferenças destacamos a seguir.

A frequência das armadilhas, fA_{rm} em função do comprimento variável *n*, para SAWm2 tem um comportamento tipo log Normal (para $N \geqslant 1000$), enquanto que em SAWm1 é sempre logarítmico, com o pico da distribuição de SAWm2 cerca de 4 vezes superior ao de SAWm1, ou seja, com muito mais armadilhas no primeiro do que no segundo método. A distribuição das armadilhas, $\%fNA_{rm}$, em função do número de armadilhas NA_{rm} (=0, 1, 2, 3 ...) mostram distribuições muito diferente nos dois

métodos, pois enquanto em SAWm1 há o predomínio de NA_{rm} pequenos (deslocados para a esquerda) em SAWm2 tais valores são deslocados de forma decrescente gradativa para a direita, confirmando que SAWm1 corrige mais eficientemente suas armadilhas.

O número máximo de armadilhas máx{NA_{rm}}, com %fNA_{rm} não nula, revela o quanto um método pode ser maximamente custoso do que o outro para a conclusão do processo de crescimento da cadeia, com diferença desfavorável a SAWm2 ao redor de duas à cinco vezes mais em N de 500 e 1000 à 2250, respectivamente. Por fim para medir o quanto de armadilhas foram encontradas durante a simulação utilizamos o número total de armadilhas (fA_{rm}) em função do comprimento N, que em SAWm1 segue um comportamento linear, enquanto em SAWm2 é quadrático.

Apesar de toda a diferença quanto aos comportamentos das armadilhas em SAWm1 e SAWm2, elas influenciam diferentemente a compactação e o tempo de enovelamento. Quanto a compactação, medida pelo raio de giração médio, em duas formas independentes de medida, não é possível diferenciar conformacionalmente os dois métodos, pois seus resultados são bastante similares. A qualidade dos resultados obtidos a partir do raio de giração foi atestadas pelo expoente de escalonamento com resultado compatível ao esperado, ou seja, superior ao de proteínas compactas e cadeia ideal, e inferior ao de polímeros lineares em bom solvente.

As diferenças observadas nos comportamentos das armadilhas em SAWm1 e SAWm2, não identificadas na compactação, ressurgem no tempo de processamento, medido pelo tempo real de CPU, em função de N, dado por comportamentos quadrático e exponencial, respectivamente, ou seja, novamente desfavorável a SAWm2, tal que em 2250 resíduos a diferença de tempo de CPU, já chega a mais de uma semana.

Embora o tempo de simulação em SAWm1, com interação de esfera rígida, ainda seja muito elevado para cadeias longas, tal tempo deve mudar pouco quando realizadas implementações adicionais, como a inserção de outras interações, pois atualmente já calculamos todas as distâncias intra-cadeia, o que costuma ser um dos ingredientes mais onerosos no tempo de CPU. Além disso, agora também conhecemos alguns comportamentos e limites para os dois métodos utilizados para o

crescimento de cadeia, o que facilita o entendimento de processos e mecanismos para o crescimento e enovelamento de proteínas desnaturadas.

4 Referências

- A. Caliri and M.A. A. da Silva, J. Chem. Phys. 106 (18), 7856-7861 (1997).

- B. Nölting in: Protein Folding Kinetics (Biophysical Methods); (Springer-Verlag Berlin, 2a edition; 17-25 (2006).

- C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr., PNAS 47(9), 1309-1314 (1961).

- C. K. Matheus, K. E. van Holde, K. G. Ahern; Biochemistry (3ª edition); 126-160; Addison Wesley Longaman (San Francisco) (2000).

- C. Lartigue, S. Vashee, M. A. Algire, R.-Y. Chuang, G. A. Benders, L. Ma, V. N. Noskov, E. A. Denisova, D. G. Gibson, N. Assad-Garcia, N. Alperovich, D. W. Thomas, C. Merryman, C. A. Hutchison III, Hamilton O. Smith, J. C. Venter, J. I. Glass. Science 325, 1693-1996 (2009).

- C. M. Guttman, Macromol. 19, 833-837 (1986).

- C. R. Cantor & P. R. Schimmel in*: Biophysical Chemistry; Part III The Behavior of Biological Macromolecules, W. H. Freeman and Company,* New York, *Chapter 18 (Configurational Statistics of Polymer Chai*ns*)*, 979-1018 (1980).

- C. Levinthal, J. Chim. Phys. 65 (1), 44-45 (1968).

- D. Ben-Avraham, S. Havlin, Diffusion and reactions in fractals and disordered systems; Cambridge University press (Cambridge), 141-156 (2000).

- D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z.-Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison III, H. O. Smith, J. C. Venter. Science 328, 1-7 (2010).

- D. L. Nelson, M.M.Cox; (Lehninger) Principles of Biochemistry (4ª edition); 1034-1080; W.H. Freeman and Company (New York) (2005).

- E. E. Meyer, K. J. Rosenberg, J. Israelachvili, Proc. Nat. Acad. Sc. USA 103, 15739-15746, 2006.

- G. A. Arteca, Phys Rev E, 49, 2417-2428 (1994).

- G. A. Arteca, Phys Rev E, 51, 2600-2610 (1995).

- G. A. Arteca, Phys Rev E, 54, 3044-3047 (1996).

- G. E. Schulz & R. H. Schirmer in: Principles of Protein Structure, ed. C. R. Cantor; (Springer-Verlag, New York); 27-45 (Chapter 3) (1990).

- G. Faure, A. Bornot, A. G. de Brevern. Biochimie 90, 626-639 (2008).

- G. Zifferer, Macromolecules 23, 3166-3172 (1990).

- G. I. Makhatadze and P. L. Privalov. Adv. Prot. Chem. 47, 308-425 (1995).

- H. C. Hamaker. Phys. IV 10, 1058-1072 (1937).

- H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov and P. E.

- Bourne. Nucl. Acids Res. 28 (1), 235-242 (2000) .

- H.-M. Lu and J. Liang. Proteins 70, 442-449 (2008).

- H. S. Chan, K. A. Dill. Phys. Today 46 (2), 24-32 (1993).

- H. Taketomi, Y. Ueda and N. Go. Int. J. Pep. Prot. Res. 7, 445-459 (1975).

- J. B. O. Mitchell, J. M. Thornton, J. Singh and S. L. Price. J. Mol. Biol. 226 (1), 251-262 (1992).

- J. C. Nelson, J. G. Saven, J. S. Moore, P. G. Wolynes. Science 277, 1793-1796 (1997).

- J. Liang and K. A. Dill, Biophys J, 81, 751-766 (2001).

- J. R. Banavar, T. X. Hoang, A. Maritan, J. Chem. Phys. 122, 234910-1-4 (2005).

- J. R. Banavar, A. Maritan, C. Micheletti, and A. Trovato, Proteins 47:315–322 (2002).

- L. Eller. O Uso do Método de Crescimento de Cadeias no Estudo do Enovelamento de Modelos de Proteínas. Dissertação (Mestrado) IBILCE Universidade Estadual Paulista "Júlio de Mesquita Filho" UNESP (2003).

- L. Hong and J. Lei, J. Pol. Sc.: Part B: Pol. Phy. 47, 207-214 (2009).

- L. F. O. Rocha, I. R. Silva, A. Caliri. Phys. A, 388 40974104 (2009).

- L. F. O. Rocha. Mecanismo de Folding de Macromoléculas: Estudo sobre o papel das especificidades topológicas e químicas das interações intramoleculares em cadeias poliméricas lineares. Tese (Doutorado) IBILCE Universidade Estadual Paulista "Júlio de Mesquita Filho" UNESP (2005).

- M. A. Moret, M. C. Santana, G. F. Zebende, and P. G. Pascutti, Phys. Rev. E 80, 041908-1-4 (2009).

- M. I. Aroyo, U. Müller and H. Wondratschek. *International Tables for Crystallography* (Springer) A1 (1.1): 2–5 (2006).

- M. Gleiser, *A Tear at the Edge of Creation: A Radical New Vision for Life in an Imperfect Universe; Free Press* (New York) , 1-273 (2010).

- M. Mann, O.N. Jensen. Nat. Biotechn. 21(3), 255-261(2003).

- M. N. Rosenbluth, A. W. Rosenbluth, J. Chem. Phys. 23 (2), 356 - 359 (1955).

- M. S. Weiss, M. Brandl, J. Sühnel, D. Pal and R. Hilgenfeld. Trends in Bioch. Scien. 26 (9), 521-523 (2001).

- N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. PNAS 98(18) 10037-10041(2001).

- N. Giovambattista, C. F. Lopez, P. J. Rossky, and P. G. Debenedetti, Proc Nat Acad Sci USA 105 (7), 2274-2279 (2008).

- N. J. A. Sloane. Scient. Amer. 250; 92-101 (1994).

- P. E. Leopold, M. Montal, J. N. Onuchic. Proc. Nat. Acad. Sci. USA 89, 8721-8725 (1992).

- R. B. Freedman in: *Protein Folding*, ed. T. E. Greighton; (W. H. Freeman and Company, New York); 455-540 (Chapter 10) (1992).

- R. V. Pappu, R. Srinivasan, and G. D. Rose, PNAS 97 (23): 12565–12570 (2000).

- S. Borman. Chem. & Engineer. News; 85 (6), 7-7 (2007).

- S. Mukherji and A. Van Oudenaarden.Nat. Rev. Genet. 10, 859-871 (2009).

- T. Greighton, Phil. Trans. R. Soc. Lond. B 348, 5-10 (1995).

- T. L. Blundell & K. Mizuguchi. Prog. in Biophys. &Mol. Biol. 73, 289-295 (2000).

- W. Rocchia, E. Alexov, and B. Honing. J. Phys. Chem. B 105, 6507-6514 (2001).

- Z. Dosztányl, C. Magyar, G. E. Tusnády, M. Cserzõ, A. Fiser and I. Simon. Nucleic Acids Res. 31 (13), 3359-3363 (2003).