

Luiz Faustino de Oliveira Rocha

Mecanismo de folding de macromoléculas

Estudo sobre o papel das especificidades topológicas e químicas
das interações intramoleculares em cadeias poliméricas lineares



1910001472



LUIZ FAUSTINO DE OLIVEIRA ROCHA

Mecanismo de *folding* de macromoléculas

**Estudo sobre o papel das especificidades topológicas e químicas
das interações intramoleculares em cadeias poliméricas lineares**

ORIENTADOR: PROF. DR. ANTONIO CALIRI

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

São José do Rio Preto – SP

2005



577.112

R672m

13/03/2006

Doação

Autor(a)

20801017

UNIVERSIDADE ESTADUAL PAULISTA



UNESP

Rocha, Luiz Faustino de Oliveira.

Mecanismo de folding de macromoléculas : estudo sobre o papel das especificidades topológicas e químicas das interações intramoleculares em cadeias poliméricas lineares / Luiz Faustino de Oliveira Rocha. - São José do Rio Preto : [s.n.] , 2005

118 f. : il. ; 30 cm.

Orientador: Antonio Caliri

Tese (doutorado) – Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas

1. Proteínas. 2. Proteínas - Simulação computacional. 3. Folding de proteínas. 4. Enovelamento. 5. Comparação (Proteínas) 6. Efeito hidrofóbico. I. Caliri, Antonio. II. Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas. III. Título..

CDU – 577.112



AGRADECIMENTOS

À Prof. Dr. Antonio Caliri
pela orientação, dedicação, estímulo, apoio e ensinamentos que vão muito além do conteúdo deste trabalho.

À Lexandra
pelo cuidado, respeito e apoio demonstrado em todos os momentos.

À minha família:

A família Caliri
pela amizade e apoio.

Às amigas e colegas:

Vanderlei, Eudélia, Eliza, Lilianete, Roosevelt, André, Lopes, Nêa, Ivan, Flávio, Sidney,
Leão, Sabrina e Luciana e a todos que contribuíram direta ou indiretamente para este
trabalho.

Às professoras do Departamento de Física do IBILCE/UNESP e Departamento de Física e
Química/FCRPA/USP
pelas ensinamentos e valiosas discussões.

Às funcionárias do Departamento de Física e Seção de Pós-graduação do IBILCE/UNESP
e do Departamento de Física e Química/FCRPA/USP e CCRPA/USP.

A FAPESP

pelo apoio financeiro sempre concedido diligentemente.

Dedico à Ariovaldo Taveira da Rocha, *in memoriam*



AGRADECIMENTOS

Ao Prof. Dr. Antonio Caliri
pela orientação, dedicação, estímulo, apoio e ensinamentos que vão muito além do conteúdo deste trabalho.

A Lessandra
pela amizade, respeito e apoio demonstrado em todos os momentos.

A minha família ...

A família Caliri
pela amizade e apoio.

Aos amigos e colegas
Vanderlei, Eulália, Elias, Eliamar, Roosevelt, André Lopes, Inês, Ivan, Flávio, Sidney, Ézio, Sabrina e Luciana e a todos que contribuíram direta ou indiretamente para este trabalho.

Aos professores do Departamento de Física do IBILCE/UNESP e Departamento de Física e Química/FCFRP/USP
pelos ensinamentos e valiosas discussões.

Aos funcionários do Departamento de Física e Seção de Pós-graduação do IBILCE/UNESP e do Departamento de Física e Química/FCFRP/USP e CIRP/USP

A FAPESP
pelo apoio financeiro sempre cumprido rigorosamente.



RESUMO

Proteínas são biomoléculas poliméricas lineares. Portanto, uma proteína globular possui, em princípio, um número astronômico de configurações possíveis. Contudo, para que seja funcional, ela precisa encontrar uma particular conformação espacial que é única e estável --sua estrutura nativa, e o faz por meio de um processo que custa um tempo relativamente pequeno. O tempo característico de *folding* varia, dependendo da proteína, de micro segundos (10^{-6} s) até horas (10^3 s). Este processo é comumente denominado *folding* protéico.

Neste trabalho estudamos o processo de *folding* de proteína por meio de um modelo simplificado do sistema composto proteína-solvente. O sistema é representado por uma cadeia linear composta por 35 unidades poliméricas envolvida por mais de doze mil moléculas de solvente. Todas as partículas, tanto do solvente como as unidades da cadeia, têm o mesmo tamanho. A evolução configuracional do sistema é determinada por meio de simulação Monte Carlo.

Os ingredientes básicos empregados no presente modelo para o estudo do mecanismo do *folding* são: (i)- especificidade estérica que no presente estágio do modelo é reduzida à restrição de esferas rígidas (impedimento de sobreposição entre partículas), e (ii)- força entrópica (efeito hidrofóbico) determinada pelos atributos hidrofóbicos (H) e polares (P) da sequência de aminoácidos da proteína alvo; esta força exerce ação direta entre monômeros da proteína e as moléculas do solvente ao seu redor. Entre as proteínas analisadas neste trabalho, num total de 27, há proteínas de mamíferos: rato (três), porco (uma), ser humano (três); de escorpiões, aranhas, tarântula e anêmona (onze); gafanhoto (duas); arroz (uma); folha de vegetal (uma); fermento (uma); do vírus HIV (duas); proteína desenhada (uma); proteína com sequência sintética não biológica (uma).

Neste trabalho, comparar adequadamente estruturas espaciais é de fundamental importância; em particular comparar estruturas moleculares de resolução espacial muito diferente. Assim, foi idealizada e aplicada uma ferramenta denominada aqui por *função discrepância configuracional*, para se estimar semelhanças e diferenças estruturais entre duas configurações quaisquer.

Somente com o emprego do potencial de esferas rígidas para o sistema cadeia-solvente, mostramos que o acesso à estrutura nativa de qualquer uma das estruturas consideradas pode ser feito num tempo relativamente pequeno e independente dos padrões estruturais (globular ou hélice) da proteína. Este fato sugere que os relativamente poucos *folds* apresentados pelas estruturas nativas das proteínas naturais conhecidas, são de fato padrões estruturais que correspondem às conformações mais prováveis de uma cadeia livre de restrições energéticas (*random coil*) embebida no solvente.

Na última etapa deste trabalho mostramos que o potencial hidrofóbico é eficaz para direcionar a cadeia para configurações vizinhas da estrutura nativa da proteína da qual extraímos a correta sequência HP utilizada no modelo. Além disso, verificamos que este potencial consegue discernir esta proteína das demais proteínas globulares. Contudo tal potencial não fornece à cadeia protéica a necessária estabilidade energética.

Os resultados aqui apresentados sugerem que o mecanismo para acesso a estrutura nativa e os condicionantes da estabilidade da cadeia nesta estrutura constituem problemas independentes. Somente na estrutura nativa todos os fatores são alinhados no sentido de conferir estabilidade ao glóbulo protéico. A busca pelo estado nativo é dirigida essencialmente pelo potencial hidrofóbico que mantém o glóbulo significativamente maleável. Assim, longe da estrutura nativa a cadeia peptídica seria mais livre energeticamente, podendo visitar configurações que lembram as configurações *random coil* de uma cadeia imersa no solvente. Mas uma vez atingida a estrutura nativa, e somente nela, todos os fatores se alinham no sentido de conferir estabilidade ao glóbulo.



ABSTRACT

Proteins are linear polymeric biomolecules. Therefore, a globular protein has, in principle, an astronomic number of possible configurations. However, in order to function properly it must find a unique specific stable conformation --its native structure, and does it through a process that takes a relatively small amount of time. The characteristic folding time changes, depending on the particular protein, from microseconds (10^{-6} s) to hours (10^3 s). This process is known as protein folding.

In this work the protein folding process is studied through a composed protein-solvent simplified model. The system is represented by a 35-mer linear chain immersed in more than twelve thousand solvent molecules. All particles, solvent molecules and chain units, are of the same size. The configurational evolution is determined by the Metropolis Monte Carlo simulation method.

The basic model ingredients employed in the present model are: (i)- steric specificity, which for the present model is reduced only to hard-core constraints (no particle overlap); and (ii)- entropic force (hydrophobic effect), which is determined by the aminoacid hydrophobic (H) and polar (P) attributes of the target protein; this force acts directly among the protein units and the surrounding solvent molecules. Among the twenty seven proteins analyzed in this work, three are from rats; one from pig; three from humans; eleven from scorpions, spider, tarantula and anemone; two from *Schistocerca gregaria*, one from rice; one from vegetable leaves; one from ferment; two from HIV virus; one designed protein; and, finally, one defined by a non biological synthetic sequence.

For the present work, to be able to do proper comparison between spatial structures is fundamental; particularly for molecular structures described by distinct spatial resolution. Therefore, it was conceived and applied a special tool, denominated here as configurational discrepancy function, in order to evaluate similarities and differences between any two configurations.

Using only hard sphere potential we were able to show that the access to native structure of all protein considered here was realized in a relatively small time window, independently of the structural patterns of the proteins (α helices, β sheets, etc.). This fact suggests that the relatively small set of folds presented by the structures of the known globular proteins are, indeed, structural patterns corresponding to those more probable conformations of an energetically free chain (random coil) immersed into the solvent.

It was also shown that the hydrophobic potential is efficient to direct the chain to the configurational neighborhood of the native structure of the protein from which the correct HP attribute was used in the chain model. Additionally it was verified that this potential is able to discern each particular globular protein from the others one. However, this potential can not provide to the chain the necessary protein stability.

Finally, the results as a whole suggest that the mechanism to access the native structure and the determinants of the chain stability in the native structure constitute independent problems. Only in the native structure all the factors are aligned in order to provide stability to the protein globule. The search for the native structure is driven essentially by the hydrophobic potential which keeps the chain malleable. Therefore, far away from the native structure, the chain is energetically free to visit configurations that remind random coil configurations of a chain immersed in the solvent. But, once in the native structure, and only in it, all the factors align up in the sense to provide stability to the globule.



ÍNDICE

	páginas
RESUMO	iii
ABSTRACT	iv
CAPÍTULO 1 - A Biomolécula Proteína e seu processo de <i>Folding</i>	1
1.1 Hipóteses básicas do processo de <i>folding</i>	2
1.2 Dos átomos às biomoléculas	3
1.3 As Biomoléculas não protéicas	6
1.3.1 Carboidratos	6
1.3.2 Ácidos nucleicos	7
1.3.3 Lipídeos	7
1.4 Aminoácidos: Os monômeros das proteínas	8
1.5 Proteínas	10
1.6 Níveis estruturais das proteínas	12
1.6.1 Estrutura primária	12
1.6.2 Estrutura secundária, supersecundária e domínios	13
1.6.3 Estrutura terciária	15
1.6.4 Estrutura quaternária	16
1.7 <i>Folding</i> de proteínas: A seqüência determina a estrutura nativa	16
1.7.1 Evidências da relação entre estruturas e funções nas proteínas	17
1.7.2 Modelos para o processo de <i>Folding</i>	18
1.7.3 Modelo simplificado para o sistema proteína-solvente	18
1.8 Sumário do conteúdo dos próximos Capítulos	20
1.9 Bibliografia	22
CAPÍTULO 2 - Comparações configuracionais: a função discrepância	24
2.1 Introdução: Comparação estrutural	24
2.2 A função discrepância: definição	26
2.3 Discrepância cruzadas entre estruturas idealizadas: semelhanças entre uma estrutura retilínea com estruturas retilíneas parcialmente dobradas e estruturas helicoidais	28
2.4 Comparações configuracionais: autodiscrepância e discrepância cruzada	33
2.4.1 Auto discrepâncias: comparações entre configurações nativas alternativas de uma mesma proteína	35
2.4.2 Discrepâncias cruzadas: comparações entre configurações nativas de diferentes proteínas	39
2.5 A função discrepância configuracional: síntese de propriedades	42
2.6 Comentários	45
2.7 Bibliografia	46
CAPÍTULO 3 - Parametrização da cadeia protéica do modelo e o fator aleatoriedade no processo de <i>folding</i> de proteínas	48



3.1	Sistema proteína-solvente: Especificação do modelo	48
3.1.1	Comportamento configuracional global: O fator temperatura	51
3.1.2	Parametrização da cadeia: A separação relativa entre monômeros consecutivos	52
3.2	Configurações <i>random coil</i> de uma cadeia imersa no solvente	54
3.2.1	Caracterização quanto à compactação e contatos intra cadeia	55
3.3	Semelhanças configuracionais: As Discrepâncias cruzadas mínimas globais ($\Delta D_{\min G}$) e médias ($\langle \Delta D_{\min G} \rangle$)	61
3.3.1	Semelhanças entre configurações alvo do modelo e configurações <i>random coil</i>	61
3.3.2	Comparações entre configurações protéicas e <i>random coil</i>	64
3.4	Resultados adicionais	66
3.4.1	Os tamanhos dos resíduos influenciando nos valores de $\langle \Delta D_{\min G} \rangle$, obtidos na simulação	66
3.4.2	Algumas implicações decorrentes de valores relativamente estáveis de $\langle \Delta D_{\min G} \rangle$	69
3.4.3	Justificativas para $\langle \Delta D_{\min G} \rangle$ assumir valor diferenciado, de acordo com a compactação da cadeia protéica	73
3.4.4	Dependência da compactação de uma proteína com o tamanho de seus resíduos	73
3.4.5	Paradoxo de Levinthal versus Semelhanças entre estruturas protéicas e configurações <i>random coil</i>	76
3.5	Comentários	77
3.6	Bibliografia	80

CAPÍTULO 4 - O Efeito hidrofóbico e o *Folding* de Proteínas 81

4.1	Forças que governam o processo de <i>folding</i> protéico	81
4.2	Efeito hidrofóbico	83
4.3	Termodinâmica do processo de <i>folding</i> de proteínas	84
4.3.1	ΔG^P : Contribuições dos grupos polares (P) para ΔG	85
4.3.2	$\Delta G^{\tilde{P}}$: Contribuições dos grupos apolares ($^{\tilde{P}}$) para ΔG	86
4.4	Modelo simplificado: Potencial Hidrofóbico e restrições estéricas	87
4.5	Propriedades configuracionais do modelo simplificado	90
4.6	Comparação entre as estruturas das proteínas alvos e conformações da cadeia previstas pelo modelo	91
4.6.1	Mapas de Distância de configurações protéicas e do modelo simulado	91
4.6.2	Mapas de Distância das proteínas e Mapas de Frequências de Contatos das configurações do modelo simulado	93
4.6.3	Discrepâncias cruzadas: comparações entre conformações protéicas e configurações do modelo simulado	94
4.7	Comentários	97
4.8	Bibliografia	99

CAPÍTULO 5 - Conclusões 101

APÊNDICE - Características gerais das proteínas analisadas neste trabalho 104



LISTA DE SÍMBOLOS E ABREVIATURAS

α, β	letras Gregas alfa e beta, utilizadas nas estrutura secundárias α -hélice e folhas β
ε	letra Grega epsilon; máxima separação entre as superfícies de dois resíduos conectados na cadeia
e.g.	do Latin <i>exempli gratia</i> (ou, por exemplo)
et al.	abreviação do Latin de <i>et alii</i> (masculino), <i>et aliae</i> (feminino), and <i>et alia</i> (neutro) indicando genericamente "e outras pessoas"
δ, Δ	letra Grega delta minúscula e maiúscula, representando variação, respectivamente em δl (elongações de hélices quadradas), $\delta_{\max}$ (amplitude máxima de movimentos) e ΔD (função discrepância configuracional).
$\Delta G, \Delta H$ e ΔS	representando, respectivamente mudanças na energia livre, entalpia e entropia
Δn	variação do número de partículas constituintes
D	Diâmetro das esferas representativas do solvente e dos aminoácidos ao longo da cadeia, cujo valor é $D = 3^{1/2}$
d_c	Distância predefinida de corte
D_{pp}	Distância ponta a ponta
e_c	Energia configuracional da cadeia
γ	letra Grega gamma; nível hidrofóbico
η	letra Grega eta; número de contatos topológico
k_B	Constante de Boltzmann
L	lado da caixa cúbica contendo o sistema cadeia-solvente
μ	letra Grega mu; para potencial químico
n_c	Número moléculas do solvente quimicamente ligada à cadeia
ϕ, ψ	letras Gregas, respectivamente phi e psi, ambas empregada para ângulos de torção
Φ	letra Grega, phi maiúsculo; energia livre
p_c	peso estatístico
P	Pressão
"PDB"	iniciais do Inglês <i>Protein Data Bank</i> (ou, Banco de dados de proteínas)
PMC	iniciais de Passo Monte Carlo, unidade de tempo
ρ	letra Grega rho; fração volumétrica de ocupação do sistema proteína-solvente
R_G	Raio de giração
RMSD	iniciais do Inglês <i>Root of Mean Square Deviation</i> (ou, raiz do desvio quadrático médio)
RMN	iniciais de Ressonância Magnética Nuclear
t, T	respectivamente tempo e temperatura absoluta
$T(A \rightarrow B)$	probabilidade de transição de uma configuração A para B
v_c	Volume da cadeia
v	letra Grega upsilon; em v_{ij} representa-se o número de resíduos ao longo da cadeia entre os resíduos i e j em contato topológico
χ	letra Grega chi; ordem de contato
Z	Função partição do sistema proteína-solvente
3D	tridimensional
∞	símbolo para "infinito".



CAPÍTULO 1

A Biomolécula Proteína e seu processo de *Folding*

Este trabalho se dedica exclusivamente ao estudo da biomolécula proteína, e por isso estas serão descritas neste Capítulo, desde seus constituintes mais básicos, átomos e aminoácidos, até as estruturas que as compõem. Em particular, será dada especial ênfase ao fato de que esta biomolécula, a partir unicamente da informação contida na sua sequência de aminoácidos alcança a sua conformação tridimensional precisa, através de um processo conhecido por *folding*¹ protéico. Tal processo é decisivo para o surgimento e manutenção da proteína, assim como para esta exercer adequadamente sua função biológica num determinado organismo.

Apesar de toda a complexidade dos diversos organismos vivos da natureza e de suas partes básicas, como as células, há características comuns a todos, como a utilização de energia em seus processos vitais e o emprego de biomoléculas similares, como as proteínas (Figura 1). Ademais, admite-se que de alguma forma todos os fenômenos envolvidos nos processos vitais não contrariam leis da física e química.

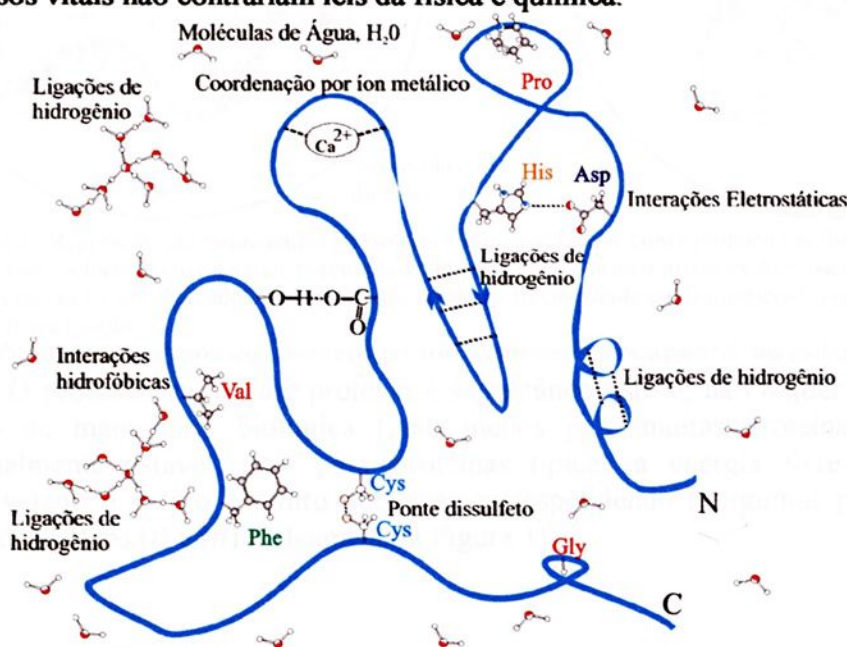


Figura 1 - As leis da Física representadas aqui pelas diversas forças, de natureza eletromagnética, estabilizam a forma tridimensional correta (ou forma nativa) das proteínas, sendo estas:

(i) Ligações de hidrogênio do esqueleto protéico, envolvendo os grupos CO e NH; nas cadeias laterais entre quaisquer doadores e aceptores de ligações de hidrogênio; (ii) Ligações de hidrogênio entre grupos polares e o solvente (água); (iii) Forças entrópicas envolvendo grupos apolares da proteína e água (Rocha, L. F. O. et al. (2004)); (iv) Forças eletrostáticas, entre grupos carregados da proteína; (v) Ligações químicas cruzadas: Pontes dissulfeto entre resíduos de Cisteínas, Cys; Coordenação por íon metálico, envolvendo ligações entre cadeias laterais e eventuais íon metálico; (vi) interação do tipo esfera rígida (representado pela interação repulsiva de van der Waals) presente por todo o sistema.

Obs.: Maiores informações serão fornecidas nas seções 1.4 e 1.5 adiante.

¹ O termo *Folding*, grafado em Inglês e em itálico, será preferido no contexto deste trabalho, pois ainda não há a uma tradução definitiva para o mesmo.

Para o seu adequado funcionamento, biomoléculas precisam adotar certas conformações específicas. Em muitas delas, o tempo de vida destas conformações deve durar somente uma fração ínfima de tempo, mas em outras, como as proteínas globulares, sua estrutura nativa pode perdurar por horas, ou dias, ou muito mais tempo, dependendo de sua função.

As proteínas são formadas por um repertório diversificado de aminoácidos, ricos em átomos de carbono, e com diferentes propriedades físicas e químicas. Os vinte aminoácidos padrões (ou naturais) apresentam diferentes tamanhos e formas, distintas polarizações elétricas e diferentes afinidades pelo solvente, água. Tais aminoácidos juntos e numa sequência específica (estrutura primária), contêm a informação básica necessária para conduzir a cadeia protéica à estrutura nativa e final, a partir de sua formação no ribossomo (Figura 2).

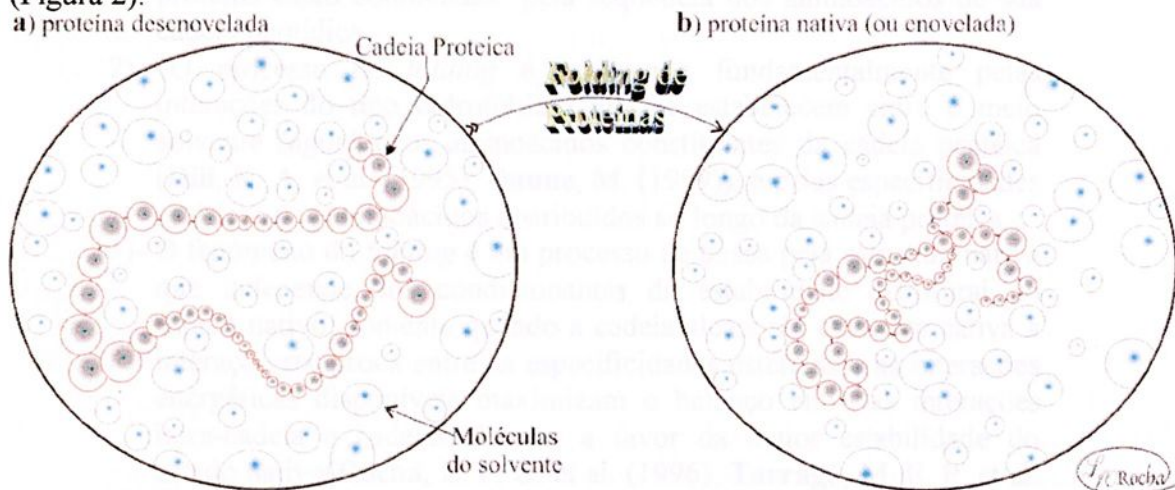


Figura 2 - Representação esquemática do sistema proteína-solvente, com a proteína nas formas (a) desenovelada e (b) nativa - A forma nativa é única, porém não rigidamente estática no transcorrer da existência da proteína, pois na forma nativa há um delicado equilíbrio entre rigidez e flexibilidade conformacional, para que a proteína possa exercer sua função.

O problema do *folding* proteico consiste em prever a estrutura nativa, a partir de sua sequência de aminoácidos.

O processo de *folding* proteico é espontâneo, isto é, não requer necessariamente o auxílio de maquinaria biológica (pelo menos para muitas proteínas pequenas), e é marginalmente estável, pois para proteínas típicas a energia livre entre os estados desenovelado e nativo é muito pequena, correspondendo a algumas poucas ligações de hidrogênio (itens (i) e (ii) na legenda da Figura 1).

1.1 Hipóteses básicas do processo de *folding*

De forma geral, o processo de *folding* que ocorre em um sistema proteína globular-solvente envolve múltiplos ingredientes. O fenômeno se processa sob parâmetros físico-químicos intensivos específicos, que definem a condição fisiológica; abrange diversos tipos de interações energéticas, envolve um rico repertório de especificidades estéricas e, *in vivo*, pode contar também com o subsídio de outras moléculas biológicas.

Mas há proteínas de todo tipo: diversos tamanhos, desde algumas poucas dezenas de aminoácidos até milhares deles, diversas topologias estruturais, e diferentes predomínios de

interações energéticas. Por exemplo, há proteínas que apresentam pontes dissulfeto e outras não; assim como proteínas em que não há pontes salinas. Inclusive, para sistemas cadeia-solvente especiais, tem sido demonstrado que mesmo as ligações de hidrogênio são dispensáveis para o processo de *folding* (Nelson, J. C. et al. (1997)).

Contudo, dois destes ingredientes são ubíquos: a interação hidrofóbica e as especificidades estéricas dos aminoácidos. São fatores comuns a todas as proteínas, pequenas ou grandes; globulares, fibrosas ou de membrana. Portanto é razoável supor que estes sejam os principais ingredientes que governam o processo do *folding* das proteínas.

Assim, neste trabalho, as seguintes premissas são admitidas:

- 1)- As instruções que finalmente determinam a estrutura nativa de uma proteína estão codificadas pela sequência dos aminoácidos de sua cadeia peptídica.
- 2)- O processo de *folding* é governado fundamentalmente pelas interações do tipo hidrofóbicas que se estabelecem entre o meio solvente (água) e os aminoácidos constituintes da cadeia protéica (Dill, K. A. et al. (1995); Daune, M. (1999)); e pelas especificidades estéricas dos aminoácidos distribuídos ao longo da cadeia protéica.
- 3)- O fenômeno do *folding* é um processo de busca pela estrutura nativa que independe dos condicionantes da estabilidade estrutural do estado nativo; somente quando a cadeia alcança a estrutura nativa a interação recíproca entre as especificidades estéricas e as interações energéticas disponíveis maximizam o balanço entre as interações intra-cadeia e cadeia-solvente, a favor da maior estabilidade do estado nativo (Rocha, L. F. O. et al. (1996); Tarragó, M. E. P. et al. (2003); Rocha, L. F. O. et al. (2004)).

Estas premissas constituem as hipóteses fundamentais deste trabalho, as quais deverão ser confirmadas pelos experimentos computacionais desenvolvidos neste trabalho.

1.2 Dos átomos às biomoléculas

Como a matéria viva adquiriu sua constituição atual é um dos grandes desafios para a imaginação humana. A teoria cosmológica mais aceita atualmente sobre a origem do universo, que teria sido originado ao redor de $1,5 \times 10^{10}$ anos atrás (segundo cientistas da Nasa $1,37 \times 10^{10}$ anos), é a do *Big Bang*² (Morris, R. (1999)); segundo tal teoria, o próprio espaço é criado por uma enorme explosão que deu origem ao universo (Figura 3a). A

² Há três principais provas para a teoria do *Big Bang*:

1ª) Universo em expansão, pela observação de galáxias distantes e seus aglomerados; 2ª) Radiação cósmica de fundo (microondas), constituída por ondas de rádio curtas que chegam a Terra com mesma intensidade proveniente de todas as direções, numa temperatura de 2,726 K, atualmente; 3ª) As quantidades de Deutério (o isótopo pesado do hidrogênio) e Hélio existentes hoje, se formaram dos prótons e nêutrons do Universo primitivo, e não nas estrelas, pois no Deutério as forças fracas entre seu próton e nêutron se romperiam nas altas temperaturas das estrelas, já o Hélio produzido, por reações nucleares nas estrelas, quando muito chegariam a 15% da quantidade existente hoje.



matéria (que inicialmente estava numa temperatura altíssima) lançada pelo *Big Bang* iniciou um processo de resfriamento, além disso, desde o início ela preenchia uniformemente todo o espaço.

Tal evolução do universo é bem compreendida a partir de 1 segundo, ou um pouco antes, da grande explosão (**Weinberg**, S. (1993)) até o período atual, sendo tal compreensão baseada em conhecimentos, como (i) os da Física Nuclear sobre as reações ocorridas no interior de estrelas; (ii) as observações astronômicas sobre a evolução de estrelas em diversos estágios de existência, desde as novas (centenas de milhares de anos), as de meia idade (como o Sol, com $0,50 \times 10^{10}$ anos), as antigas (com 2% a 3% de Hélio (He) a mais que o do Sol) e as já inexistentes e ainda contempladas atualmente apenas por seu calor residual.

Espalhados por este universo recém criado (Figura 3b) haviam fótons (luz), elétrons e prótons, partículas α (núcleos de He), poucos núcleos de dêuterons (núcleos de deutérios) e vestígios de outros núcleos (de outros tipos de He e Li (Lítio)) e talvez partículas como o neutrino (**Penrose**, R. (2001)).

Com a redução da temperatura do Universo (Figura 3c) os elétrons livres, que formavam uma densa neblina, juntaram-se as partículas nucleares (prótons e nêutrons) formando os primeiros átomos, os mais simples, como o Hidrogênio (H), Hélio e vestígio de Lítio. Estes átomos simples originaram os outros átomos mais pesados por: (a) reações termonucleares no interior das estrelas; (b) explosões de supernovas; (c) ação dos raios cósmicos fora das estrelas. Particularmente os isótopos dos átomos de importância biológica, no caso o Carbono (C), Nitrogênio (N), Oxigênio (O), Fósforo (P) e Enxofre (S), possuem núcleos estáveis.

Principalmente os prótons e elétrons se combinaram para formar o gás difuso (Figura 3d), que se distribuiu de alguma forma irregular para constituir as galáxias com suas estrelas. A existência de nuvens de gases (basicamente de H e He), ligeiramente mais densas que a média do Universo, propiciou a formação de galáxias, pois as forças gravitacionais faziam com que tais nuvens se contráíssem, se tornando ainda mais densas, por milhões de anos, até o ponto de ser contra balanceada pela força centrífuga da rotação da nuvem. Uma variação na densidade das nuvens de gás de uma parte em 10^5 seria suficiente para produzir as galáxias e seus aglomerados.

As nuvens de gases e poeira interestrelar maior que a média no interior das galáxias se contráíram formando as estrelas.

A datação radioativa de elementos instáveis indica que o planeta Terra surgiu ao redor de $0,45 \times 10^{10}$ anos atrás, provavelmente teve sua atmosfera primitiva constituída por compostos simples, como H_2O líquida e vapor, H_2 , CH_4 , NH_3 , H_2S , CO e CO_2 . Na década de 30, independentemente, A. I. Oparin e J. B. S. Haldane propuseram que os primeiros compostos orgânicos (ou compostos de Carbono), como as biomoléculas simples, se formaram a partir de reações entre as moléculas da atmosfera primitiva terrestre submetida a descargas elétricas atmosféricas ou pela radiação ultravioleta solar (Figura 3e). Na década de 50, experimentalmente, S. Miller e H. Urey (**Miller**, S. J. & Orgel, L. E. (1974)) mostraram que os compostos da atmosfera primitiva da Terra podiam reagir abioticamente (ou na ausência de vida) entre si originando inclusive componentes de proteínas, particularmente os aminoácidos Glicina e Alanina, respectivamente nas proporções de 2,1% e 1,7%, apenas inferiores ao ácido fórmico, 4%, e o ácido glicólico, 1,9%.

A pronta formação de aminoácidos em condições abióticas, como demonstrada no experimento de Miller-Urey, ao contrário dos nucleotídeos (formados com maiores

dificuldades) é uma das fortes evidências da importância das proteínas (no caso, proteinóides, com propriedades comparáveis de proteínas reais, e com agregados formados nos oceanos, ou em suas margens, na Terra primitiva) para o surgimento da vida (Cairns-Smith, A. G. (1985)). Tais agregados de proteinóides juntamente com outras biomoléculas formadas abioticamente formariam protocélulas, precursoras das células verdadeiras; modelos de protocélulas têm sido obtidos por vários pesquisadores (Campbell, M. K., Capítulo 1, pg. 37-45 (2000)). A mais antiga evidência de vida na Terra é de um microfóssil, com $0,35 \times 10^{10}$ anos, de uma célula de bactéria filamentosa.

Nossos resultados com um sistema cadeia-solvente, no qual somente o efeito de volume excluído é considerado, evidenciam que mesmo num intervalo de tempo relativamente pequeno, por busca aleatória, é possível acessar a estrutura nativa de todas as proteínas consideradas neste trabalho (Capítulo 3). Tais resultados podem contribuir para um melhor entendimento da teoria do surgimento da vida a partir de proteinóides e também com o processo de *folding* protéico, realizado no ambiente celular, merecendo tais suposições estudos adicionais de nossa parte.

Os eventos descritos até aqui, desde o surgimento do Universo até a formação das primeiras biomoléculas, são resumidamente descritos na Figura 3.

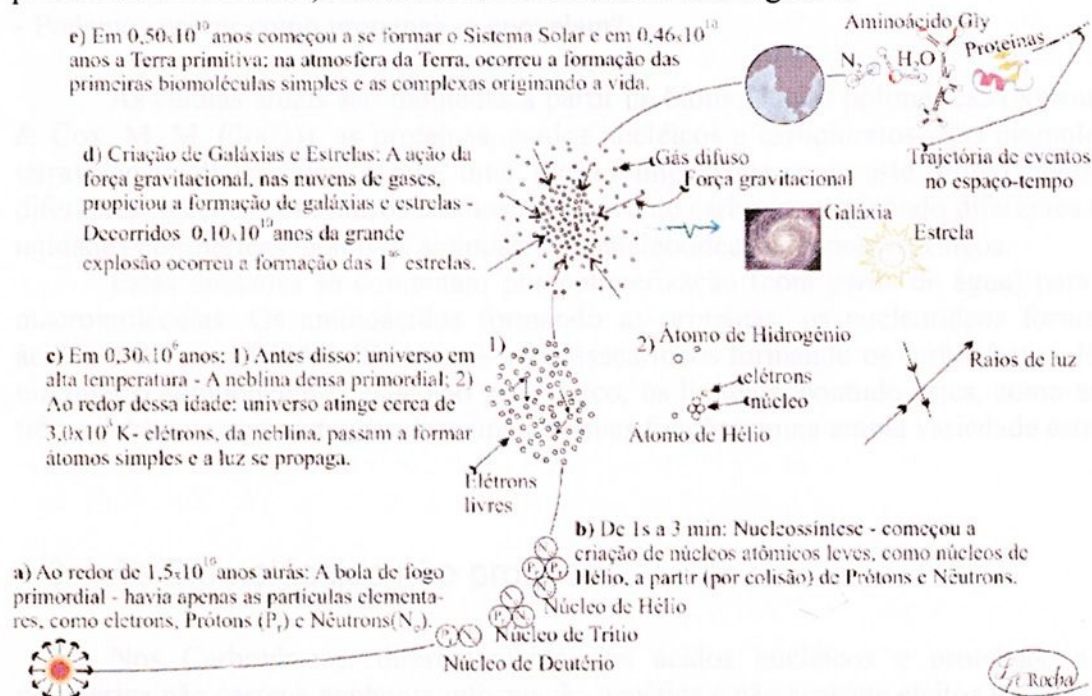


Figura 3 - Cronologia dos principais eventos, desde a criação do Universo, até a formação das biomoléculas, sendo estes:

- Até cerca de 1 segundo (1s) após a criação, as descrições sobre o Universo são puramente teóricas (Guth, A. (1998)), mas a partir de 1s, baseados em evidências (observações e/ou medidas), conhecemos o que aconteceu.
- O Universo de 1s a 3 minutos de idade: a temperatura do Universo com idade de 1s era quase 10^{10} K, mas esfriou rapidamente, até que com a idade de 3 minutos as reações nucleares não podiam mais ocorrer, a partir daí o Universo simplesmente esfriou e expandiu.
- 1) Antes de $0,30 \times 10^6$ anos, na densa neblina de elétrons livres os raios luminosos não se propagavam, por serem absorvidos ou espalhados pelos elétrons; 2) Após $0,30 \times 10^6$ anos o universo atinge cerca de $3,0 \times 10^4$ K os elétrons passam a formar átomos simples e a luz (radiação cósmica de fundo) pode então se propagar.
- Irregularidades na distribuição das nuvens de gás e com o tipo adequado de concentração de gravidade teve início a constituição das galáxias e estrelas.
- Na atmosfera primitiva da Terra, provavelmente a partir de precursores simples, como H_2O , H_2 , CO_2 e N começou a síntese das primeiras biomoléculas simples seguida pelas biomoléculas mais complexas.

Tópicos como o surgimento e evolução do Universo ocupam vários ramos científicos e constituem grandes mistérios a serem desvendados. Em geral são de natureza interdisciplinar e ainda sem respostas satisfatórias e portanto deverão ser abordados pelos cientistas do século 21 (**Science** (2005)); dentre estes, são destacadas as questões:

- De que o Universo é feito?
- Como e onde a vida na Terra surgiu?
- O nosso é o único Universo?
- O que causou a inflação do cosmos?
- Quando e como as primeiras estrelas e galáxias se formaram?
- De onde vêm os raios cósmicos de energia ultra-alta?
- Como os planetas se formam?

São questões gerais, mas há grandes desafios de toda ordem (**Science** (2005)), como de tópicos intimamente relacionados ao tema do atual trabalho, como:

- Qual é a estrutura da água?
- Qual é a origem da homociralidade na natureza?
- Quantas proteínas há nos seres humanos?
- Podemos prever como proteínas se enovelam?

As células atuais são montadas a partir de biomoléculas poliméricas (**Nelson, D. L. & Cox, M. M.** (2002)): as proteínas, ácidos nucleicos e carboidratos. Nas biomoléculas a tetravalência do carbono exerce uma função importante, pois este átomo pode formar diferentes ligações com outros átomos, inclusive de carbono, originando diferentes tipos de unidades poliméricas, como os aminoácidos, nucleotídeos e monossacarídeos.

Estas unidades se combinam por polimerização (com perda de água) para formar macromoléculas. Os aminoácidos formando as proteínas, os nucleotídeos formando os ácidos nucleicos (DNA e RNA) e os monossacarídeos formando os carboidratos. Há ainda um quarto componente celular não polimérico, os lipídeos, contudo estes, como as outras três, também se agregam para desempenhar suas funções numa ampla variedade estrutural.

1.3 As Biomoléculas não protéicas

Nos Carboidratos, diferentemente dos ácidos nucleicos e proteínas, a cadeia polimérica não carrega nenhuma informação genética e não acarreta efeitos importantes nas propriedades do polímero. Contudo a diversidade estrutural das unidades poliméricas nos carboidratos é fundamental para suas funções biológicas. Já os Lipídeos devido a sua constituição predominantemente hidrofóbica tendem a se agrupar formando membranas, por exemplo.

1.3.1 Carboidratos

Carboidratos ou sacarídeos (do Grego, *sakcharon*, açúcar) são as mais abundantes biomoléculas. Desempenham vários papéis: atuam como fonte de energia; são fundamentais para processos que ocorrem na superfície da célula, como nas interações célula-célula e no reconhecimento molecular nos sistemas de imunidade; são também



componentes estruturais em várias classes de organismos, como em muitos vegetais e nas paredes celulares de bactérias. Os carboidratos mais simples são os monossacarídeos (como a D-Glicose, e podem ser aldose ou cetose), os quais podem ser unidos entre si por ligação glicosídica, com perda de uma molécula de água, para formar oligossacarídeos (como os dissacarídeos sacarose, lactose e maltose) e polissacarídeos (como a celulose, quitina, amido e glicogênio). Estes últimos tendem a ser muito heterogêneos, em tamanho e composição, pois não são construídos de acordo com um molde genético, como os ácidos nucleicos e as proteínas.

Os carboidratos também são constituintes de proteínas denominadas glicoproteínas, tendo estas um conteúdo variável de carboidratos, desde menos que 1% a mais de 90% em peso. As glicoproteínas também exercem diversas funções específicas, como as demais proteínas.

1.3.2 Ácidos nucleicos

As unidades básicas dos ácidos nucleicos são os nucleotídeos, sendo estes constituídos por uma base nitrogenada ou nucleobases (podendo ser pirimidina (Citosina, Timina e Uracila) e purinas (Adenina e Guanina)), um açúcar (desoxirribose e ribose, respectivamente no DNA e RNA) e um grupo fosfato. Além das cinco bases nitrogenadas, existem outras “não usuais”, de ocorrências menos comuns, como a Hipoxantina, N⁶-Dimetiladenina, 5-Metilcitosina, 5,6-Diidrouracila, Inosina. Tais bases são encontradas, mas não exclusivamente, no RNA de transferência.

Os ácidos nucleicos, DNA e RNA (iniciais, em Inglês, para respectivamente Ácido DesorriboNucleico e Ácido RiboNucleico), são constituídos por nucleotídeos unidos por ligação fosfodiéster, formando um esqueleto de açúcar-fosfato no polinucleotídeo, com as nucleobases unidas ao açúcar deste esqueleto por ligação glicosídica. A sequência de bases nitrogenadas contém a informação genética, e a sequência dos açúcares e fosfatos tem uma função estrutural formando o esqueleto do polinucleotídeo, originando assim uma estrutura regular na forma de dupla hélice para a molécula de DNA, mantidas por ligações de hidrogênio. O conhecimento da estrutura dos ácidos nucleicos permitiu a elucidação do código genético e o entendimento dos mecanismos de transmissão da informação genética das células mãe para as células filhas e a síntese de proteínas.

1.3.3 Lipídeos

Os Lipídeos (do Grego *lipos*, gordura) são constituídos por compostos similares estruturalmente, com predominância de grupamentos apolares (ou hidrofóbicos), por isso poucos solúveis em água e solúveis em solventes orgânicos, como clorofórmio, acetona, metanol, éter. Os lipídeos podem ser classificados em dois tipos, no primeiro há compostos de cadeias abertas, com cabeças polares e longas caudas apolares, como é o caso dos ácidos graxos, triacilgliceróis (conhecido como gordura e antigo triglicerídeos), fosfoacilgliceróis (ou fosfolipídeos), esfingolipídeos e os glicolipídeos (compostos formados por um carboidrato unido por ligação glicosídica a um grupamento álcool de um lipídeo). No segundo tipo de lipídeos há compostos de cadeia cíclica (com vários anéis fechados), como os esteróides (e.g., o colesterol). Os lipídeos formam as membranas celulares de organismos, hormônios esteróides, ácidos biliares e armazenam energia.

As membranas celulares (ou membranas plasmáticas) têm componentes lipídicos (em bicamada) e protéicos, sendo a interação entre dois componentes determinantes da função destas membranas, a saber, (i) para a separação entre interior e exterior da célula;



(ii) transporte de substâncias específicas para dentro e fora da célula; (iii) catálise feita por enzimas (tipo particular de proteína) ligadas à membrana; (iv) ligação de substâncias biologicamente importantes a proteínas receptoras na membrana. O arranjo em bicamada é mantido por interações não covalentes, como a força de van der Waals e interações hidrofóbicas, estando à cabeça polar em contato com o interior e exterior aquoso da célula, e a cauda apolar no interior da membrana.

A existência de membranas biológicas separando o interior e exterior celular, é fundamental para o desenvolvimento das células vivas, pois permitiu um contato mais próximo entre moléculas codificadoras (ácidos nucleicos) e catalíticas (proteínas) e sendo isoladas as células puderam competir e desenvolver-se no processo natural de seleção. Uma das teorias recente da origem da vida, a teoria de dupla-origem, tem procurado separar tal origem por: a) um desenvolvimento de catálise, feita pelo surgimento de agregados de moléculas capazes de catalisar reações; b) um sistema codificador, baseado em ácidos nucleicos; e a combinação destes dois processos teria originado a vida como a conhecemos hoje. Tem sido mostrado, em peptídeos isolados (Lee, D. et al. (1996)), que é possível a codificação e catálise ser executada tanto por peptídeos, quanto por DNA.

1.4 Aminoácidos: Os monômeros das proteínas

As unidades básicas, ou monômeros, das proteínas (do Grego, *proteios*, primário) são os aminoácidos, num total de 20 tipos, os quais são normalmente encontrados nas proteínas. Os aminoácidos comuns (conhecidos como α -aminoácidos) têm em sua estrutura (Figura 4a) um grupo amina e um grupo carboxila, ambos ligados a um átomo de carbono, denominado carbono α ; estando tal átomo ainda ligado a um hidrogênio e a uma cadeia lateral (representada pelo grupo *R*), diferente para cada um dos aminoácidos padrões. Os 20 aminoácidos possuem distintas propriedades físicas e químicas, isso leva, em parte, as proteínas a não apresentarem estruturas regulares e uniformes, como as já descritas para os ácidos nucleicos.

Os aminoácidos são “objetos” quirais, por não poderem ser sobrepostos às suas imagens especulares, como o que ocorre com as mãos esquerda e direita. Nos aminoácidos (a exceção da Glicina que é não quiral, devido seu grupo R ser um hidrogênio), o carbono α é um centro quiral ou assimétrico, com 4 diferentes grupos ligados a ele, originando estereoisômeros nas formas “L” e “D” (iniciais das palavras em Latin, respectivamente *Laevus* (que está do lado esquerdo) e *Dexter* (que está do lado direito)), cujas imagens especulares não se sobrepõem. Nas proteínas são encontrados somente aminoácidos na forma L (homoquiralidade), por exemplo, L-Alanina.

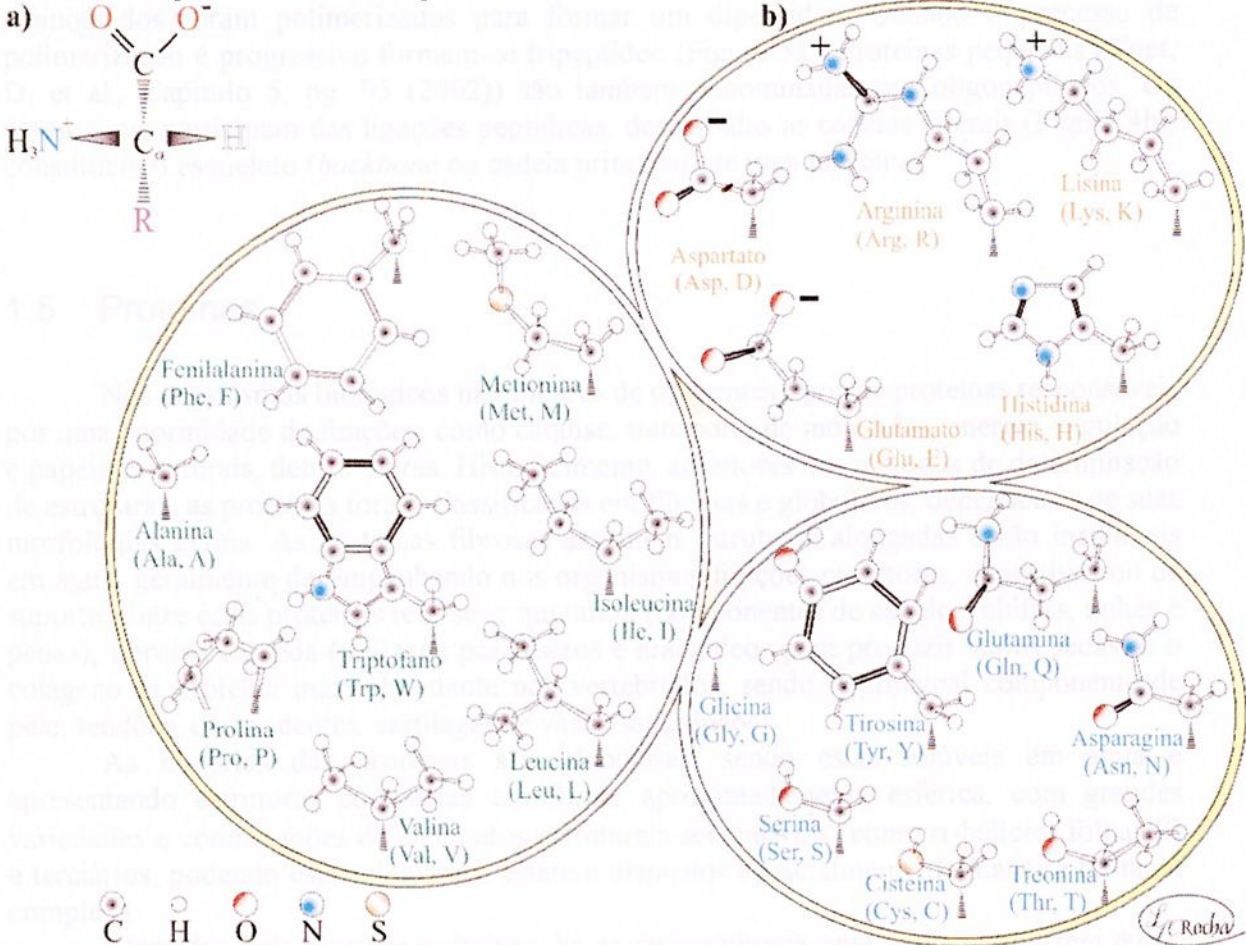
Além de quirais os aminoácidos são anfóteros, pois podem agir como ácidos ou como bases, já que seus grupos amino e carboxílico prontamente se ionizam, nunca assumindo a forma neutra em solução aquosa. Para os aminoácidos com cadeia lateral ionizável há três grupos ácidos base. Estes grupos ionizáveis são medidos por valores de pK (Voet, D. et al., Capítulo 4, pg. 79-85 (2002)), podendo estes serem influenciados por grupos vizinhos e alterados quando o aminoácido estiver numa cadeia polipeptídica.

Dentre os critérios utilizados para classificar os aminoácidos, se destacam as polaridades de suas cadeias laterais (Figura 4b). Alguns aminoácidos têm suas cadeias



laterais apolares, são eles a Alanina, Valina, Leucina, Isoleucina, Prolina, Fenilalanina, Triptofano e a Metionina, sendo estes aminoácidos geralmente encontrados no interior protéico, afastado do ambiente aquoso, devido ao efeito hidrofóbico (Capítulo 4).

Outros aminoácidos possuem cadeias laterais polares eletricamente neutras, em pH fisiológico; são eles a Glicina, Serina, Treonina, Tirosina, Cisteína, Glutamina e a Asparagina. Alguns destes aminoácidos podem estar preferencialmente no interior e outros na superfície da proteína. Por fim há aminoácidos com cadeias laterais polares carregadas, podendo tais cadeias conter grupos ácidos ou básicos; no primeiro caso têm-se os ácidos Glutâmico (ou Glutamato) e Aspártico (ou Aspartato) e no segundo a Histidina, Lisina e Arginina, tais aminoácidos geralmente preferem o contato com o solvente aquoso, assim ficando na superfície externa da proteína.



Quanto à polaridade, é comum classificar (Capítulo 4), mais resumidamente ainda, os aminoácidos, apenas em apolares (ou hidrofóbicos) e polares (ou hidrofílicos).

Além dos 20 aminoácidos citados anteriormente (Figura 4b), existem outros aminoácidos “incomuns”, que ocorrem em algumas proteínas, e são derivados dos aminoácidos em mudanças após a proteína ter sido sintetizada (em alterações pós-tradução); como é o caso da hidroxiprolina e a hidroxilisina (encontradas em algumas proteínas do tecido conjuntivo, como o colágeno) e a tiroxina (achada somente na glândula tireóide, na proteína tiroglobulina).

O grupo carboxila de um aminoácido pode reagir com o grupo amina de outro aminoácido, formando uma ligação amida e eliminando uma molécula de água, constituindo-se assim uma ligação peptídica entre aminoácidos; neste caso os dois aminoácidos foram polimerizados para formar um dipeptídeo. Quando o processo de polimerização é progressivo formam-se tripeptídeo (Figura 5a). Proteínas pequenas (Voet, D. et al., Capítulo 5, pg. 95 (2002)) são também denominadas por oligopeptídeos. Os átomos que participam das ligações peptídicas, descartadas as cadeias laterais (Figura 4b), constituem o esqueleto (*backbone* ou cadeia principal) de uma proteína.

1.5 Proteínas

Nos organismos biológicos há milhares de diferentes tipos de proteínas responsáveis por uma enormidade de funções, como catálise, transporte de moléculas, energia, regulação e papéis estruturais, dentre outras. Historicamente, anteriores aos métodos de determinação de estruturas, as proteínas foram classificadas em fibrosas e globulares, dependendo de suas morfologias gerais. As proteínas fibrosas assumem estruturas alongadas e são insolúveis em água, geralmente desempenhando nos organismos funções protetoras, conectivas ou de suporte. Entre estas proteínas tem-se: a queratina (componentes de cabelos, chifres, unhas e penas), fibroína da seda (utilizada por insetos e aracnídeos para produzir várias sedas) e o colágeno (a proteína mais abundante nos vertebrados, sendo o principal componente de pele, tendões, ossos, dentes, cartilagem e vasos sanguíneos).

As maiorias das proteínas são globulares, sendo estas solúveis em água e apresentando estruturas compactas em forma aproximadamente esférica, com grandes variedades e combinações de elementos estruturais secundários (como α -hélices, folhas β) e terciários, podendo estes elementos estarem dispostos espacialmente de maneira bastante complexa.

Além dos dois tipos de proteínas, há as de membrana, que também ocorrem numa enorme variedade de formas e tamanhos e de acordo com a interação com a membrana, estas proteínas podendo ser integrais ou intrínsecas (fortemente ligadas as membranas e para serem separadas, rompem com a membrana), periféricas ou extrínsecas (também ligadas as membranas, mas podem ser separadas destas, mantendo a membrana intacta) e as ligadas a lipídeos. Assim estas proteínas se localizam na membrana celular e mediam trocas de moléculas e informações através da fronteira celular.

Devido à movimentação na posição de um par de elétrons do grupo peptídico, a ligação peptídica, no esqueleto protéico, apresenta-se como uma estrutura de ressonância



híbrida (Figura 5b), tendo um caráter parcial (ao redor de 40%) de dupla ligação, em decorrência disto o grupo peptídico, que forma a união de dois aminoácidos, tem uma estrutura planar rígida (Figura 5c).

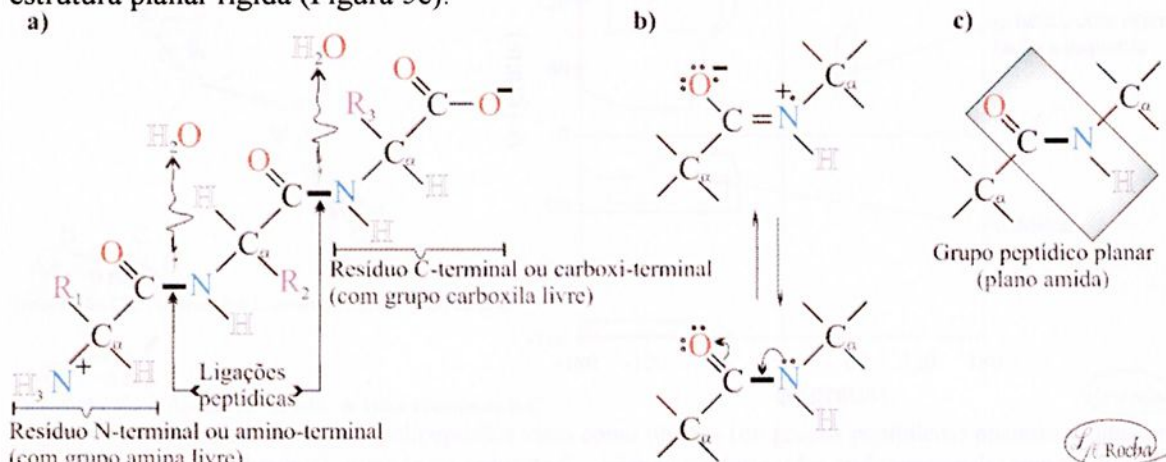


Figura 5 - a) Formação de tripeptídeo (constituído por três resíduos) com eliminação de duas molécula de água e formação de duas ligações peptídicas (C-N); b) A ligação peptídica é uma estrutura híbrida de estruturas de ressonância; c) Grupo peptídico rígido e planar, sendo que a conformação planar maximiza as sobreposições π que por sua vez respondem pela rigidez do grupo peptídico.

Obs. - Na nomenclatura de proteínas, tem-se primeiro o resíduo N terminal e por último o resíduo C terminal.

- O termo "resíduo" de aminoácido se refere a um aminoácido sem um Hidrogênio e/ou um grupo hidroxila.

O esqueleto de uma proteína pode ser descrito como uma sequência de grupos peptídicos rígidos e planares, não havendo rotação significativa em torno da ligação peptídica e nota-se rotações razoavelmente livres em torno das ligações que envolvem o átomo de carbono-alfa, em C_{α} -N (medida pelo ângulo ϕ) e C_{α} -C (medida pelo ângulo ψ). Portanto cada configuração³ do esqueleto pode ser descrita pela especificação dos ângulos ϕ e ψ , denominados ângulos de torção (ou ângulos diedrais, ou ângulos de rotação, ou ainda ângulos de Ramachandran, em homenagem a G.N.Ramachandran que sistematizou a representação configuracional por tais ângulos). Muitas das rotações nas ligações C_{α} -N (ϕ) e C_{α} -C (ψ) podem causar interferência estérica, com colisão e sobreposição entre átomos em diferentes resíduos de aminoácidos (Fitzkee, N. C. & Rose, G. D. (2004)). Os valores estericamente permitidos dos ângulos de torção ϕ e ψ indicam quais conformações da cadeia polipeptídica são permitidas e podem ser resumidos num diagrama, o denominado diagrama de Ramachandran (Figura 6).

³ As variações das estruturas 3D das biomoléculas são descritas em termos de configurações e conformações, tendo estes dois termos diferentes significados (Nelson, D. L. & Cox, M. M., Capítulo 3: Biomoléculas 41-63 (2002)). Neste trabalho nos referimos a configuração, como um dos possíveis arranjos espaciais da molécula. Por outro lado o termo conformação é utilizada para designar um conjunto de configurações que conferem uma certa forma à cadeia.

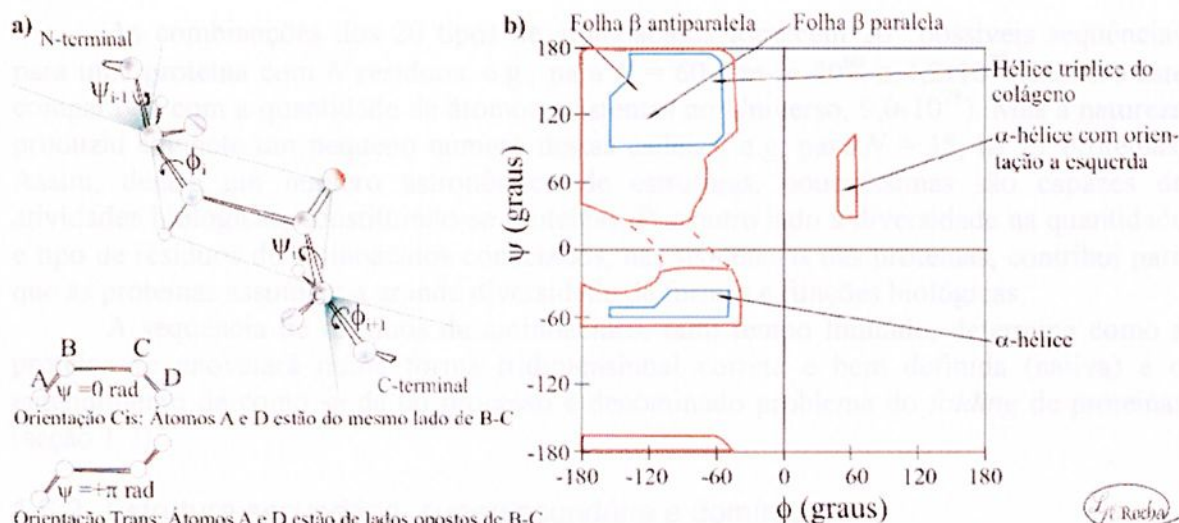


Figura 6 - a) Esqueleto da cadeia polipeptídica visto como objetos (ou grupos peptídicos) planares unidos em cantos opostos por dobradiças, girando ao redor de C_{α} , criando conformações tridimensionais, como α -hélices. Abaixo são destacados, pelo ângulo de torção " ψ ", para um sistema genérico ABCD (, e.g. $C_{\alpha}CNC_{\alpha}$), as orientações cis ($\psi = 0$ rad) e trans ($\psi = +\pi$ rad), sendo que os grupos peptídicos (Figura 5c) quase sempre adotam a orientação trans, mais estável que a cis.

b) Mapa de Ramachandran com valores de ϕ , ψ - Os contornos em cores delimitam regiões para todos os resíduos, exceto a Gly e Pro; em regiões completamente permitidas, em azul, e as parcialmente permitidas, em vermelho; já a região branca é decorrente de conflitos estéricos em muitas das combinações entre ϕ e ψ não verificadas em proteínas. São destacadas os ângulos de torções de várias estruturas secundárias.

Obs. - A orientação ou conformação cis é menos estável (~ 8 kJ/mol) do que a trans, devido a interferência espacial entre cadeias laterais de resíduos vizinhos, mas a Prolina, nas proteínas, tem $\sim 10\%$ de orientação cis.

- O mapa de Ramachandran foi publicado em 1963, quando nenhuma estrutura de proteína havia sido resolvida.

- Os grupos R (Figura 4b) são representados, no item "a", por esferas com cores púrpura e branca.

As combinações dos grupos peptídicos planares e das rotações dadas pelos ângulos de torções ϕ e ψ , para cada resíduo, tem importantes implicações para a maneira como conformação do esqueleto de uma proteína se dobra e para seus níveis de complexidades estruturais protéicos.

1.6 Níveis estruturais das proteínas

As proteínas, da mesma forma que outras moléculas poliméricas, como os ácidos nucleicos, podem ser analisadas em categorias estruturais, denominadas estruturas primária, secundária, terciária e quaternária.

1.6.1 Estrutura primária

A estrutura primária consiste simplesmente da ordem sequencial dos resíduos de aminoácidos ao longo da cadeia polipeptídica, o que confere o caráter unidimensional à cadeia polipeptídica. O conhecimento da estrutura primária combinada com a localização das pontes dissulfeto (ou ligação entre resíduos de Cisteínas), se tais pontes existirem, fornecem a estrutura covalente completa da proteína.

As combinações dos 20 tipos de aminoácidos fornecem 20^N possíveis seqüências para uma proteína com N resíduos, e.g., para $N = 60$ tem-se $20^{60} \cong 1,2 \times 10^{78}$ (número este comparável com a quantidade de átomos existentes no Universo, $9,0 \times 10^{78}$). Mas a natureza produziu somente um pequeno número destas cadeias, e.g, para $N = 35$, há 27 proteínas. Assim, dentre um número astronômico de estruturas, pouquíssimas são capazes de atividades biológicas, constituindo-se proteínas. Por outro lado a diversidade na quantidade e tipo de resíduos de aminoácidos conectados, nas seqüências das proteínas, contribui para que as proteínas assumam a grande diversidade de formas e funções biológicas.

A seqüência de resíduos de aminoácidos, num tempo limitado, determina como a proteína se enovelará numa forma tridimensional correta e bem definida (nativa) e o entendimento de como se dá tal processo é denominado problema do *folding* de proteínas (seção 1.7).

1.6.2 Estrutura secundária, supersecundária e domínios

Os átomos formadores do esqueleto protéico (portanto desconsiderando a cadeia lateral R) podem se acomodar, por meio das ligações de hidrogênio, em estruturas denominadas secundárias regulares, como hélices, folhas pregueadas e as voltas; sendo as mais importantes as α -hélices e folhas β (ou folhas β pregueadas), que são periódicas, apresentando valores repetidos dos ângulos de torção ϕ e ψ (Figura 6b).

Na α -hélice há pontes de hidrogênio, ao longo do eixo da hélice, entre os grupos C=O de um resíduo i ao grupo N-H do resíduo $i+4$ (Figura 7a); cada volta da hélice tem 3,6 resíduos por volta, numa distância (ou passo da hélice) de $5,4\text{\AA}$. Na estrutura helicoidal os átomos dos resíduos i e $i+4$ podem estar alinhados, assim conferindo força máxima as pontes de hidrogênio lineares, daí a grande estabilidade nestas estruturas. Os átomos das cadeias laterais projetam-se para fora da hélice, evitando assim interferência estérica, entre si e os átomos do esqueleto polipeptídico aglutinados na parte central da hélice.

Nas folhas β (Figura 7b) o esqueleto polipeptídico encontra-se distendido, podendo suas ligações de hidrogênio estar em diferentes partes do esqueleto de uma mesma cadeia (ou intra cadeia) da proteína ou entre diferentes cadeias (inter cadeias), estando as cadeias laterais se estendendo alternadamente para lados opostos, com distâncias de $7,0\text{\AA}$, entre dois resíduos. Se as cadeias ligadas por ligações de hidrogênio se estenderem nas mesmas direções e sentidos (das extremidades N-terminal para C-terminal), então haverá folhas β paralelas; mas se as cadeias estenderem em sentidos opostos, então as folhas β serão antiparalelas. Em folha β paralela há a necessidade de ligações cruzadas mais extensas, entre suas fitas, e fora do plano das folhas β , enquanto a folha β antiparalela pode ser conectada por apenas uma pequena alça.



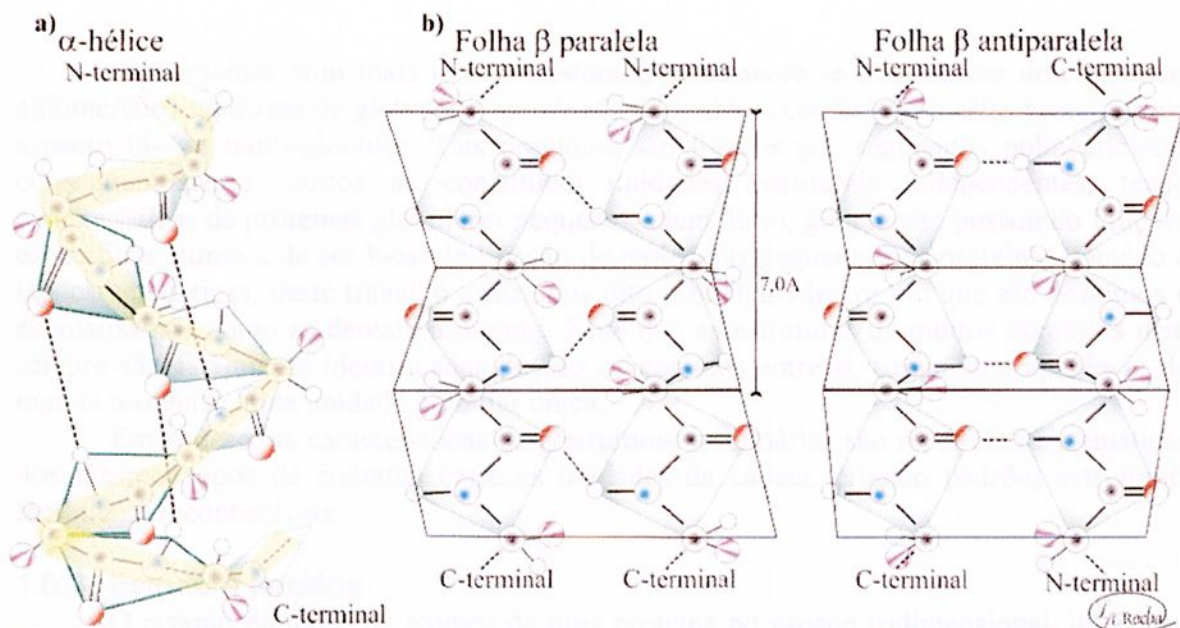


Figura 7 - As duas estruturas secundárias regulares proteicas mais importantes - a) α -hélice orientada a direita; b) folhas β pregueadas. Nas proteínas tais estruturas podem estar distorcidas das conformações ideais, mostradas nesta figura, e.g. isto ocorre com frequência nos giros iniciais e finais de uma α -hélice.

As ligações de hidrogênio nas α -hélices são paralelas às direções (de N para C terminal) da cadeia proteica e não perpendiculares a elas, como nas folhas β .

Obs. - As conformações em folhas β tem aparência enrugada ou pregueada nas bordas, daí folhas β pregueadas.

- As folhas β paralelas, com as ligações de hidrogênio mais distorcidas, são menos estáveis que as antiparalelas.

Em cadeias polipeptídicas muitas estruturas ao mudarem de sentido o fazem por estruturas secundárias, denominadas voltas reversas (ou curvaturas β), sendo que devido impedimentos estéricos os resíduos Gly e Pro costumam se acomodar mais adequadamente, sendo mais comuns, nestas estruturas.

Muitas proteínas não possuem estruturas regulares, como hélices e folhas pregueadas, daí tradicionalmente se utilizam os termos segmentos de estruturas aleatórias ou enovelamento aleatório (do Inglês *random coil*); contudo o termo “aleatório(a)”, neste contexto, parece ser inadequado (Richards, F. M. (1992)), pois tais estruturas complexas, além de exatas para a precisa função da proteína, também tem a mesma conformação nativa em diferentes moléculas de uma mesma proteína.

Neste trabalho (Capítulo 3) estaremos empregando o termo *random coil* para estruturas aleatórias, geradas na simulação Monte Carlo, apenas com o impedimento estérico (ou esfera rígida, representando a interação repulsiva do potencial de van der Waals) entre as unidades do sistema cadeia-solvente.

As folhas β e α -hélices podem ter uma ampla combinação a medida que a cadeia polipeptídica se dobra, produzindo vários tipos de estruturas supersecundárias (ou motivos), como exemplo destas estruturas tem-se as unidades $\beta\alpha\beta$, na qual um segmento de α -hélice conecta duas fitas de folhas β ; numa unidade $\alpha\alpha$ há duas α -hélices antiparalelas, com interação energética entre as cadeias laterais de seus resíduos. Há outros arranjos, como o meandro β , constituído por fitas de folhas β antiparalelas conectadas por voltas reversas; e arranjos em barril β , e.g. formado por uma série de meandros β conectados, ou por unidades $\beta\alpha\beta$ alternadas.

As proteínas com mais de 200 resíduos, geralmente se dobram em dois ou mais aglomerados na forma de glóbulo, denominados domínios, conferindo a estas proteínas um aspecto bi- ou multi-globular. Tais domínios são ligados por segmentos polipeptídicos, conseqüentemente muitos se constituem unidades estruturais independentes, tendo características de proteínas globulares pequenas; além disso, geralmente possuindo funções específicas, como a de ser local de ligação de moléculas pequenas nas proteínas. Devido a tais características, neste trabalho utilizamos oito proteínas (dentre 27) que são domínios e as analisamos como as demais proteínas. Note que as estruturas de muitos domínios nem sempre são facilmente identificadas devido os contatos entre si, tornando a aparência de muitas proteínas, uma unidade globular única.

Em síntese, as características das estruturas secundárias são repetições sistemáticas dos mesmos tipos de contatos entre as unidades da cadeia, criando padrões estruturais facilmente reconhecíveis.

1.6.3 Estrutura terciária

O arranjo de todos os átomos de uma proteína no espaço tridimensional, incluindo as posições dos átomos das cadeias laterais e de quaisquer grupos não padrões (não pertencentes aos 20 aminoácidos), bem como os dobramentos de elementos estruturais secundários, constituem as estruturas terciárias das proteínas. Detalhes de cerca de 35.000 destas estruturas, nos dias atuais - como as coordenadas de seus átomos - estão contidos num banco de dados, denominado "PDB" (iniciais do Inglês *Protein Data Bank*, **Berman**, H.M. et al. (2000)).

A separação entre estruturas secundária e terciária das proteínas não são fáceis de serem feitas (**Thornton**, J. M. (1992)), pois o dobramento do esqueleto protéico (estrutura secundária) pode influenciar, e ser influenciado, pelas cadeias laterais (parte da estrutura terciária) dos resíduos de aminoácidos.

A maneira como uma proteína se enovela para atingir sua conformação nativa não é ainda conhecida. Tem sido considerado como impraticável que a referida busca ocorra aleatoriamente, dentro do vasto espaço configuracional disponível para a cadeia protéica, como proposto originalmente por C. Levinthal (**Levinthal**, C. (1968)) e conhecido como paradoxo de Levinthal, conforme evidenciado pelo seguinte raciocínio:

Seja uma proteína de N resíduos, com apenas 2 valores de ângulos de torções ϕ e ψ , por resíduos; há assim $2N$ ângulos de torções para a proteína. Se cada resíduo tiver somente três configurações estáveis (desconsiderando as configurações das cadeias laterais, como gauche^+ , trans e gauche^-), então a proteína terá $3^{2N} \sim 10^N$ estruturas possíveis. Caso cada uma destas estruturas seja acessada num tempo mínimo de 10^{-13} s (tempo em que as ligações simples são reorientadas), então o tempo total t , para uma proteína experimentar ao acaso todas as estruturas possíveis, será da ordem de $t = 10^{N-13}$ s.

Por exemplo, para $N = 35$, tem-se $t = 10^{22}$ s, ou seja, mesmo para uma pequena proteína, o tempo típico de acesso à conformação nativa seria superior a idade do Universo ($1,5 \times 10^{10}$ anos ou $4,7 \times 10^{17}$ s, Figura 3), contudo as proteínas se enovelam nas suas formas funcionais nativas, numa escala temporal, geralmente entre $1\mu\text{s}$ (ou 10^{-6} segundo) até algumas horas. Consequentemente, hoje é amplamente aceito que, de alguma forma, a informação sobre como atingir rapidamente a estrutura nativa está inerentemente contida na estrutura primária da proteína.



1.6.4 Estrutura quaternária

Muitas proteínas (geralmente com peso molecular acima de 100 kDaltons) possuem múltiplas cadeias polipeptídicas ou subunidades, idênticas ou diferentes, e em diferentes quantidades. Utiliza-se o termo oligômero, quando uma proteína tem mais de uma subunidade (particularmente quando há duas, três e quatro cadeias, têm-se, respectivamente, dímero, trímero e tetrâmero). O arranjo espacial estável destas subunidades é identificado por estrutura quaternária das proteínas.

As subunidades podem interagir entre si por interações não covalentes, como ligações de hidrogênio e interações hidrofóbicas.

Quanto as forças de ligações dos diferentes níveis estruturais das proteínas, tem-se que: as estruturas primárias são mantidas por interações covalentes entre os resíduos de aminoácidos, contudo as outras estruturas superiores, secundárias, terciárias e quaternárias (Figura 8), são estabilizadas por interações não covalentes, como as interações hidrofóbicas e ligações de hidrogênio e em particular nas terciárias, por pontes dissulfeto.

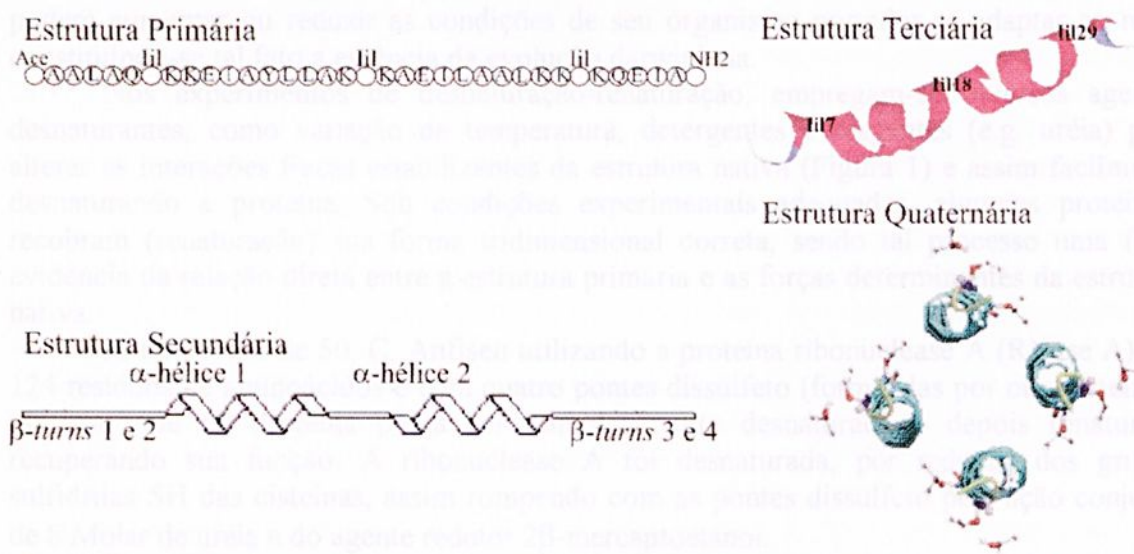


Figura 8 - Ilustração dos níveis estruturais protéicos, a partir de uma das proteínas, 1rh4, analisadas neste trabalho - Estruturas primária, secundária e terciária da 1rh4, sendo tal proteína um tetrâmero (estrutura quaternária).

Obs. Na estrutura primária da 1rh4 os nomes dos resíduos e do heterogrupa Iso-IsoLeucina (lil) estão, respectivamente num código de 1 e 3 letras.

As estruturas nativas das proteínas são estabilizadas pelo balanço delicado de diversas forças de origem não covalentes, que às vezes são concorrentes e outras vezes cooperativas entre si. Tais estruturas são, então, marginalmente estáveis, nas condições fisiológicas.

1.7 Folding de proteínas: A sequência determina a estrutura nativa

Em princípio, é tradicionalmente aceito que, sob condições fisiológicas as sequências de aminoácidos, de proteínas pequenas, contêm toda a informação necessária

para especificar como a proteína se enovela e se mantém em sua forma nativa única para exercer adequadamente sua função biológica (**Richards**, F.M. (1991); **Chan**, H.S. & Dill, K.A. (1993)). Esta assertiva tem sido corroborada por experimentos de mutação e desnaturação-renaturação.

1.7.1 Evidências da relação entre estruturas e funções nas proteínas

No caso de mutações de aminoácidos na estrutura primária é estudado o efeito provocado nas conformações nativas das proteínas. Verifica-se que dependendo da posição na sequência e da natureza dos resíduos substituídos, a mutação pode produzir desde efeitos insignificantes até a perda total da função da atividade da proteína. Um exemplo bem conhecido é o caso da anemia falciforme, em que a hemoglobina passa a ter, na posição seis, o ácido Glutâmico substituindo a Valina, em ambas as cadeias β , de sua estrutura quaternária, levando à agregação da hemoglobina e as hemácias a se deformarem na forma de foice, que dá o nome à doença.

Há outras centenas de formas mutantes conhecidas, devido a substituição de apenas um resíduo nas sequências das proteínas (**Goldenberg**, D. P. (1992)). Mutações em geral, podem aumentar ou reduzir as condições de seu organismo portador se adaptar ao meio constituindo-se tal fato a essência da evolução darwiniana.

Nos experimentos de desnaturação-renaturação, empregam-se diversos agentes desnaturantes, como variação de temperatura, detergentes e reagentes (e.g. uréia) para alterar as interações fracas estabilizantes da estrutura nativa (Figura 1) e assim facilmente desnaturando a proteína. Sob condições experimentais adequadas, algumas proteínas, recobram (renaturação) sua forma tridimensional correta, sendo tal processo uma forte evidência da relação direta entre a estrutura primária e as forças determinantes da estrutura nativa.

Já na década de 50, C. Anfinsen utilizando a proteína ribonuclease A (RNase A), de 124 resíduos de aminoácidos e com quatro pontes dissulfeto (formadas por oito cisteínas) mostrou que tal proteína podia ser completamente desnaturada e depois renaturada recuperando sua função. A ribonuclease A foi desnaturada, por redução dos grupos sulfidrilas SH das cisteínas, assim rompendo com as pontes dissulfeto pela ação conjunta de 8 Molar de uréia e do agente redutor 2 β -mercaptoetanol.

Na forma desnaturada os oito resíduos de cisteínas podem se ligar de 105 (ou 7!!) maneiras diferentes, porém somente uma destas combinações conduz às quatro pontes dissulfeto corretas. Tal fato acontece, quando os desnaturantes uréia e o β -mercaptoetanol são removidos, por diálise; em seguida a proteína é exposta a solução de O_2 em pH 8, oxidando os grupos SH, que passam a formar as pontes dissulfeto corretas, assim renaturando de forma espontânea a ribonuclease A, que passa a ter sua atividade enzimática recuperada. Em nosso atual trabalho há algumas proteínas, que como a ribonuclease, possuem 8 resíduos de cisteínas, formando 4 pontes dissulfeto.

Proteínas alteradas por mutação e experimentos de desnaturação-renaturação, assim como comparações configuracionais entre estruturas protéicas (Capítulo 2), oferecem uma oportunidade para analisar a relação entre suas estruturas e funções protéicas. No caso das comparações configuracionais, realizada por meio de métodos computacionais, podem-se, entre outros estudos, observar similaridades entre estruturas nativas; assim revelando detalhes sobre o *folding* e funções biológicas das proteínas e também o parentesco evolutivo das proteínas e os organismos que as geraram, desde milhões de anos atrás até os



dias atuais (**Holm, L. & Sander, C. (1996)**), como parece ser o caso do citocromo c (proteína responsável pelo transporte de elétrons), **Voet, D. et al. pg. 146-147 (2002)**.

1.7.2 Modelos para o processo de *Folding*

O mecanismo envolvido no processo de *folding* protéico, pode ser expresso por diversos modelos genéricos (**Baldwin, R. L. (1994)**; **Mirny, L. & Shakhnovich, E. (2001)**; **Dill, K. A. et al. (1995)**) que procuram estabelecer os mecanismos básicos do processo. Não há, atualmente, uma unificação destes modelos, pois o processo está sujeito a particularidades das seqüências das proteínas, ou seja, de diferentes forças que podem originar mecanismos distintos para dirigir as proteínas à suas formas nativas. Dentre tais modelos citamos:

- Modelo *framework*: As interações locais (ou entre resíduos próximos na seqüência das proteínas) interagem formando estruturas secundárias (hélices e *turns*), precocemente no processo de *folding*, que se juntam, por aglomeração, para determinar a estrutura nativa protéica.

- Modelo de colisão-difusão: As micro estruturas (ou micro domínios, constituído por segmentos contínuos de resíduos) se formam rápida e independentemente e as colisões e difusão destas estruturas, dão origem a subestruturas maiores (como estruturas secundárias parciais), que por uma ou várias trajetórias iniciais direcionam a estrutura nativa. O enovelamento por colisão-difusão é similar ao enovelamento no modelo *framework*.

- Modelo do colapso: Interações hidrofóbicas (**Baldwin, R. L. (2002)**) entre resíduos distantes, ao longo da cadeia protéica, conduzem a um rápido colapso da cadeia (com a formação de um glóbulo hidrofóbico), que pode preceder ou ocorrer concomitantemente com a formação de estruturas secundárias, e ambos levam ao estado nativo. As interações hidrofóbicas são consideradas um dos principais fatores determinantes do processo de *folding*.

- Modelo de nucleação (**Daggett, V. & Fersht, A. R. (2003)**): Baseado principalmente na cinética do enovelamento em dois estados, desenovelado e enovelado (**Baldwin, R. L. & Rose, G. D. (1999)**). As conformações na região de transição, destes dois estados, é o limiar entre o espaço conformacional multidimensional do *folding* e a base de atração para a estrutura nativa, formada por configurações concorrentes a um rápido enovelamento. O estado parcialmente enovelado, precedendo a forma nativa, denominado núcleo do enovelamento, é estabilizado por interações terciárias (estado *molten globule* ou glóbulo fundido), podendo conter elementos de estrutura secundária, e por meio de um processo cooperativo se chega à forma nativa.

Há outros modelos propostos para o processo de *folding* protéico, como o enovelamento hierárquico (**Srinivasan, R. & Rose, G. D. (1995)**) e o do funil (**Onuchic, J.N. et al. (1995)**).

Para analisarmos questões, como a compactação da cadeia protéica, empregamos um modelo simples, por simulação computacional, que trata o solvente de forma explícita.

1.7.3 Modelo simplificado para o sistema proteína-solvente

A simulação do sistema proteína-solvente conduzida neste trabalho, se dá no espaço contínuo (ou fora de uma rede de Bravais), com a cadeia protéica representada topologicamente, por um “colar de pérolas”, com 35 resíduos representados por esferas rígidas de diâmetro D e conectadas aos seus vizinhos por cordas idealmente flexíveis de comprimento definido. As moléculas do solvente, num total de 12.565, também são

representadas por “esferas” rígidas de diâmetro D . Detalhes e especificações do modelo são providos no Capítulo 3.

Para caracterizar grandezas relacionadas à atividade conformacional global das cadeias do modelo face as proteínas reais, geralmente são utilizadas a distância ponta a ponta, D_{pp} (distância euclidiana entre os centros de massas do primeiro e último monômeros) e o raio de giração, $R_{G,\alpha}$, já que ambas grandezas estão relacionadas com a forma e o tamanho destas cadeias (Doi, M. (1996)). Para uma cadeia de N monômeros, o raio de giração é definido como a raiz quadrada da média, do quadrado, das distâncias, $d_{\alpha,i}$, entre o i -ésimo monômero (ou centro de massa de um conjunto de átomos) e o centro de massa da cadeia em uma configuração particular α . Considerando monômeros de mesma massa tem-se que:

$$R_{G,\alpha}^2 = \frac{1}{N} \sum_{i=1}^N d_{\alpha,i}^2 \quad \text{ou} \quad R_{G,\alpha}^2 = \frac{1}{N^2} \sum_{i=j+1}^N \sum_{j=1}^{N-1} d_{\alpha,i,j}^2 \quad (1)$$

Para uma cadeia com N monômeros, têm-se que quanto mais compactas, menores são seus valores de $R_{G,\alpha}$, assim como quanto mais distendidas, maiores são seus valores de $R_{G,\alpha}$.

Experimentalmente, devido a intensidade de espalhamento de Raio X ser medida em função de $R_{G,\alpha}$, este é mais conveniente para expressar a dimensão de uma cadeia polimérica por $R_{G,\alpha}$, do que D_{pp} ; também teoricamente o emprego de $R_{G,\alpha}$ é mais conveniente, pois ao invés de se utilizar apenas uma distância, como em D_{pp} , têm-se todas as M distâncias (não repetidas) entre pares de monômeros i e j , numa combinação simples, dada por

$$M = \binom{N}{2} \quad (2)$$

, para $N = 35$ e não excluindo vizinho entre os monômeros, haverá um total de 595 distâncias (ou $35! / (33! 2!)$) intra cadeia para $R_{G,\alpha}$.

Para as demais grandezas, empregadas nos próximos Capítulos - a saber, função discrepância configuracional (Capítulos 2, 3 e 4), ordens de contato (Capítulo 3) e variação de vizinhos dos monômeros na cadeia polimérica (Capítulo 4) - serão excluídos dois vizinhos ao longo da cadeia, assim $M = 528$ e quando os monômeros estiverem distâncias abaixo de um valor predefinido d (Figura 1, Capítulo 3) estes estarão em contatos topológicos, sendo considerados vizinhos topológicos.

Quanto às proteínas reais, utilizadas neste trabalho, estas são simplifcamente representadas apenas pelas coordenadas dos centros de massas de seus resíduos de aminoácidos (Figura 9).

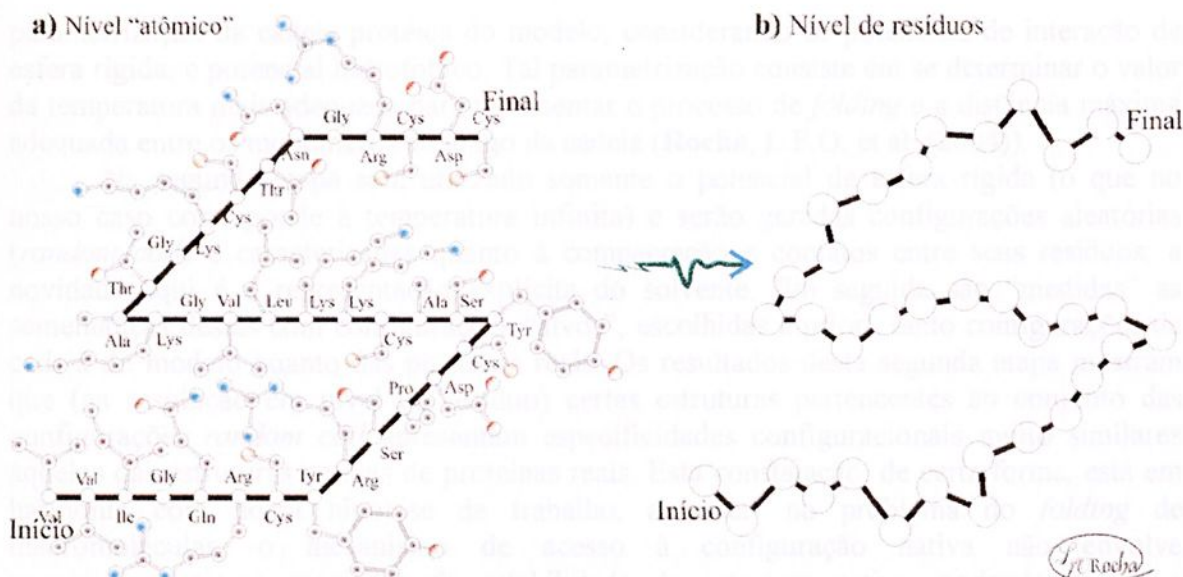


Figura 9 - Representação esquemática da estrutura nativa da proteína Itsk, a nível atômico e pelos centro de massa de seus resíduos - (a) A Itsk é mostrada na forma de "U" com sua cadeia principal representada apenas por um círculo branco, C_{α} , e nas cadeias laterais há os átomos de cada resíduo (Figura 4b), a menos dos Hidrogênios; (b) As proteínas, nos próximos três Capítulos, serão representadas apenas pelas coordenadas dos centros de massas de cada um de seus resíduos.

A seguir discorreremos sucintamente sobre os próximos Capítulos deste trabalho. Aliás todos os resultados mostrados nestes Capítulos foram obtidos a partir de programas desenvolvidos por nós e codificados em FORTRAN.

1.8 Sumário do conteúdo dos próximos Capítulos

No **Capítulo 2** introduzimos o ferramental técnico para nosso estudo do processo de *folding*, especificamente para tratar da comparação entre estruturas, de proteínas. De fato este assunto trata de um problema desafiador e atual, mas já com significativos progressos realizados nas últimas décadas, feito pelo emprego de vários métodos computacionais, em diferentes aplicações, como a de quantificar semelhanças e dessemelhanças estruturais. A comparação de estruturas protéicas (e/ou seus agrupamentos) por similaridade é também um dos tópicos importantes e atuais nas descobertas e desenvolvimento de novas drogas e outras pesquisas biomédicas.

Aqui definiremos e proporemos uma grandeza, a denominada função *discrepância configuracional*, para a comparação estrutural. Tal função será aplicada na quantificação de (des)semelhança entre estruturas idealizadas e entre estruturas de proteínas reais. Esta função, por ser simples conceitualmente, automática e rápida de ser estimada, além de geral quanto a sua aplicação, se torna adequada e conveniente para as comparações estruturais realizadas neste e nos próximos dois capítulos deste trabalho, principalmente para comparar estruturas de proteínas reais com configurações e conformações resultantes de modelos com pouca resolução espacial, como é o nosso caso.

No **Capítulo 3** o sistema cadeia protéica em solução aquosa será abordado por um modelo simplificado (simulação Monte Carlo), com resolução em nível de resíduos, e sua apresentação será considerada em duas etapas. Na primeira etapa é realizada a

parametrização da cadeia protéica do modelo, considerando os potenciais de interação de esfera rígida, e potencial hidrofóbico. Tal parametrização consiste em se determinar o valor da temperatura mais adequada para representar o processo de *folding* e a distância máxima adequada entre os monômeros ao longo da cadeia (Rocha, L.F.O. et al. (2004)).

Na segunda etapa será utilizado somente o potencial de esfera rígida (o que no nosso caso corresponde à temperatura infinita) e serão geradas configurações aleatórias (*random coil*) e caracterizadas quanto à compactação e contatos entre seus resíduos; a novidade aqui é a representação explícita do solvente. Em seguida são “medidas” as semelhanças destas com configurações “alvos”, escolhidas a priori, tanto configurações da cadeia do modelo quanto das proteínas reais. Os resultados desta segunda etapa mostram que (na resolução em nível de resíduo) certas estruturas pertencentes ao conjunto das configurações *random coil* apresentam especificidades configuracionais muito similares àquelas das estruturas nativas de proteínas reais. Esta constatação, de certa forma, está em harmonia com nossa hipótese de trabalho, a saber: no problema do *folding* de macromoléculas, o mecanismo de acesso à configuração nativa não envolve necessariamente os requisitos de estabilidade da estrutura nativa, podendo assim, o mecanismo de acesso e estabilidade, serem tratados independentemente.

Adicionalmente, a classificação do tamanho dos resíduos em apenas dois grupos, em resíduos “pequenos” e “grandes” e a sua distribuição ao longo da sequência das proteínas reais, fornece importantes *insights* para o entendimento da ocorrência da semelhança entre estruturas de proteínas reais e estruturas *random coil*.

No **Capítulo 4**, utilizamos o potencial hidrofóbico para o sistema proteína-solvente do modelo, baseados principalmente em evidências teóricas (Miyazawa, S. & Jernigan, R. L. (1996)) e experimentais (Kamtekar, S. et al. (1993)) que têm mostrado ser a força hidrofóbica (ou entrópica) a principal força responsável por guiar o processo de *folding* protéico. Tal força reflete a tendência de alguns aminoácidos em se aglutinar, reduzindo assim a exposição total da superfície hidrofóbica ao solvente. Este efeito acaba por induzir a proteína a uma forma compacta, com um núcleo hidrofóbico. Neste caso os 20 aminoácidos padrões, de acordo com a afinidade por água, podem ser qualitativamente classificados em hidrofóbicos (H) e polares (P).

Aqui atentaremos a aspectos essenciais do efeito hidrofóbico, através de (i) incorporar nos monômeros da cadeia protéica, do modelo, os padrões H ou P da sequência de proteínas globulares reais; (ii) as moléculas do solvente são incluídas explicitamente, para que o efeito hidrofóbico “puro” surja naturalmente da interação monômero-solvente. Nosso principal objetivo é avaliar em que extensão o padrão HP da sequência de uma proteína está relacionado com sua correspondente conformação nativa. Nossa hipótese é que tal padrão HP, desde o início do processo de *folding*, induz a proteína através de bem sucedidas trajetórias configuracionais, conduzindo a cadeia a sua conformação nativa única e biologicamente ativa.

Por fim, no **Capítulo 5** sintetizamos as principais conclusões do presente trabalho.

As proteínas utilizadas neste trabalho, num total de 27, independentes de seus tipos (fibrosas, globulares ou de membrana), têm em comum o mesmo número de aminoácidos, sendo suas informações, retiradas do “PDB”. Tais proteínas, ordenadas segundo seus raios de giração em ordem crescente, juntamente com demais detalhes de suas sequências e conformação nativa são fornecidos num Apêndice, após o Capítulo 5.

1.9 Bibliografia

- Baldwin**, R. L., *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions* (Serie B: Physics Vol. **325**) / S. Doniach Editor, Plenum Press, New York, Pathways and Mechanism of Protein Folding, 1-5 (1994).
- Baldwin**, R. L. & Rose, G. D. Is protein folding hierarchic? I. Local structure and peptide folding, *TIBS* **24**, 26-33 (1999).
- Baldwin**, R. L., Making a Network of Hydrophobic Clusters, *Science* **295**, 1657-1658 (2002).
- Berman**, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: *The Protein Data Bank*. *Nucleic Acids Research*, **28**, 235 (2000).
- Cairns-Smith**, A. G., The First Organisms, *Scientific American* **252** (6), 90-100 (1985).
- Campbell**, M. K. em: *Bioquímica*, 3ª edição ARTMED, Capítulo 4 (A Estrutura Tridimensional das Proteínas) (2000).
- Chan**, H.S. & Dill, K.A., The Protein Folding Problem, *Physics Today*, 24-32 (February 1993).
- Daggett**, V. & Fersht, A. R., Is there a unifying mechanism for protein folding?, *TRENDS in Biochemical Sciences* **28**, 18-25(2003).
- Daune**, M. in: *Molecular Biophysics Structures in Motion*, Oxford University Press Inc., New York, Chapter 13: Hydration of Protein, 277-289 (1999).
- Dill**, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P.D. and Chan, H. S., Principles of protein folding - A perspective form simple exact models, *Protein Science* **4**, 561-602 (1995).
- Doi**, M. in: *Introduction to polymer physics*, Oxford Science Publications, New York, Chapter 1: Properties of an isolated polymer molecules 1-19; Capítulo 2: Concentrated solutions and melts 20-44(1996).
- Editores** da revista Science e vários autores, What Don't We Know ?, *Science* **309**, 75-100 (2005).
- Holm**, L. & Sander, C., Mapping the protein universe, *Science* **273**, 595-602 (1996).
- Fitzkee**, N. C. & Rose, G. D., Steric restrictions in protein folding: An α -helix cannot be followed by a contiguous β -strand, *Protein Science* **13**, 633 (2004).
- Goldenberg**, D. P., *Protein Folding* / T. E. Greighton Editor, W. H. Freeman and Company, New York, Chapter 8: Mutational Analysis of Protein Folding and Stability, 353 - 403 (1992).
- Guth**, A., O Universo inflacionário, Editora Campus (1998).
- Kamtekar**, S., Schiffer, J.M., Xiong, H., Babik, J.M., Hecht, M.H. Protein design by binary patterning of polar and nonpolar amino acids, *Science* **262**, 1680-1685 (1993).
- Lee**, D. et al, Self-replicating Peptide, *Nature* **382**, 525-528 (1996).
- Levinthal**, C., Are there Pathways for Protein Foldign ? , *J. Chim. Phys* **65**, 44 (1968).
- Miller**, S. J. & Orgel, L. E. *The origin of Life on Earth*, Prentice-Hall, Englewood Clifts, New York, pg.85 (1974).
- Mirny**, L. & Shakhnovich, E., Protein Folding Theory: From Lattice to All-Atom Models, *Annu. Rev. Biophys. Biomol. Struct.* **30**, 361-396 (2001).
- Miyazawa**, S. & Jernigan, R. L., Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading , *J. Mol. Biol.* **256**, 623-644 (1996).



- Morris, R.**, *O que sabemos sobre o Universo*, Jorge Zahar Editor, Rio de Janeiro, Parte I (1999).
- Nelson, D. L. & Cox, M. M.**, *Lehninger Princípios de Bioquímica*, Sarvier, São Paulo, 3ª edição, Capítulos 5: Aminoácidos, Peptídeos e Proteínas 89-123; 9: Carboidratos e Glicoconjugados 225-249; 10: Nucleotídeos e Ácidos Nucléicos 250-279; 11: Lipídeos 280-300 (2002).
- Nelson, J. C., Saven, J. G., Moore, J. S., Wolynes, P. G.**, Solvophobic Driven Folding of Nonbiological Oligomers, *Science*, **277**, 1793-1796 (1997).
- Onuchic, J. N., Wolynes, P. G., Luthey-Schulten Z., Socci, N. D.**, Toward an outline of the topography of a realistic protein folding funnel, *Proc. Natl. Acad. Sci. USA* **92**, 3626-3630 (1995).
- Penrose, R.**, *A mente Nova do Rei: computadores, mentes e as leis da física*, Editora Campus Ltda, Rio de Janeiro, Capítulo 7 (2001).
- Richards, F. M.**, The Protein Folding Problem, *Scientific American*, 54- 63 (January 1991).
- Richards, F. M.** in: *Protein Folding*, ed. by T. E. Creighton, W. H. Freeman and Company, New York, Chapter 1: Folded and Unfolded Proteins: An Introduction, 1-55 (1992).
- Rocha, L. F. O., da Silva, M. A. A., Caliri A.**, Entropic force and folding of macromolecules, *Phys. Lett. A* **220**, 178-182 (1996).
- Rocha, L. F. O., Tarragó, M. E. P., Caliri, A.**, The Water Factor in the Protein-Folding Problem, *Brazilian Journal of Physics*, **34**, 90-101 (2004).
- Srinivasan, R. & Rose, G. D.**, LINUS: A Hierarchic Procedure to Predict the Fold of a Protein, *PROTEINS: Struct. Funct. and Genet.* **22**, 81-99 (1995).
- Tarragó, M. E. P., Rocha, L. F. O., daSilva, R. A., Caliri, A.**, Steric constraints as folding coadjuvant, *Physical Review E* **67**, 031901-1 a 031901-7 (2003).
- Thornton, J. M.**, *Protein Folding* / T. E. Creighton Editor, W. H. Freeman and Company, New York, Chapter 2: Protein Structures: The End Point of the Folding Pathway, 59 - 81 (1992).
- Voet, D., Voet, J. G., Pratt, C. W.**, *Fundamentos de Bioquímica*, ArtMed Editora, 1ª reimpressão (2002).
- Weinberg, S.**, *The first three minutes: A modern view of the origin of the Universe*, BasicBooks, New York, Chapter V, 1993.



CAPÍTULO 2

Comparações configuracionais: a função discrepância

No mundo macroscópico conseguimos, principalmente pelo sentido da visão, perceber razoavelmente as semelhanças, diferenças e complementaridade entre objetos, o que pode ser vantajoso em inúmeras situações cotidianas, além de úteis em atividades de lazer e em várias áreas científicas e tecnológicas em geral. Já a comparação entre configurações pode ser bastante dificultada, pois, por exemplo, precisa-se quantificar numericamente as semelhanças e diferenças entre as estruturas comparadas de forma muito rápida e consistente. Assim empregam-se alguns recursos para esta tarefa que pode variar enormemente dependendo da precisão almejada, daquilo que será medido e os objetivos a serem alcançados com a medida.

2.1 Comparação estrutural

O problema da comparação estrutural, vem de longa data. Já na década de 60, **Perutz, M. F. et al. (1960)**, por meio de alinhamento estrutural, tornou possível verificar similaridades estruturais entre duas proteínas diferentes, a saber, hemoglobina e mioglobina. A comparação entre estruturas tem contribuído muito para o atual entendimento da relação entre a estrutura e a função das proteínas.

A partir da comparação entre configurações de proteínas pode-se, por exemplo: (i) determinar similaridades, diferenças e complementaridades entre elas, contribuindo assim para o entendimento do relacionamento entre estas proteínas; (ii) ajudar a identificar domínios e similaridades estruturais em proteínas sem identidades significativas nas seqüências; (iii) medir proximidades funcionais entre proteínas, devido a dependência de suas funções com suas estruturas; (iv) procurar por estruturas similares a estrutura de uma nova proteína depositada no “PDB”, a partir de outras estruturas conhecidas, ajudando assim a entender o papel biológico da nova proteína desconhecida; (v) organizar o crescente número de milhares de formas assumidas pelas estruturas de proteínas depositadas no “PDB”, possibilitando a identificação de novos tipos de arquiteturas e relacionamentos evolutivo entre estas proteínas; (vi) contribuir para o entendimento do processo de *folding* que ocorre espontaneamente nos organismos biológicos.

Há diversos métodos para comparação de estruturas protéicas (Figura 1), mas não há como dizer qual seja o mais seguro e adequado para tal fim, pois cada medida se baseia numa concepção particular de similaridade estrutural, com o emprego de diferentes estratégias algorítmicas (**Koehl, P. (2001)** ; **Kolodny, R. et al. (2005)** ; **Ye, Y. & Godzik, A. (2004)**). Uma maneira tradicional de verificar aspectos equivalentes entre duas configurações é a inspeção visual destas configurações tentando sobrepô-las, normalmente com o auxílio da minimização de alguma grandeza, como o RMSD (iniciais do Inglês *Root of Mean Square Deviations*, **Koehl, P. (2001)**), ou com diferenças entre ângulos diedrais ou de valências, entre os mesmos pares de pontos correspondentes nas duas configurações.



Desta forma, pode-se comparar estruturas como um todo, ou por segmentos destas. Contudo a comparação visual pode apresentar inconvenientes, como as de projetar num plano (bidimensional) configurações tridimensionais, além de às vezes não ser possível identificar semelhanças em algumas estruturas, por estas estarem, de uma forma complexa, por exemplo, giradas umas em relação às outras. Outro exemplo é ilustrado pelo caso de alguns segmentos de resíduos em estruturas secundárias, tipo *loop* e *turn*, que podem dificultar a identificação visual de configurações parecidas. Por visualização direta, também não é possível identificar configurações que podem ser estruturalmente bastante diferentes, como pode ser o caso das configurações cujas distâncias pequenas são deslocadas em relação aos mesmos pares de distâncias em outras proteínas.

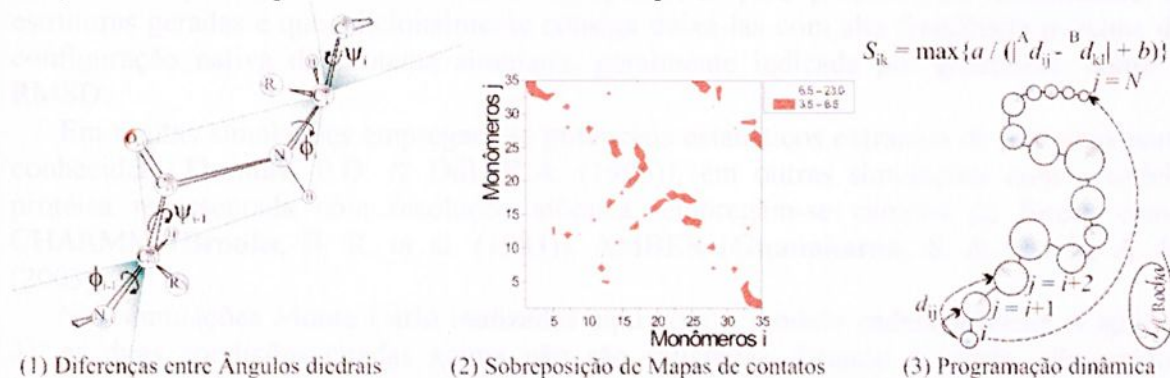


Figura 1 - Ilustração de alguns métodos para se obter a melhor correspondência estrutural possível (alinhamento estrutural) entre proteínas, sendo estes baseados em propriedades geométricas do conjuntos de pontos (coordenadas de átomos ou resíduos) nestas estruturas.

(1) Diferença entre cada par de ângulos diedrais entre duas estruturas A e B ($\phi^A - \phi^B$; $\psi^A - \psi^B$), Srinivasan, R. & Rose, G. D. (1985). Na Figura acima são mostrados dois valores dos ângulos diedrais, ϕ e ψ , de uma estrutura protéica hipotética, e.g. A.

(2) Sobreposição (ou alinhamento) entre mapas de contatos das estruturas A e B, Pelta, D. et al. (2005). O mapa de contato acima (da proteína 1tsk, nº 4) é a representação de todos os contatos (pontos em vermelho) e em não contatos (em branco) de uma estrutura. Cada par de resíduos (dado por seus centros de massas), não consecutivos (e.g. excluindo i com $i \pm 1$ e $i \pm 2$), estarão em contato, quando suas distâncias euclidianas forem inferiores a uma distância limiar, no caso 6,5 Å.

(3) Através de uma função de pontuação de similaridade (como S_{ik} , Taylor, W.R. & Orengo, C. A. (1986)) compara-se as respectivas matrizes distâncias interatômicas (d_{ij}^A, d_{kl}^B) nas estruturas ou subestruturas A e B tentando minimizar suas distâncias intermoleculares, e portanto maximar S_{ik} , daí a notação $\max \{ \dots \}$ em S_{ik} . Na Figura acima há a representação de uma estrutura hipotética A, de tamanho N , com distâncias interatômicas d_{ij}^A . Obs. Na função S_{ik} , tem-se a, b como duas constantes e d_{ij} as distância entre os átomos i com os j na estrutura A acima. Há outras expressões alternativas para S_{ik} .

A comparação entre duas estruturas pode ser reformulada, de forma geral, por meio do problema de comparar dois conjuntos quaisquer de pontos dispostos no espaço tridimensional (3D), como a comparação entre estruturas idealizadas (seção 2.3 abaixo). Contudo, geralmente a comparação configuracional é feita para casos particulares, como os realizados na seção 2.4 deste Capítulo.

No geral há três classes de comparações configuracionais. Na primeira classe são feitas comparações entre conformações obtidas experimentalmente da mesma proteína, e.g., por diferentes resultados de cristalografia de Raio X, ou entre as diversas estruturas obtidas por espectroscopia RMN, como ilustrado na seção 2.4, deste Capítulo. Na segunda classe há o confronto entre configurações obtidas experimentalmente com as obtidas por modelos (geradas por simulação computacional, como pelo método Monte Carlo), como aplicado no

Capítulo 3. A terceira classe refere-se a comparações estruturais realizadas numa estrutura por vez (Jacobs, D.J. et al. (2001)), especificamente para averiguar segmentos (ou regiões) flexíveis e rígidos numa dada proteína.

Particularmente nos métodos tradicionais, a realização de comparações configuracionais entre proteínas e modelos (2ª classe acima), geradas numa simulação computacional, quase sempre, requerem as seguintes condições:

- As estruturas comparadas devem estar na mesma escala, isto é com os mesmos tamanhos de suas unidades (à nível de resíduo ou nível atômico), ângulos e comprimentos de ligações entre as unidades.

- Na simulação, geralmente utiliza-se um potencial para proporcionar estabilidade às estruturas geradas e que adicionalmente consiga deixá-las com alta frequência próxima da configuração nativa da proteína simulada, geralmente indicada por grandezas, como o RMSD.

Em muitas simulações empregam-se potenciais estatísticos extraídos de proteínas reais conhecidas (Thomas, P.D. & Dill, K.A. (1996)); em outras simulações com a cadeia protéica representada com resolução atômica, empregam-se campos de forças como CHARMM (Brooks, B. R. et al. (1983)), AMBER (Gnanakaran, S. & Garcia, A. E. (2003)).

Nas simulações Monte Carlo realizadas aqui com o modelo cadeia-solvente (Capítulo 3), as duas condições citadas acima não são satisfeitas: Quanto à escala, são usados monômeros todos de mesmo tamanho e também não são utilizados os ângulos e comprimentos de ligações verificados nas proteínas; já quanto ao potencial no presente caso (Capítulo 3) somente o potencial de esfera rígida é analisado. Apesar das limitações, estes dois potenciais serão utilizados para entendermos de forma gradual, progressiva e separadamente, suas influências diretas no processo de *folding* protéico.

Devido às peculiaridades do modelo utilizado, torna-se inadequado o emprego dos métodos tradicionais para comparações configuracionais. Por isso, a função discrepância, representada por ΔD , foi introduzida para quantificar adequadamente, conforme nossas necessidades, as diferenças entre duas configurações quaisquer. O valor de ΔD depende da topologia configuracional das configurações envolvidas, e este ponto é o centro das análises que serão feitas neste Capítulo.

2.2 A função discrepância configuracional: definição

De forma geral, necessita-se fazer comparações entre quaisquer duas estruturas (ou dois arranjos espaciais de monômeros conectados), inclusive entre partes destas, mas para isso as distâncias (d') entre os monômeros em cada um dos casos devem ser apropriadamente re-escaladas e normalizadas pela relação

$$d = (d' - d'_{min}) / (d'_{max} - d'_{min}) \quad (1)$$

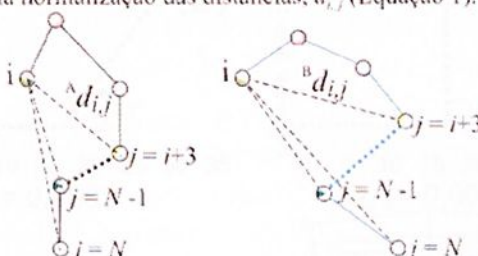
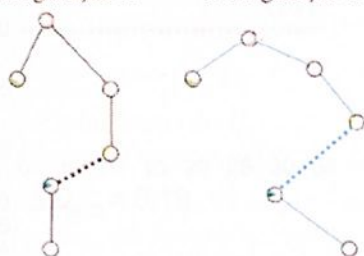
, onde d é a nova distância normalizada, que varia no intervalo $[0 - 1]$, e $d'_{min} \neq d'_{max}$, são, respectivamente, as distâncias mínimas e máximas entre os pares de monômeros separadamente em cada uma das configurações a serem comparadas. No caso das configurações protéicas e modelo (simulação) as distâncias (d') são medidas, respectivamente entre os centros de massa de cada resíduo de aminoácido e entre os centros dos monômeros esféricos.

1) Escolha de duas configurações (ou estruturas) A, B, com N monômeros, a serem comparadas

2) Determinação de todas as distâncias, $d'_{i,j}$, entre os monômeros (i, j) nas duas configurações, eliminados dois pares primeiros vizinhos ao longo da cadeia, ou seja, $n = 2$ e $M = 528$; seguida pela normalização das distâncias, $d_{i,j}$ (Equação 1).

Configuração A

Configuração B



3) Cálculo da função discrepância ΔD (Equação 5) entre as configurações A e B, para todas as M distâncias, dada por:

$$\Delta D(A, B) = \frac{1}{528} \sum_{i=1}^{N-3} \sum_{j=i+3}^N |d_{i,j}^A - d_{i,j}^B|$$

Figura 2 - A comparação estrutural entre as configurações A e B quaisquer é dada pelo valor da função discrepância ΔD (A, B), ou apenas ΔD , sendo esta uma grandeza global envolvendo duas configurações, e se utiliza de todas as M distâncias (não repetidas, $d_{i,j} = d_{j,i}$) entre pares de monômeros. Para o cálculo da função ΔD (A, B) é feito um alinhamento estrutural das duas configurações nos procedimentos 1, 2 e 3 acima.

Para quantificar semelhanças e dessemelhanças entre pares de estruturas, com a função discrepância configuracional, ΔD , estas são organizadas em estrutura mais alongadas e mais compactas, sendo estas: 1) configurações ideais: na forma de segmentos de linhas retas e hélices; 2) configurações de proteínas: escolhidas do "PDB". No capítulo 3, também com a função ΔD , verificamos semelhanças entre configurações do modelo e proteínas reais.

As semelhanças entre estruturas idealizadas podem ser bastante úteis no entendimento de outras estruturas mais diretamente relacionadas com as proteínas, como as hélices, ou ainda combinações diversificadas de estruturas secundárias, como hélices, folhas, *turns* e *loops* em proteínas reais. Tais estruturas idealizadas são simplificadas na forma de hélices triangulares e quadradas, e segmentos de linhas com dobras de diferentes inclinações. As hélices triangular e quadrada têm essas denominações por possuírem, respectivamente três e quatro unidades por volta, sendo estes valores próximos aos 3,6 resíduos por volta numa alfa-hélice.

2.3 Discrepância cruzada entre estruturas idealizadas: semelhanças entre uma estrutura retilínea com estruturas retilíneas parcialmente dobradas e estruturas helicoidais

A forma mais simples para dispor N monômeros, separados por distâncias não nulas, é numa estrutura retilínea, por isso tal estrutura será tomada como referência para ser comparada, através da função discrepância ΔD , com outras estruturas também idealizadas (Figura 3), sendo estas últimas retas inclinadas ao meio (inclinadas de um ângulo fixo a partir do 19ª unidade), e helicoidais (nas formas triangular e quadrada). Para todas as estruturas idealizadas foi mantida fixa a distância unitária entre unidades consecutivas, sendo todas as unidades pontuais.

Após normalizadas as distâncias; para quantificar o quanto ocorrem convergência entre duas configurações, utilizamos uma grandeza baseada em conceitos estatísticos. Seja um par de monômeros (i, j) numa configuração A e o mesmo par numa outra configuração B. Então definimos

$$\Delta D_{i,j} = |^A d_{i,j} - ^B d_{i,j}|^\gamma \quad (2)$$

, onde $^A d_{i,j}$ e $^B d_{i,j}$ são, respectivamente, os valores medidos das distâncias normalizadas entre os monômeros i e j nas duas configurações e assim $\Delta D_{i,j}$ é a discrepância entre estas medidas. O expoente γ foi adotado aqui como unitário (ou seja, $\gamma = 1$) para não privilegiar valores de distâncias grandes ou pequenos de $\Delta D_{i,j}$.

Para uma sequência de N monômeros, se forem excluídos os n pares primeiros vizinhos, de cada monômero ao longo da cadeia, teremos, M distâncias (não repetidas) entre pares de monômeros, numa combinação simples, dada por

$$M = \binom{N-n}{2} \quad (3)$$

Assim, para $N = 35$, se não forem excluídos nenhum vizinho ($n = 0$) haverá o total de 595 distâncias (ou $35!/(33!2!)$); caso seja excluído os pares primeiros vizinhos ($n = 1$) terá $M = \binom{N-1}{2} = 561$, e assim sucessivamente. No nosso caso adotamos $n = 2$, daí $M = 528$. Com

isso, a discrepância não tem influência direta dos monômeros próximos na sequência da cadeia, pois naturalmente suas distâncias são pequenas (ou seja, há altas correlações entre elas), independentes da topologia da configuração no espaço 3D. Assim, a função discrepância configuracional é uma função definida pelo seguinte valor médio

$$\Delta D = \frac{1}{M} \sum_{i=1}^{N-\alpha} \sum_{j=i+\beta}^N \Delta D_{i,j} \quad (4)$$

, e corresponde a uma característica global entre duas configurações A e B quaisquer.

Em nossas comparações configuracionais temos usado $\alpha = \beta = 3$ e $M = 528$. O valor da discrepância ΔD , satisfaz $0 \leq \Delta D \leq 1$, e sua função é mostrar (Figura 2) o quanto a configuração qualquer A, por exemplo, geradas na simulação se assemelha a outra configuração B (como, a configuração nativa de uma proteína real). Quanto menor ΔD , maior é a semelhança entre as configurações. Naturalmente para $\Delta D = 0$ corresponde a configurações idênticas. A função discrepância configuracional será utilizada em várias comparações configuracionais neste trabalho, e nestes casos sempre repetiremos o procedimento acima, ou seja, normalizamos as distâncias separadamente em cada estrutura (equação 1) e em seguida calculamos ΔD (equações 2, 3 e 4). Eliminando os dois pares primeiros vizinhos ($n = 2$) ao longo das cadeias, a equação 4, se torna

$$\Delta D_{A,B} = \frac{1}{528} \sum_{i=1}^{N-3} \sum_{j=i+3}^N |^A d_{i,j} - ^B d_{i,j}| \quad (5)$$

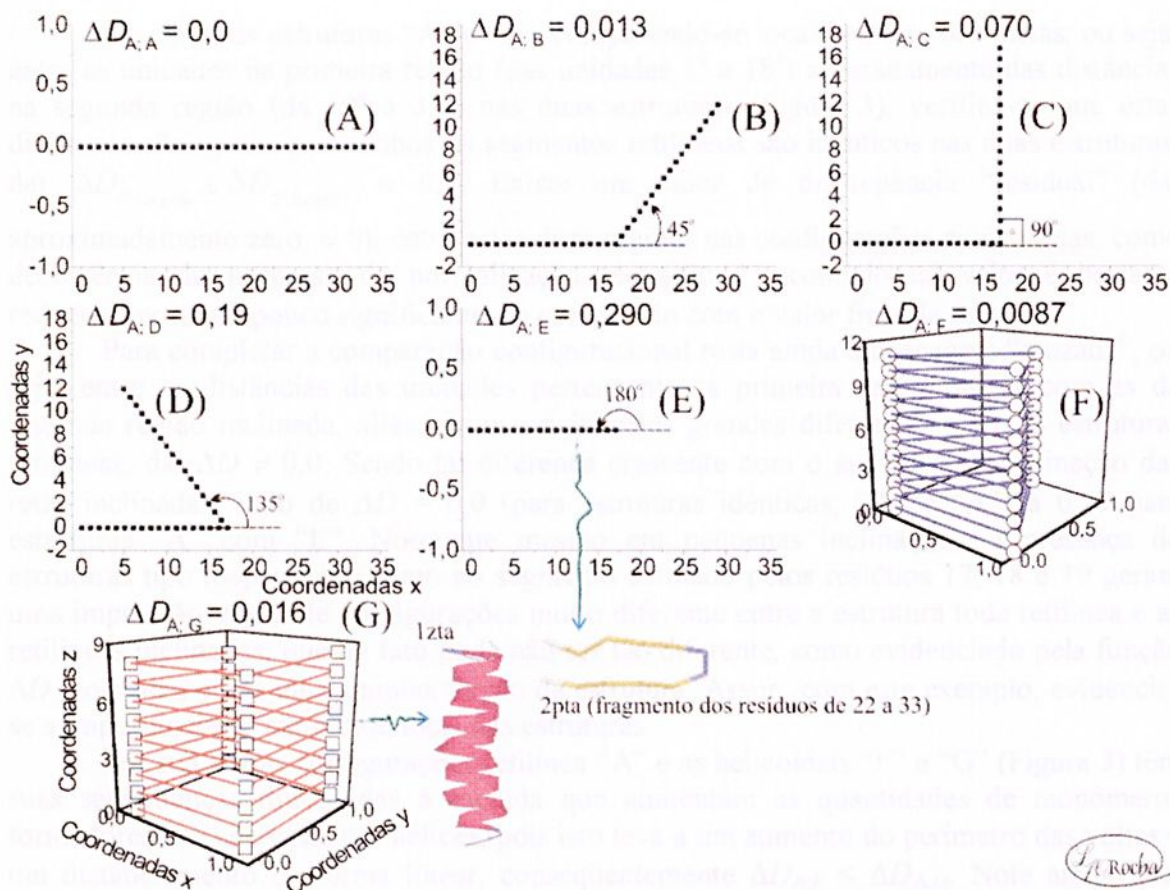


Figura 3 - Aplicando a função discrepância, $\Delta D_{A:A, B, \dots, G}$ em comparações configuracionais entre a estrutura, alvo, toda retilínea "A" com as estruturas retilíneas inclinadas "B" até "E" (a partir da 19ª unidade) e helicoidais triangular "F" e quadrada "G" - As semelhanças indicadas acima de cada estrutura, vai desde zero (entre as estruturas idênticas "A" e "A") até cerca de 0,3 (entre as estruturas "A" e "E"). Ao redor das estruturas "E" e "G" são mostrados os equivalentes destas estruturas nas proteínas reais, respectivamente folhas betas antiparalelas (em amarelo unidas por uma volta beta, em azul, num fragmento da proteína 2pta) e uma alfa-hélice (em rosa, em quase toda a extensão da proteína 1zta).

Obs. - Em todas as estruturas acima, as vezes em diferentes escalas dos eixos cartesianos (x, y, z), a primeira unidade se encontra na origem (0,0,0). As estruturas de "A" a "E" são planas, assim o valor de z é uma constante arbitrária qualquer.

- A estrutura "E" assemelha-se a uma folha beta antiparalela, tendo esta uma grande discrepância com a estrutura retilínea, conforme expresso por $\Delta D_{A:E}$. Para deixar a estrutura "E" mais parecida com uma folha beta, basta deixar uma distância unitária entre as unidades 1 a 18 e as unidades 19 a 35, porém isso abaixa apenas em um centésimo $\Delta D_{A:E}$, ou seja, de 0,29 para 0,28; neste caso a semelhança entre uma folha beta e uma alfa-hélice (tipo "G") seria $\Delta D_{E:G} = 0,275 \approx \Delta D_{A:E}$.

Ao observar as estruturas mostradas na Figura 3, pode-se indagar:

Qual a discrepância entre "A" e "B"? Em quanto a discrepância entre "A" e "B" é menor do que entre "A" e "G"?

Ainda que a comparação configuracional entre as estruturas possa ser feita globalmente, sua análise pode ser realizada localmente; neste caso pode ser adequado observar duas regiões distintas, sendo a primeira, idêntica nas duas estruturas e a segunda muito diferente, como no caso da comparação entre "A" e "B" da Figura 3.

No caso das estruturas “A” e “B”, comparando-se localmente as distâncias, ou seja, entre as unidades na primeira região (das unidades 1ª a 18ª) separadamente das distâncias na segunda região (da 19ª a 35ª) nas duas estruturas (Figura 3), verifica-se que estas distâncias são iguais, pois ambos os segmentos retilíneos são idênticos nas duas estruturas, daí $\Delta D_{1^a \text{ Região}} + \Delta D_{2^a \text{ Região}} \approx 0,0$. Existe um valor de discrepância “residual” (daí aproximadamente zero, ≈ 0), entre estas duas regiões nas configurações comparadas, como decorrência do processo de normalização (equação 1), contudo tal valor é bastante reduzido, portanto pouco significativo, se comparado com o valor final de ΔD .

Para completar a comparação configuracional resta ainda a checagem “cruzada”, ou seja, entre as distâncias das unidades pertencentes à primeira região linear com as da segunda região inclinada, aliás, daqui originam as grandes diferenças entre as estruturas retilíneas, daí $\Delta D \neq 0,0$. Sendo tal diferença crescente com o aumento da inclinação das retas inclinadas, indo de $\Delta D = 0,0$ (para estruturas idênticas, “A” e “A”) à 0,29 para estruturas “A” com “E”. Note que mesmo em pequenas inclinações, a presença de estruturas tipo *loop* e *turn*, como no segmento formado pelos resíduos 17, 18 e 19 geram uma impressão visual de configurações muito diferente entre a estrutura toda retilínea e as retilíneas inclinadas, que de fato pode não ser tão diferente, como evidenciado pela função ΔD , pois trata-se de uma simples flexão da estrutura. Assim, com este exemplo, evidencia-se a importância da similitude local das estruturas.

Por sua vez as configurações retilínea “A” e as helicoidais “F” e “G” (Figura 3) têm suas semelhanças diminuídas à medida que aumentam as quantidades de monômeros formadores de cada volta das hélices, pois isto leva a um aumento do perímetro das voltas e um distanciamento da forma linear, conseqüentemente $\Delta D_{A:F} < \Delta D_{A:G}$. Note ainda que, configurações visualmente bastante diferentes podem refletir globalmente valores bastante próximos da função discrepância. Novamente aqui temos o peso das similitudes locais das estruturas atuando com peso significativo, mas em direção diferente do caso anterior, sendo $\Delta D_{A:B} \cong \Delta D_{A:G}$.

A α -hélice possui 3,6 resíduos por volta, ou seja, quantidade intermediária de resíduos por volta das hélices triangular e quadrada. Portanto, supõe-se a partir dos resultados obtidos para $\Delta D_{A:F}$ e $\Delta D_{A:G}$, que configurações protéicas, com longas α -hélices, ou mesmo uma fita simples de DNA, possam resultar em pequenas discrepâncias estruturais entre si. Será mostrado, no Capítulo 3, que estruturas distendidas, não retilíneas, do modelo, quando comparadas com proteínas reais, contendo grandes segmentos contínuos de α -hélices, possuem discrepâncias somente um pouco acima de 0,02 ($\cong \Delta D_{A:G}$).

Para estruturas helicoidais a elongação de suas voltas pode influenciar na semelhança com outras estruturas, como entre a retilínea “A” e a hélice quadrada “G” em diferentes elongações (Figura 4).

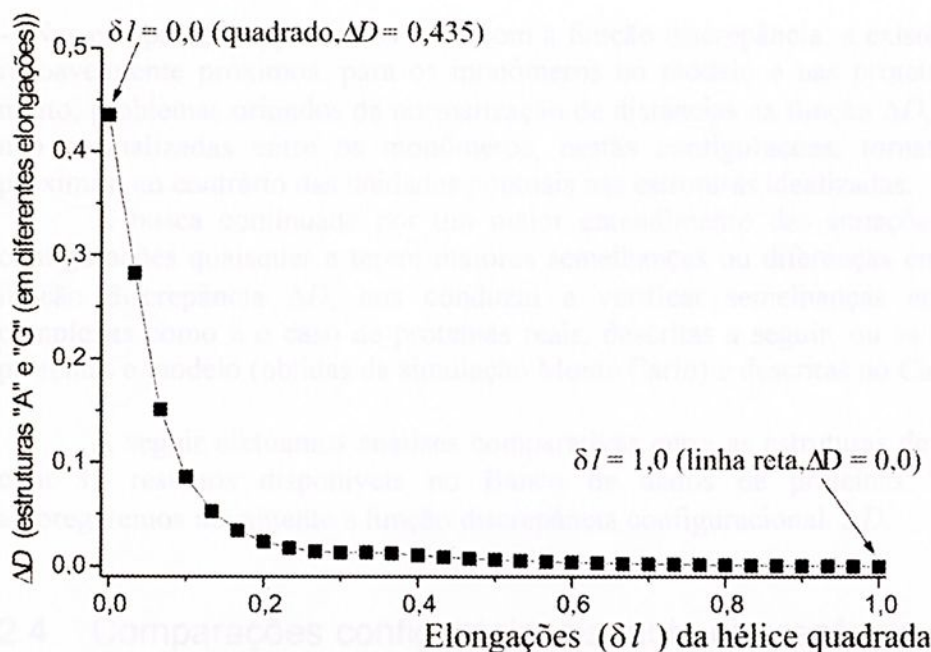


Figura 5 - Semelhanças entre a estrutura retilínea "A" com a hélice quadrada "G" em diferentes elongações - a medida que a hélice quadrada aumenta sua elongação, a semelhança com a configuração retilínea aumenta drasticamente, saindo de $\Delta D = 0,44$ em $\delta l = 0,0$ (quadrado) e atingindo quase $\Delta D = 0,0$ (valor este exato para $\delta l = 1,0$), a partir de $\delta l = 0,80$, como mostrado no gráfico de $\Delta D \times \delta l$.

Observando a variação da discrepância ΔD , com a elongação δl , nota-se um decaimento tipo exponencial (Figura 5), tendo em $\delta l = 0,0$ (hélice comprimida num quadrado) a discrepância máxima, $\Delta D = 0,44$; à medida que a hélice começa a ficar esticada ($\delta l \neq 0,0$) a discrepância com a configuração retilínea diminui drasticamente, assim, a diferença máxima entre a hélice (usada aqui) e a sequência retilínea de monômeros é 0,44. Mesmo para pequenas elongações $\delta l \cong 0,1$ a discrepância cai para $\Delta D \leq 0,1$, pois a partir desta elongação começam existir grandes diversidades nas distâncias, entre as unidades da hélice, compatíveis com as existentes na configuração retilínea. As diferenças significativas entre as estruturas "A" e "G", ocorrem apenas em pequenas elongações ($0,0 \leq \delta l \leq 0,40$), fora desta faixa de elongação a discrepância, se torna independente de δl permanecendo praticamente constante próxima de 0,0.

Os resultados da comparação configuracional entre estruturas idealizadas, evidência que deve se ter cuidado na utilização da função discrepância, pois em casos extremos, como hélices comprimidas o processo de normalização de distâncias (ingrediente fundamental no cálculo de ΔD) unifica as distâncias (no intervalo 0 - 1) nas duas configurações muito diferentes tornando-as muito parecidas; este parece ser o caso para $0,8 \leq \delta l \leq 0,1$ nas hélices quadradas. Situações como estas não desqualificam a função ΔD para comparações configuracionais, pois:

- Esta função como todas as outras utilizadas com o mesmo objetivo, por mais genéricas que possam ser, há situações específicas, como a mostrada acima, onde se deve ter ponderações quanto aos resultados fornecidos, pois a função calcula, por meio de um re-escalamento de distâncias, semelhanças e diferenças solicitadas entre quaisquer duas estruturas.
- Um dos nossos objetivos aqui é evidenciar não apenas o quanto à função ΔD funciona bem em comparações configuracionais, mas também suas limitações.

- Nas comparações que temos feito com a função discrepância, a existência de tamanhos, razoavelmente próximos, para os monômeros no modelo e nas proteínas diminuem, em muito, problemas oriundos da normalização de distâncias na função ΔD , pois as distâncias, não normalizadas entre os monômeros, nestas configurações, tornam-se naturalmente próximas, ao contrário das unidades pontuais nas estruturas idealizadas.

A busca continuada por um maior entendimento das situações que levam duas configurações quaisquer a terem maiores semelhanças ou diferenças entre si, utilizando a função discrepância ΔD , nos conduziu a verificar semelhanças em estruturas mais complexas como é o caso de proteínas reais, descritas a seguir, ou as comparações entre proteínas e modelo (obtidas da simulação Monte Carlo) e descritas no Capítulo 3.

A seguir efetuamos análises comparativas entre as estruturas de todas as proteínas com 35 resíduos disponíveis no Banco de dados de proteínas "PDB", para isto empregaremos novamente a função discrepância configuracional, ΔD .

2.4 Comparações configuracionais: auto discrepância e discrepância cruzada.

A natureza das proteínas, podem ter um delicado balanço entre Flexibilidade (**Frauenfelder**, H.(1989)) $\longleftrightarrow$ Rigidez conformacional, estando ambas num (para exercer uma dada função) (para sustentar especificidade funcional) estado estável termodinamicamente, como é o caso das enzimas (**Schulz**, G. E. & Schirmer, R. H. (1990)).

Experimentalmente o parâmetro de ordem S^2 e o fator B (ou temperatura), respectivamente, em espectroscopia "RMN" (**Sahu**, S.C. et al. (2000)) e cristalografia de Raio X (**Smith**, D.K. et al. (2003)), tem sido utilizados para identificar regiões flexíveis nas estruturas protéicas. Também há muitas proteínas intrinsecamente desordenadas (ou "IDP", iniciais do inglês *Intrinsically Disordered (or Unstructured) Proteins*), que no estado nativo estão totalmente ou parcialmente em desordem, existindo mais de uma centena de tais proteínas, como os Prions, Tao (**Linding**, R. et al. (2003)).

O conhecimento do quanto às estruturas das proteínas são flexíveis pode ser importante, entre outros, para o entendimento de suas funções, devido à relação estrutura/função, e também por que tais flexibilidades parecem ser diferentes para proteínas globulares e não globulares, conforme será evidenciado no Capítulo 3.

As proteínas utilizadas aqui, num total de 27 proteínas, tiveram suas estruturas, determinadas por ressonância magnética nuclear ("RMN" - em solução), a exceção de uma proteína determinada por Raio X. Como para cada proteína há mais de uma solução à sua estrutura (a exceção da obtida por Raio X de estrutura única), podemos medir o quanto tais estruturas são diferentes entre si, ou o quanto elas são flexíveis (Figura 6). O critério para a escolha destas proteínas, foi apenas a ocorrência exata de 35 resíduos em suas seqüências (Apêndice, após o Capítulo 5), assim sendo o mais geral possível, pois algumas proteínas possuem pontes dissulfeto (condição não analisada atualmente), outras contêm grupos prostéticos e outras são subunidades e outras não possuem nenhum destes três atributos.



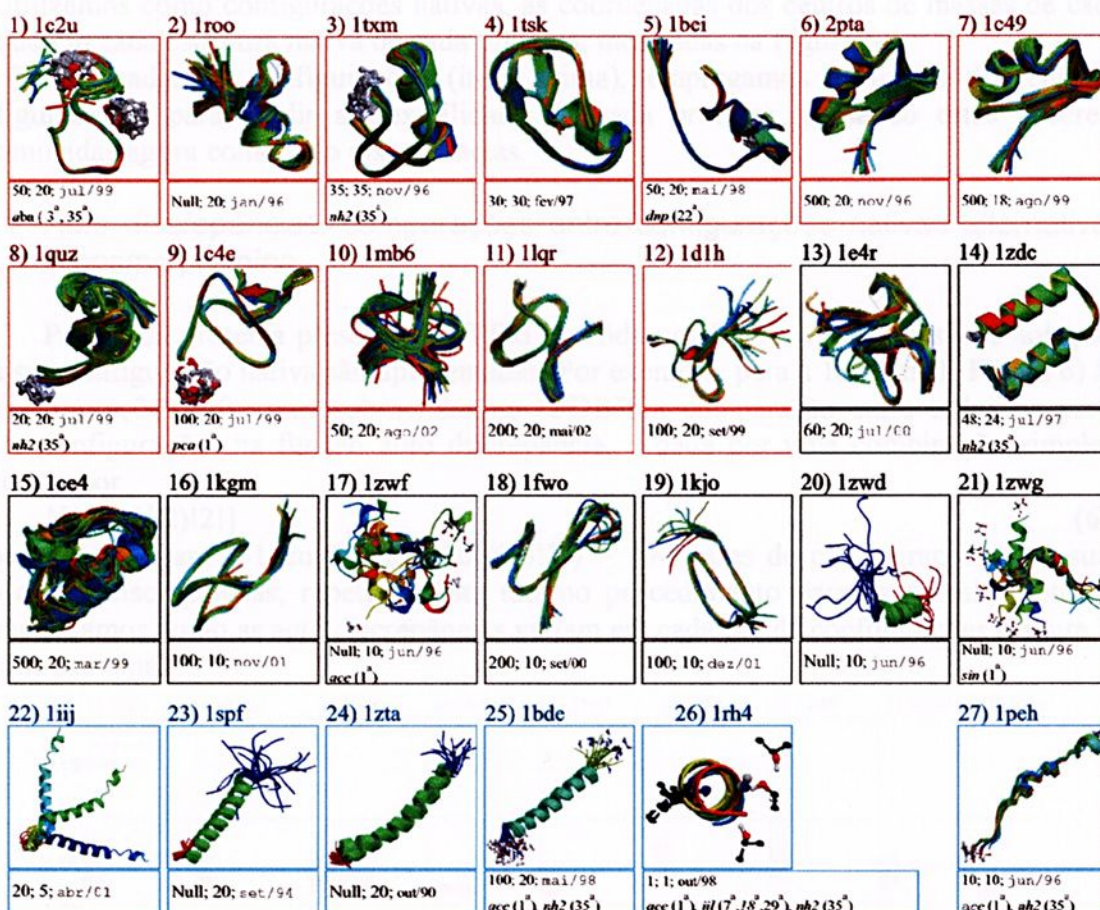


Figura 6 - “Flexibilidade estrutural” das 27 proteínas (dispostas na forma de fitas, representação *Molscript*), de 35 resíduos, organizadas sequencialmente da mais compacta, nº 1, 1c2u, até a menos compacta, nº 27, 1peh. Em cada proteína, quanto mais difusas as configurações (ou estruturas, ou *conformers*), maior a incerteza sobre sua conformação. As cores vermelhas, pretas e azuis nos nomes das proteínas são referentes as proteínas enquadradas por faixa de compactação (detalhes no Capítulo 3), respectivamente com altas, intermediárias e baixas compactações. Na 1ª linha abaixo de cada proteína consta, respectivamente, o número de estruturas calculadas e submetidas (número de diferentes fitas em cada proteína acima) por RMN e a data de depósito (ou processamento) no banco de dados de proteínas, “PDB”. Quando uma proteína possui hetero grupos, na 2ª linha, consta os nomes destes grupos e suas posições na sequência da proteína, podendo muitos destes serem identificados, nas estruturas como esferas conectadas.

Obs. A proteína nº 26 (1rh4) é a única que possui apenas uma estrutura calculada e submetida (um cristal), pois esta proteína foi obtida por difração de Raio X. O valor “Null” para as estruturas calculadas de algumas proteínas, é devido tal informação não ter sido fornecida pelo depositante.

Como a flexibilidade estrutural é uma das bases do nosso modelo (detalhado no início do Capítulo 3), pois estamos também estudando o mecanismo que induz a cadeia à sua configuração nativa e um de nossos principais interesses é a comparação entre configurações, surge naturalmente um questionamento sobre a variabilidade das configurações das proteínas:

A que nível a estrutura nativa de uma proteína pode variar ?

Esta questão pode ser abordada graças aos dados disponíveis no “PDB” e para observarmos o quanto as proteínas analisadas são flexíveis,

- (i) Utilizamos como configurações nativas, as coordenadas dos centros de massas de cada resíduo em cada estrutura nativa de cada proteína, mostradas na Figura 6;
- (ii) Determinadas as configurações (item acima), empregamos a função discrepância configuracional para medir a flexibilidade de cada proteína, passando estas a serem denominadas agora como auto discrepâncias.

2.4.1 Auto discrepâncias: comparações entre configurações nativas alternativas de uma mesma proteína

Para cada proteína presente no “PDB”, obtida por RMN, várias possíveis soluções para sua configuração nativa são apresentadas. Por exemplo, para a 1c2u (nº 1, Figura 6) há 20 configurações diferentes submetidas ao “PDB”; a comparação entre todos os pares destas configurações na função auto discrepância, é dada por uma combinação simples, fornecida por

$$N! / [(N-2)!2!] \tag{6}$$

, por exemplo, para a 1c2u teremos $20!/(18!2!) = 190$ pares de configurações com suas respectivas discrepâncias; repetindo este mesmo procedimento para as demais proteínas acompanhamos como as auto discrepâncias variam em cada par de configurações (Figura 7) de cada proteína.

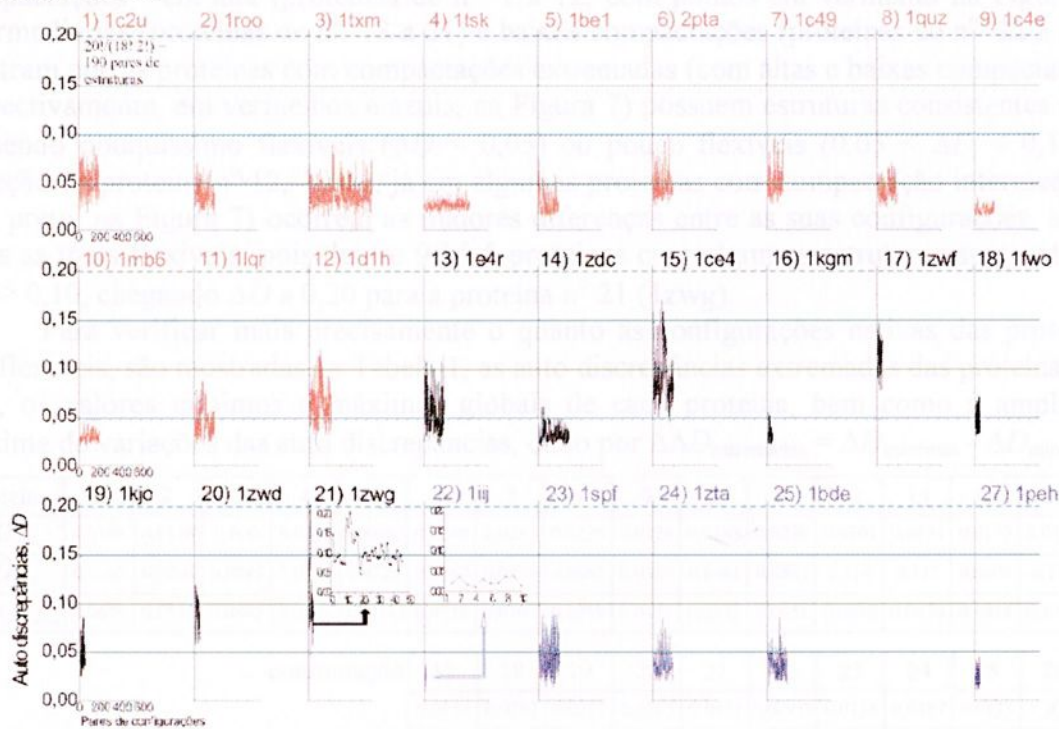


Figura 7 - Auto discrepâncias entre pares de configurações nativas de cada proteína - Note que os resultados acima estão em sintonia com os mostrados na Figura 6, isto é, quanto mais consistentes forem as diversas configurações, menor a discrepância entre elas. As cores vermelhas, pretas e azuis nos gráficos acima seguem o critério de compactação das proteínas (já comentado na legenda da Figura 6).

Obs. - A proteína nº 21, com poucas estruturas, no total de 10, foi destacada no gráfico pequeno devido sua maior variação de ΔD , que chega a 0,20; já a proteína nº 22, também devido suas poucas estruturas, no total de 5, tem poucos valores de ΔD , no total de 10 (ou seja, combinação simples de 5 estruturas agrupadas de duas em duas), daí a necessidade do gráfico pequeno, mostrando as variações de suas ΔD .

- A proteína nº 26 tem estrutura nativa única (obtida por cristalografia de Raio X), assim não possuindo auto discrepâncias ΔD , daí a ausência de seu gráfico.

Pelas variações das auto discrepâncias, ΔD , entre os pares de configurações, mostrados na Figura 7, observamos o quanto as estruturas nativas das proteínas, podem variar, como evidenciado por suas flexibilidades, sendo a maior flexibilidade para a proteína nº 21, 1zww, devido sua maior variação de ΔD , com máxima em 0,20 (destacada num gráfico pequeno na Figura 7). A proteína nº 22, 1lij (também destacada por um gráfico pequeno na Figura 7), tem auto discrepâncias baixas, apesar de visualmente (na Figura 6) suas estruturas serem muito díspares como decorrências destas estruturas estarem giradas uma em relação as outras, ou devido a existência de alguns de seus resíduos formando um segmento em estrutura tipo *loop* e *turn*, assim gerando a impressão visual de estruturas muito diferentes (como nas estruturas toda retilínea e as retilíneas inclinadas ao meio, Figura 3), portanto este é mais um exemplo da importância de se quantificar semelhanças configuracionais, como feito por ΔD e não apenas visualizá-las espacialmente.

As auto discrepâncias ΔD , mostram que algumas proteínas têm, em solução, estruturas bem definidas, enquanto outras apresentam grande variabilidade conformacional, e em apenas duas estruturas da proteína nº 3, 1txm, tivemos $\Delta D = 0,0$ (caracterizando configuração nativa única). Tais resultados mostram quantitativamente o observado para as flexibilidades conformacionais, mostrados na Figura 6.

As flexibilidades das configurações nativas das proteínas observadas quanto a compactações - em alta (proteínas de nºs 1 a 12, com pontos em vermelho na Figura 7), intermediárias (proteínas de nºs 13 a 21) e baixas compactações (proteínas de nºs 22 a 27) - mostram que as proteínas com compactações extremadas (com altas e baixas compactações, respectivamente, em vermelhos e azuis, na Figura 7) possuem estruturas consistentes entre si, sendo pouquíssimo flexíveis ($\Delta D < 0,05$) ou pouco flexíveis ($0,05 < \Delta D < 0,10$), a exceção da proteína nº 12, 1dlh, já em algumas proteínas com compactação intermediária (em preto, na Figura 7) ocorrem as maiores diferenças entre as suas configurações, sendo estas as mais flexíveis, pois dentre 9 há 5 proteínas com algumas estruturas apresentando $\Delta D > 0,10$, chegando ΔD a 0,20 para a proteína nº 21 (1zww).

Para verificar mais precisamente o quanto as configurações nativas das proteínas são flexíveis, são mostradas na Tabela 1, as auto discrepâncias extremadas das proteínas, ou seja, os valores mínimos e máximos globais de cada proteína, bem como a amplitude máxima de variações das auto discrepâncias, dado por $\Delta\Delta D_{\text{extremadas}} = \Delta D_{\text{máximas}} - \Delta D_{\text{mínimas}}$.

Proteínas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\Delta D_{\text{mínG}}$	0,0186	0,0156	0,000	0,0151	0,00883	0,0249	0,0267	0,0230	0,0125	0,0188	0,0230	0,0201	0,0194	0,0192	0,0426	0,0222
$\Delta D_{\text{máxG}}$	0,0865	0,0701	0,0885	0,0457	0,0723	0,0987	0,0974	0,0685	0,0344	0,0461	0,0881	0,119	0,117	0,0606	0,170	0,0549
$\Delta\Delta D_{\text{extremadas}}$	0,0679	0,0545	0,0885	0,0306	0,06347	0,0738	0,0707	0,0219	0,0455	0,0273	0,0651	0,0989	0,0976	0,0414	0,1274	0,0327
... continuação																
	17	18	19	20	21	22	23	24	25	26	27					
	0,0758	0,0297	0,0257	0,0563	0,0451	0,00918	0,0128	0,0187	0,0135	×	0,0102					
	0,146	0,0712	0,0858	0,119	0,202	0,0364	0,0878	0,0835	0,0859	×	0,0433					
	0,0702	0,0415	0,0601	0,0627	0,1569	0,02722	0,075	0,0648	0,0724	×	0,0331					

Tabela 1 - Destacando as auto discrepâncias extremadas (mínimas e máximas globais, respectivamente, $\Delta D_{\text{mínG}}$ e $\Delta D_{\text{máxG}}$) e a amplitude máxima de variações das auto discrepâncias, $\Delta\Delta D_{\text{extremadas}} = \Delta D_{\text{máxG}} - \Delta D_{\text{mínG}}$, entre as estruturas de cada uma das 27 proteínas. Tanto os valores extremados, $\Delta D_{\text{mínG}}$ e $\Delta D_{\text{máxG}}$, quanto os demais valores das auto discrepâncias, para cada proteína, estão contidas nos gráficos da Figura 7.



Observando-se as amplitudes máximas de variações das auto discrepâncias (dada por $\Delta\Delta D_{\text{extremadas}}$, na Tabela 1) nota-se que no geral esta fica entre cinco e dez centésimos, com exceção de algumas proteínas, nas três regiões de compactação, pouquíssimas flexíveis com $\Delta\Delta D_{\text{extremadas}}$ abaixo de cinco centésimos, no outro extremo há duas proteínas (n^{os} 15 e 21) com $\Delta\Delta D_{\text{extremadas}}$ acima de dez centésimos, caracterizando altas flexibilidades estruturais.

A flexibilidade protéica, observada mais em algumas proteínas do que em outras (Figuras 6, 7 e Tabela 1) pode ser relacionada a diversos fatores, como (a) segmentos contendo quantidades de resíduos variáveis, em estruturas como os *loops* ou *coils*; (b) cadeias laterais de resíduos que não participem, em redes de ligações de hidrogênios com outros grupos (como moléculas do solvente ou outros resíduos), **Livesay, D.R. et al. (2004)**; (c) segmentos flexíveis nas extremidades Amina e Carboxila terminal (**Vihinen, M. et al. (1994)**), aliás estas extremidades podem ser regiões de desordens, ocasionando freqüentes dificuldades na expressão, purificação e cristalização das proteínas, **Livesay, D.R. et al. (2004)**.

Para exemplificar a flexibilidade das proteínas, medida por outro método, citamos **Smith, D.K. et al. (2003)**, que utilizando fatores B, da cristalografia de Raio X, mostram que de um conjunto de 290 proteínas não idênticas, há 92% (266) contendo um total de 1011 segmentos flexíveis, de comprimentos variáveis, com aproximadamente 61% destes segmentos contendo um ou dois resíduos, tendo o maior segmento flexível 53 resíduos na extremidade Carboxila terminal (proteína 1FUR). Ainda com este conjunto de proteínas, há 29% (84) com ambas as extremidades flexíveis, enquanto 75% (217) têm pelo menos uma das extremidades flexíveis. Ainda das 290 cadeias protéicas há em torno de 46% (132) cadeias contendo pelo menos uma extremidade ausente da densidade eletrônica, do Raio X.

Após observar o quanto as proteínas são flexíveis (Figuras 6, 7 e Tabela 1) surge outra questão:

Quais as semelhanças e diferenças configuracionais entre as estruturas nativas das diferentes proteínas ?

Antes de abordarmos esta questão faz-se necessário verificar a homologia entre cada par de proteínas (Figura 8), ou qual a quantidade de mesmos resíduos nas mesmas posições ao longo da sequência destas proteínas, pois é presumível a relação entre altas homologias e baixas discrepâncias entre proteínas, já que quanto maior a homologia entre duas proteínas provavelmente mais parecidas são elas, conseqüentemente menor sua discrepância ΔD .



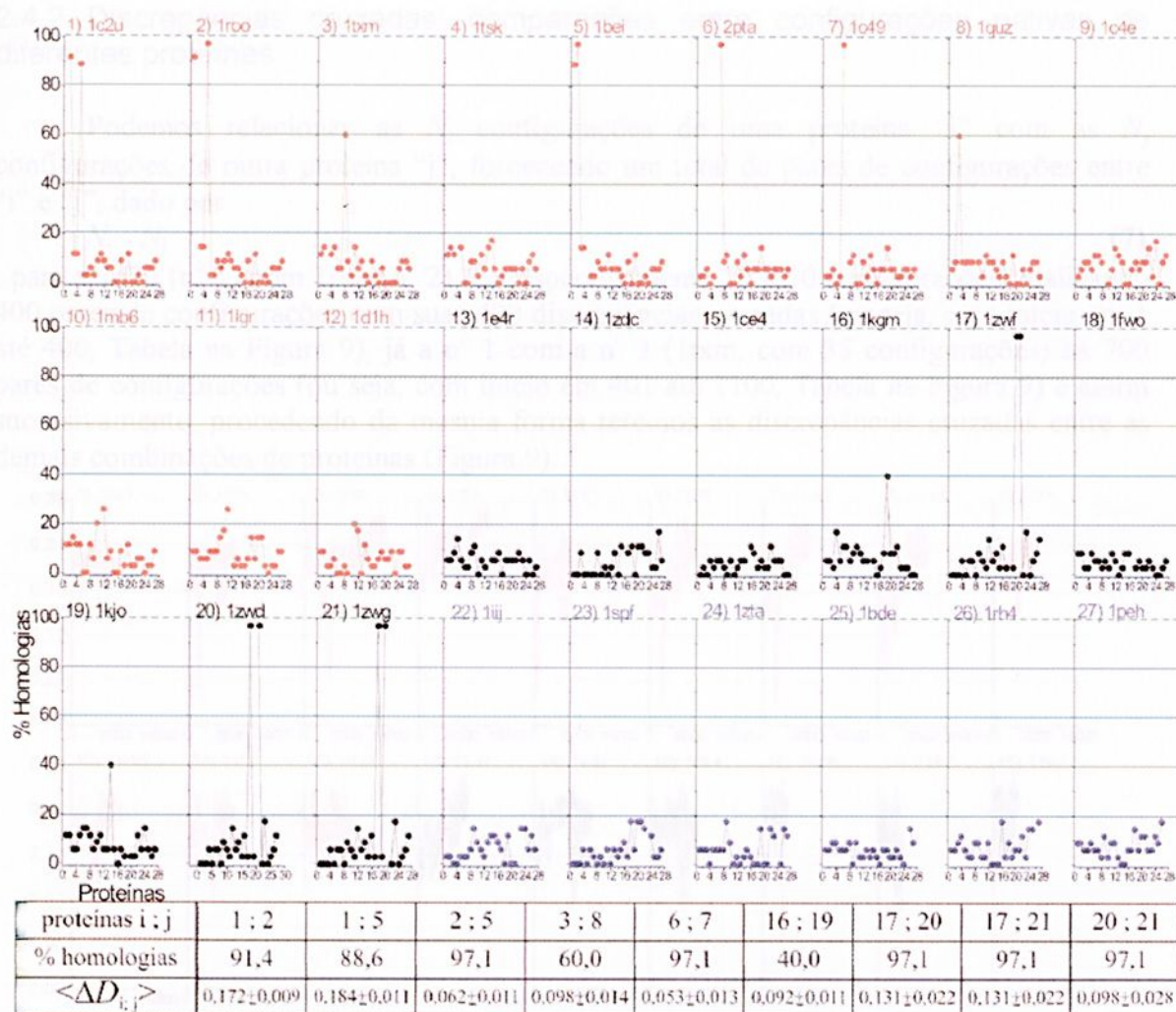


Figura 8 - Homologias percentuais entre as 27 proteínas (i ; j) de 35 resíduos - No geral as proteínas tem homologias abaixo, de 25%, ou seja, em suas seqüências há menos de 25% (ou de 6 dentre 35) dos mesmos resíduos coincidentes nas mesmas posições em cada par de proteína (i ; j). As proteínas com percentagem de homologias acima de 25% tem suas discrepâncias médias destacadas na tabela acima.

Obs. - Não são mostradas homologias entre uma mesma proteína (i ; i), que seria 100%, daí nestes casos a ausência de pontos nos gráficos acima, quebrando a seqüência de pontos conectados por retas.

Note que no geral as 27 proteínas são não idênticas, devido seus baixos níveis de identidades nas seqüências (com homologias abaixo de 25%, **Smith, D.K. et al. (2003)**), o que é desejável para verificar descorrelacionadamente semelhanças e diferenças configuracionais, entre suas estruturas nativas, conforme faremos a seguir. Note ainda que alta homologia não necessariamente leva a grande semelhança configuracional, pois algumas proteínas tem altas homologias (n^{os} 1 com 2 e 5; 17 com 20 e 21), mas suas discrepâncias médias (dos valores a serem mostrados na Figura 9 abaixo) não são tão baixas ($\Delta D > 0,10$, Tabela na Figura 8) quanto se poderia esperar.

Para observarmos as semelhanças e diferenças configuracionais entre as estruturas nativas do nosso conjunto de 27 proteínas não idênticas (última pergunta acima), empregamos a função discrepância inter proteínas, agora denominada discrepância cruzada.

2.4.2 Discrepâncias cruzadas: comparações entre configurações nativas de diferentes proteínas

Podemos relacionar as N_i configurações de uma proteína “i” com as N_j configurações de outra proteína “j”, fornecendo um total de pares de configurações entre “i” e “j”, dado por

$$N_i \times N_j \tag{7}$$

, para a 1c2u (nº 1) com 1roo (nº 2) há, respectivamente 20 e 20 configurações totalizando 400 pares de configurações com suas 400 discrepâncias cruzadas (ou seja, com início em 1 até 400, Tabela na Figura 9), já a nº 1 com a nº 3 (1txm, com 35 configurações) há 700 pares de configurações (ou seja, com início em 401 até 1100, Tabela na Figura 9) e assim sucessivamente; procedendo da mesma forma teremos as discrepâncias cruzadas entre as demais combinações de proteínas (Figura 9).

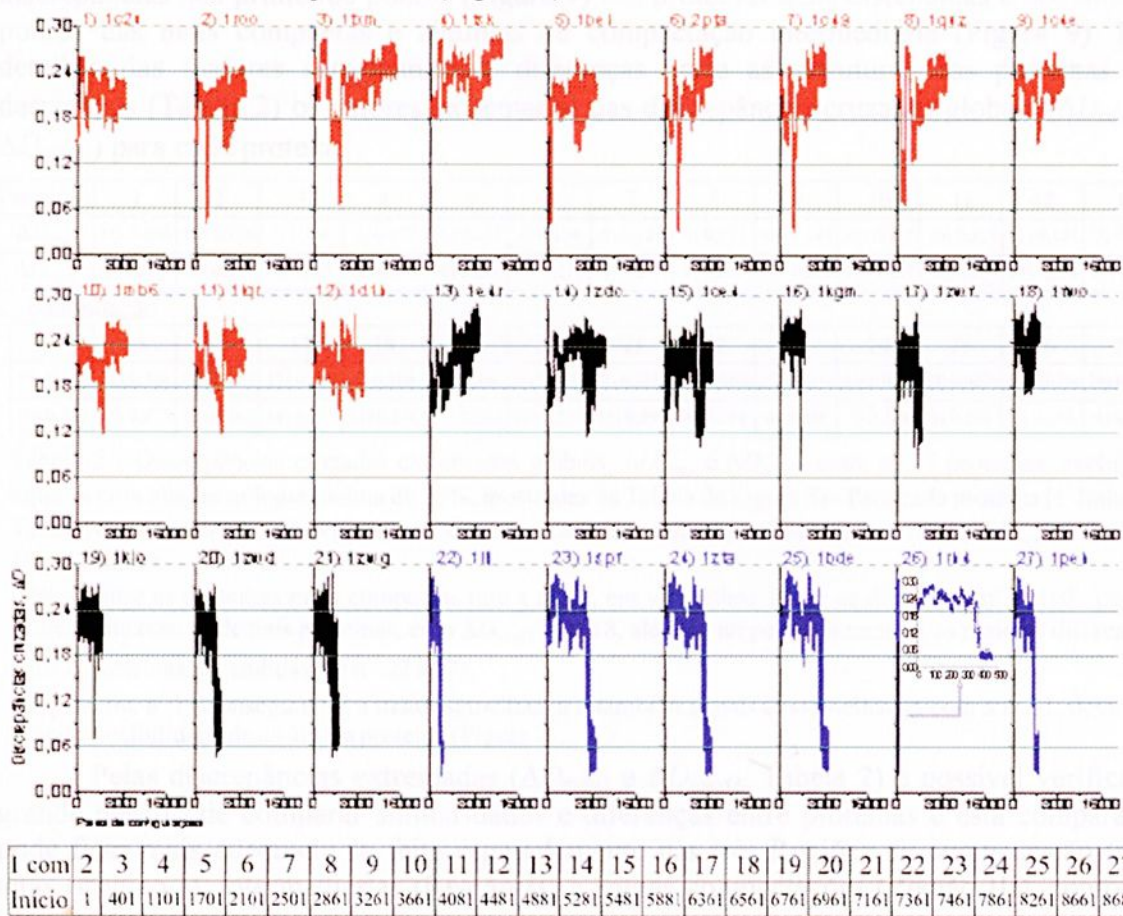


Figura 9 - Discrepâncias cruzadas, ΔD , entre pares de configurações nativas de diferentes proteínas - No geral, se forem excluídas as discrepâncias cruzadas das proteínas com alta homologias (Tabela na Figura 8), nota-se que as maiores semelhanças/diferenças configuracionais ocorrem com valores de ΔD , respectivamente, abaixo de 0,12 (ou 0,10) e acima de 0,26. Para dispor cada gráfico acima foi escolhido manter fixa uma configuração N_i e variar as outras N_j e calcular suas ΔD , tal escolha, as vezes, pode gerar alguns “degraus” nos valores de ΔD .

A proteína nº 26 tem estrutura nativa única (obtida por cristalografia de Raio X) e assim poucas ΔD cruzadas, no total 463, numa escala de 16.000, por isso suas discrepâncias cruzadas são destacadas num gráfico pequeno.

Na Tabela acima consta o início das quantidades de pares de estruturas da proteína nº 1 (1c2u) com as outras 26 proteínas, sendo tal procedimento utilizado para a construção dos gráficos acima.



Para as 27 proteínas analisadas, as maiores semelhanças e diferenças configuracionais ocorrem, com os menores e maiores valores de discrepâncias cruzadas, no geral estando estas, respectivamente, abaixo de 0,18 (e para alguns casos abaixo de 0,12 ou 0,10) e um pouco acima de 0,26. Quanto as semelhanças entre algumas proteínas, com ΔD abaixo de 0,18 há de se destacar dois casos distintos, sendo (i) algumas proteínas (as de n^{os} 1, 2, 3, 5, 6, 7, 8, 16, 17, 19, 20, 21) apresentam estes valores de ΔD decorrente da alta homologia com outras (ver Tabela da Figura 8); (ii) as duas últimas proteínas com compactação intermediária (de n^{os} 20 e 21) e as bastante distendidas a partir da n^o 22 possuem valores de discrepâncias abaixo de 0,12 devido as semelhanças entre proteínas também muito distendidas. Por outro lado as proteínas mais distendidas (a partir da n^o 20) possuem as maiores diferenças configuracionais (representada por $\Delta D \geq 0,26$) com as proteínas mais compactas e algumas de compactação intermediárias, daí os altos valores de discrepâncias nos primeiros pontos (Figura 9) das proteínas mais distendidas e nos últimos pontos das mais compactas e algumas de compactação intermediária (Figura 9). Para detalhes das maiores semelhanças e diferenças entre as estruturas das proteínas são destacados (Tabela 2) os valores extremados das discrepâncias cruzadas globais ($\Delta D_{\min G}$ e $\Delta D_{\max G}$) para cada proteína.

Proteínas	1	2	3	4	5	6	7	8	9	10	11	12	13
$\Delta D_{\min G}$	18) 0,160	15) 0,136	7) 0,146	2) 0,177	15) 0,137	8) 0,124	8) 0,128	6) 0,124	10) 0,147	11) 0,117	10) 0,117	11) 0,138	2) 0,142
$\Delta D_{\max G}$	3) 0,260	16) 0,254	4) 0,285	25) 0,292	16) 0,260	18) 0,276	9) 0,277	4) 0,276	21) 0,267	21) 0,267	16) 0,274	21) 0,274	25) 0,290

... continuação

14	15	16	17	18	19	20	21	22	23	24	25	26	27
15) 0,115	17) 0,101	18) 0,171	15) 0,101	1) 0,160	21) 0,161	25) 0,9547	25) 0,6554	23) 0,0216	22) 0,0216	23) 0,0218	24) 0,0275	24) 0,0256	26) 0,0267
16) 0,274	18) 0,277	21) 0,285	18) 0,278	21) 0,290	21) 0,273	18) 0,274	18) 0,290	4) 0,284	4) 0,286	4) 0,288	4) 0,292	4) 0,288	4) 0,284

Tabela 2 - Discrepâncias cruzadas extremadas globais, $\Delta D_{\min G}$ e $\Delta D_{\max G}$, entre as 27 proteínas, excluídas aquelas com alta homologia (acima de 25%, mostradas na Tabela da Figura 8) - Para cada proteína (1^a linha da Tabela) consta com que outras proteínas ocorrem, e os valores, das discrepâncias extremadas $\Delta D_{\min G}$ (2^a linha) e $\Delta D_{\max G}$ (3^a linha).

Obs. - Entre as proteínas mais compactas (até a n^o 12, em vermelho), há de se destacar a n^o 4, 1tsk, pouco semelhante com as demais proteínas, com $\Delta D_{\min G} \cong 0,18$, além de ter persistentemente as maiores diferenças com as proteínas distendidas (de n^{os} 22 a 27).

- A proteína n^o 19 conseguiu ter a maior semelhança e também a mais dessemelhança com a n^o 21, devido a grande flexibilidade desta última proteína (Figura 7).

Pelas discrepâncias extremadas ($\Delta D_{\min G}$ e $\Delta D_{\max G}$, Tabela 2) é possível verificar o grande desafio de comparar similaridades e diferenças entre proteínas e esta comparação pode ficar mais adequada se feita separadamente nas três Regiões de compactação (com altas (Região I), intermediárias (Região II) e baixas compactações (Região III), conforme legenda da Figura 6), sendo assim

- Região I (proteínas de n^{os} 1 a 12): As maiores semelhanças, dada por $\Delta D_{\min G}$ (entre 0,12 e 0,18), ocorrem entre as proteínas de compactação próximas, na mesma Região I, a exceção das proteínas 1, 2 e 5 mais semelhantes com proteínas de compactação intermediárias (Região II). Já as maiores diferenças, dadas por $\Delta D_{\max G}$ (de 0,26 e 0,29), ocorrem entre proteínas das três regiões de compactações, sendo este fato uma decorrência direta da alta complexidade estrutural observada nas proteínas compactas desta Região.

- Região II (proteínas de n^{os} 13 a 21): As maiores semelhanças ocorrem com as maiores variações de $\Delta D_{\min G}$ (de 0,050 (n^{os} 20 e 21 com 25) a 0,17 (n^{os} 16 com 18)) e com proteínas



das três regiões; já as maiores diferenças configuracionais ocorrem com outras proteínas da mesma Região de compactação (com exceção da proteína nº 13). Portanto, proteínas de compactações semelhantes não indicam maior probabilidade de grandes semelhanças, pois os específicos contatos intra cadeia dos resíduos podem ser bastante diferentes nestas proteínas, acarretando grandes diferenças configuracionais.

- Região III (proteínas de nº 22 a 27): Nestas proteínas distendidas, a presença de muitas α -hélices (resíduos próximos na sequência) diminuem o valor da discrepância, daí naturalmente as maiores semelhanças com outras proteínas da mesma região e por isso também as maiores diferenças com proteínas compactas (com grande composição de outras estruturas secundárias além de α -hélice), particularmente a nº 4, 1tsk.

As maiores semelhanças e diferenças configuracionais, das configurações nativas das proteínas, ocorrem com as menores e maiores discrepâncias cruzadas, isto pode ser confirmado e melhor entendido se forem visualizadas as estruturas das 27 proteínas pelo conjunto de todas as distâncias entre seus resíduos na proteína, dado pelos mapas de distâncias (Figura 10).

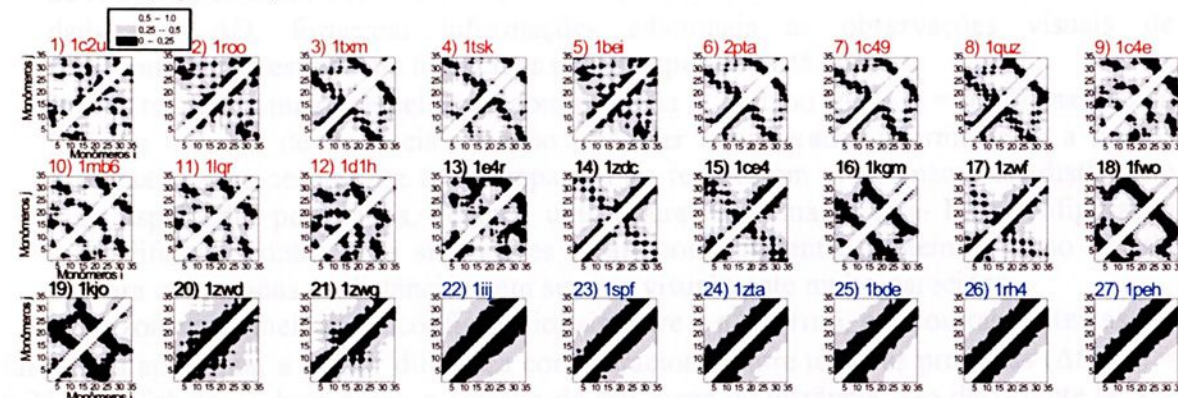


Figura 10 - Mapas de distâncias das 27 proteínas, referentes à primeira configuração nativa dentre as submetidas ao “PDB” (Figura 6) - As distâncias entre resíduos (centro de massa), mostradas nos mapas, estão normalizadas (equações 1) e mostradas em cores preta, cinza e branca, respectivamente para distâncias menores que 0,25, de 0,25 a 0,50 e as maiores que 0,50 nas proteínas. A cor branca ao longo da diagonal principal se deve a exclusão de distâncias entre mesmos resíduos (i com i) e primeiros e segundo vizinhos (i com $i-1$, $i-2$, $i+1$, $i+2$).

Sempre que for necessário visualizar configurações cujas (des)semelhanças sejam previstas pela função ΔD , recorreremos aos mapas de distâncias, pois com estes é possível, confirmar os resultados previstos pela função ΔD , como é o caso dos valores de discrepâncias das proteínas com altas homologias com outras proteínas (Figura 8) e discrepâncias cruzadas (Figura 9 e Tabela 2).

Obs.- Foi selecionada a primeira configuração de cada proteína, pois em muitas vezes (nos arquivos das proteínas no “PDB”) estas são consideradas as estruturas mais representativas das proteínas.

- Os mapas de distâncias são parecidos, contudo mais gerais, que os mapas de contatos (Figura 1)2), pois nos primeiros os resíduos não precisam estar em contato.

Os mapas de distâncias capturam a estrutura tridimensional das proteínas, ou de outras configurações quaisquer, e certos aspectos estruturais são visíveis, como as α -hélices (resíduos próximos na sequência) representadas por bandas largas (devido a aglomeração de vários pontos) pretas ao longo da diagonal principal, e folhas betas (resíduos distantes na sequência) mostrados por bandas largas pretas ao longo da diagonal secundária (ou perpendicular a diagonal principal), já pontos isolados podem conter informações de *loops* e *turns*, ou representar proximidade entre resíduos de Cisteínas em pontes dissulfeto. Mapas de distâncias parecidos, geralmente são referentes a proteínas também parecidas, por

exemplo, as proteínas com altas homologias (como as nº 1 com 2 e 5; 3 com 8, ver Tabela da Figura 8) e as proteínas com baixos valores de ΔD (e.g., os valores de $\Delta D_{\min G}$ na Tabela 2) apresentam mapas de distâncias semelhantes. Por outro lado, grandes valores de discrepâncias (como os dados por $\Delta D_{\max G}$ na Tabela 2), tendem a terem mapas de distância muito diferentes, normalmente complementares, pois, nestes casos, grandes distâncias (em branco) entre resíduos num mapa têm pequenas distâncias (em preto) no outro mapa e vice-versa.

Observando os mapas de distâncias (Figura 10) da proteína nº 4, 1tsk, e que apresenta a menor semelhança entre todas as proteínas ($\Delta D_{\min G} = 0,177$, na Tabela 2) pode-se entender um pouco mais das informações fornecidas pela função ΔD , também aplicáveis à outras proteínas, assim

(i) entre a 1tsk e a nº 7, 1c49, nota-se muitas semelhanças no conjunto das distâncias, porém com regiões em preto deslocadas nos mapas – com as pequenas distâncias da 1c49 abaixo das mesmas regiões na 1tsk – levando a função ΔD a assumir grandes valores (no caso $\Delta D = 0,231$), pois esta função leva em conta as distância entre cada par de resíduos, conseqüentemente as quantificações de (des)semelhanças configuracionais, dada por ΔD , fornecem informações adicionais as observações visuais de (des)semelhança estruturais fornecidas pelos mapas de distâncias;

(ii) a 1tsk tem a maior semelhança com proteína nº 2, 1roo ($\Delta D_{\min G} = 0,177$, na Tabela 2), pois o mapa de distância da 1roo pode ser considerado, intermediário a muitas proteínas, como se percebe ao acompanhar as regiões em preto (pequenas distâncias) mais espalhadas pelo mapa, gerando uma figura na forma de seta. Esta configuração mais difusa encontra mais similitudes locais com diferentes proteínas, como a 1tsk, embora com mapas de distâncias nem sempre visualmente muito parecidos.

As razoáveis semelhanças configuracionais entre a nº 2, 1roo, com outras proteínas e o fato desta apresentar a menor diferença configuracional entre todas as proteínas ($\Delta D_{\max G} = 0,254$, na Tabela 2), bem como o aspecto de seu mapa de distância, são decorrente de sua estrutura secundária apresentar muitas voltas ou *turns*; além da presença de muitas alfa-hélices (resíduos próximos na sequência) que diminuem o valor da discrepância, pois as distâncias nestes casos são naturalmente próximas levando a discrepâncias locais baixas.

Após realizarmos comparações estruturais intra e inter proteínas, com a função ΔD , a aplicaremos ao estudo do problema do *folding* protéico no Capítulo 3, para as cadeias modelo (obtidas durante a simulação) com as proteínas (cujos mapas de distâncias estão na Figura 10), isso será realizado pelas discrepâncias cruzadas modelos e proteínas.

A função discrepância ΔD , tem sido empregada, em diversas comparações configuracionais, sendo algumas destas apresentadas neste Capítulo e outras nos próximos dois capítulos, mas já é possível destacar as características mais importantes desta função.

2.5 A função discrepância configuracional: síntese de propriedades

Das definições (seção 2.2) e aplicações da função ΔD em quantificar semelhanças e dessemelhanças entre estruturas (seções 2.3 e 2.4) podem ser destacadas as principais características desta função ΔD , sendo estas

— Característica principal: Para efetuarmos a comparação entre duas configurações quaisquer utilizando ΔD , devemos antes normalizar, equação 1, todas as distâncias envolvidas e este procedimento torna ΔD invariável por escala, translações e rotações (Figura 11) entre as configurações comparadas, podendo esta ser considerada a principal característica de ΔD .

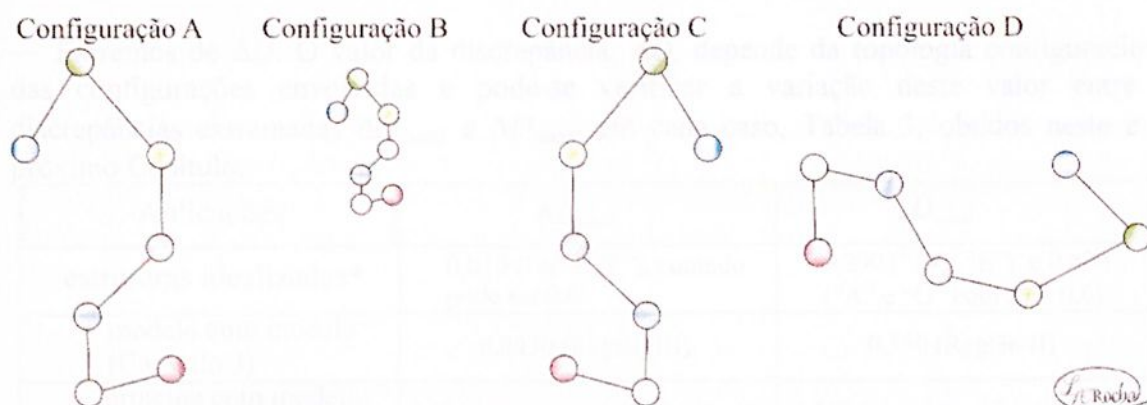


Figura 11 - A função discrepância, ΔD , é invariante ao tamanho ("escala"), translação e rotação, conseqüentemente assumindo o mesmo valor, $\Delta D = 0,0$, entre a configuração A com B, C e D.

— Nível de Detalhe: A função ΔD é razoavelmente rigorosa em comparações configuracionais, pois configurações com mapas de distâncias parecidos, mas deslocados nos contatos entre pares de resíduos (como nas proteínas 1tsk e 1c49, na Figura 10), levam à grandes discrepâncias; este fato será aproveitado nos resultados de discrepância cruzada proteínas com modelo, mostrando a eficiência de nosso potencial hidrofóbico utilizado na simulação. Contudo normalmente grandes ΔD , são devido a grandes diferenças entre configurações, como pode ser verificado confrontando os valores de $\Delta D_{\max G}$, na Tabela 2, com os mapas de distâncias da Figura 10.

— Invariância: A função $\Delta D_{\text{extremada}}$ (Tabela 2) não é invariante para proteínas p e q , como pode ser verificado para muitas proteínas, por exemplo, a proteína nº 2 tem a máxima semelhança (expressa por $\Delta D_{\min G}$) com a proteína nº 15, mas a máxima semelhança da proteínas nº 15 é com a nº 18, o mesmo ocorrendo para $\Delta D_{\max G}$ da proteína nº 2. Tal invariância de $\Delta D_{\text{extremada}}$ é grandemente decorrente da existência de diversas configurações alternativas (ou *conformers*) representativas da conformação nativa de cada proteína, como fica evidente nos valores extremados das auto discrepâncias (Figura 7).

— Cautela: A função ΔD é muito sensível, quando comparada com valores limiares " α " (no Capítulo 3), por isso é preciso muita cautela no seu emprego, nestes casos, conforme mostraremos nas discrepâncias cruzadas das proteínas com as configurações modelo. Tal valor de α está estreitamente relacionado a $\Delta D_{\min G}$ (Tabela 2) que também pode ser tomado como referência para nossas simulações realizadas nos próximos Capítulos.

— Limitação: Os resultados fornecidos por ΔD , por influência direta do processo de normalização de distâncias empregado em ΔD podem não ser condizentes com o esperado para a semelhança entre duas estruturas, se estas tiverem as distâncias entre suas unidades consecutivas muito incompatíveis entre si. Tal fato ocorreu na comparação entre a estrutura

idealizadas retilínea e hélices quadradas comprimidas ($0,8 \leq \delta / \leq 0,1$, Figura 5), contudo este não é o caso das comparações entre proteínas com proteínas e das proteínas com modelo, pois os tamanhos dos resíduos, nestes casos, são razoavelmente equivalentes, tornando as distâncias entre os resíduos consecutivos compatíveis, ao contrário das estruturas idealizadas cujos resíduos são pontuais.

— Extremos de ΔD : O valor da discrepância, ΔD , depende da topologia configuracional das configurações envolvidas e pode-se verificar a variação deste valor entre as discrepâncias extremadas $\Delta D_{\min G}$ e $\Delta D_{\max G}$ em cada caso, Tabela 3, obtidos neste e no próximo Capítulo.

Aplicações	$\Delta D_{\min G}$	$\Delta D_{\max G}$
estruturas idealizadas*	0,010 ("A" e "F"), contudo pode ser 0,0	0,290 ("A" / "E") e 0,430 ("A" e "G" com $\delta / = 0,0$)
modelo com modelo (Capítulo 3)	0,0430 (Região III)	0,350 (Região II)
proteína com modelo (Capítulo 3)	0,0320 (proteína n° 22, Região III)	0,360 (proteína n° 24, Região III)
proteína A com proteína B	0,0216 (proteínas n°s 22 e 23)	0,292 (proteínas n°s 4 e 25)
proteína A com proteína A	0,000 (proteína n° 3 / n° 3) e 0,00883 (proteína n° 5 / n° 5)	0,202 (proteína n° 21 / n° 21)

* Ver Figura 3 para referenciar-se aos símbolos A, B, ... G.

Tabela 3 - Diversas aplicações e seus respectivos valores de funções discrepâncias configuracionais extremadas globais, mínimas globais ($\Delta D_{\min G}$) e máximas globais ($\Delta D_{\max G}$).

Observando as diversas aplicações com ΔD nota-se as maiores semelhanças, dadas por $\Delta D_{\min G}$, em discrepâncias cruzadas, modelo com modelo e proteína com modelo (Capítulo 3) são compatíveis com as discrepâncias cruzadas proteína A com proteína B. Assim como há $\Delta D_{\min G}$ próxima entre estruturas idealizadas e proteína A com proteína A. As maiores diferenças estruturais ocorrem com valores de discrepâncias $\Delta D_{\max}$, entre 0,20 e 0,30 embora haja casos excepcionais como os evidenciados por $\Delta D_{\max} \cong 0,35$ e até de 0,43 $\Delta D_{\max G}^{A/G}$.

— Máximo de ΔD : A função discrepância $\Delta D_{\max G}$ pode assumir hipoteticamente o valor máximo 1,0, bastando para isso que duas estruturas sejam extremamente diferentes. Na prática tal fato somente é constatado se for reduzindo as quantidades de M distâncias distintas entre pares de monômeros, de várias centenas (no caso tivemos $M = 528$ (Equação 5)) para abaixo de uma dezena; adicionalmente devendo necessariamente tais distâncias entre os mesmos pares de resíduos nas duas configurações serem muito diferentes, por isso a 3ª coluna na Tabela 3 tem $\Delta D_{\max G}$ chegando no máximo a apenas 0,43.

— Confirmando resultados previstos por ΔD : Os resultados das discrepâncias, ΔD , podem ser confirmados pelos mapas de distâncias (Figura 10), sendo este um procedimento adotado por nós, principalmente quando de boas previsões de similaridades feitas por ΔD .



Da quantificação entre estruturas idealizadas e entre estruturas intra e inter proteínas, mantendo o tamanho destas estruturas fixo em 35 monômeros, já é possível destacar alguns comentários finais referentes à própria função discrepância configuracional, bem como de sua versatilidade (e conceitualmente simples), em realizar tais quantificações.

2.6 Comentários

A função discrepância configuracional foi, neste Capítulo, definida (seção 2.2), aplicada a duas situações bastante gerais (seção 2.3 e 2.4) e caracterizada, destacando suas principais limitações e aplicações (seção 2.4), mostrando-se adequada, para nossas necessidades atuais, para comparação configuracional, como continuará a ser feito no Capítulo 3.

As comparações entre configurações, por meio da função discrepância ΔD também evidenciam estratégias e dificuldades de forma similares a vários outros métodos utilizados em programas bastante sofisticados como LINUS (Srinivasan, R. & Rose, G. D. (1995)), FIRST (Jacobs, D.J. et al. (2001)), FATCAT (Ye, Y. & Godzik, A. (2004)) dentre outros, sendo que cada um destes apresentam uma melhor performance para casos específicos, além de possuírem limitações quanto à aplicações.

As auto discrepâncias para proteínas de diferentes compactações, Figura 7, confirmam a constatação de que muitas funções importantes de proteínas grandes ocorrem fora do domínio globular, tendo os subespaços globulares (mais compactos) e não globulares (menos compactos) diferentes flexibilidades (Linding, R. et al. (2003) ; Wright, P. & Dyson, H. (1999) ; Dunker, A. et al. (2002)).

A identificação de segmentos altamente móveis (em desordem ou desestruturados), dentre outros fatores, é essencial para comparações adequadas entre diferentes cadeias protéicas (Smith, D.K. et al. (2003)), com diversas aplicações (ver itens i a vi no 2º parágrafo da seção 2.1). Já a previsão destes segmentos pode acarretar aumento da expressão, foldabilidade e estabilidade de uma proteína (Linding, R. et al. (2003)). Em nosso caso, como a flexibilidade estrutural é uma das bases do nosso modelo (representada pela separação variável entre os monômeros consecutivos na cadeia, conforme detalhado no início do Capítulo 3), então surge a necessidade de quantificar configurações e de se conhecer em detalhes a função ΔD desenvolvida para este fim, daí suas variadas aplicações, como nas estruturas idealizadas e das configurações intra e inter proteínas, inclusive por que tais quantificações poderão servir de referenciais para resultados das discrepâncias cruzadas entre estruturas proteína com modelo (geradas, ao longo da simulação Monte Carlo), no Capítulo 3.



2.7 Bibliografia

- Brooks, B. R.,** Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations *J. Comp. Chem.* **4**, 187–217 (1983)
- Dunker, A.,** Brown, C., Lawson, J., Iakoucheva, L., Obradovic, Z., Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582 (2002).
- Frauenfelder, H.,** New looks at protein motions, *Nature* **338**, 623–624 (1989).
- Gnanakaran, S. & Garcia, A. E.,** Validation of an All-Atom Protein Force Field: From Dipeptides to Larger Peptides, *J. Phys. Chem. B* **107**, 12555–12557 (2003)
- Jacobs, D.J.,** Rader, A.J., Kuhn, L.A., Thorpe, M.F., Protein Flexibility Predictions Using Graph Theory, *PROTEINS: Struct. Funct. and Genet.* **44**, 150–165 (2001)).
- Koehl, P.,** Protein structure similarities, *Current Opinion in Structural Biology* **11**, 348–353 (2001).
- Kolodny, R.,** Koehl, P., Levitt, M., Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures, *J. Mol. Biol.* **346**, 1173–1188 (2005).
- Linding, R.,** Jensen, L. J., Diella F., Bork, P., Gibson, T. J., Russell, R. B., Protein Disorder Prediction: Implications for Structural Proteomics, *Structure* **11**, 1453–1459 (2003).
- Livesay, D.R.,** Dallakyan, S., Wood, G.G., Jacobs, D.J., An flexible approach for understanding protein stability, *FEBS Letters* **576**, 468–476 (2004).
- Pelta, D.,** Krasnogor, N., Bousono-Calzon, C., Verdegay, J. L., Hirst, J., Burke, E., A fuzzy sets based generalization of contact maps for the overlap of protein structures, *Fuzzy Sets and Systems* **152**, 103–123 (2005).
- Perutz, M. F.,** Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T., Structure of myoglobin: a three-dimensional Fourier synthesis at 5.5 Angstrom resolution, obtained by X-ray analysis, *Nature*, **185**, 416–422 (1960).
- Rocha, L. F. O.,** da Silva, M. A. A., Caliri A., Entropic force and folding of macromolecules, *Phys. Lett. A* **220**, 178–182 (1996).
- Sahu, S.C.,** Bhuyan, A.K., Majumdar, A., Udgaonkar, J.B., Backbone Dynamics of Barstar: A ¹⁵N NMR Relaxation Study, *PROTEINS: Structure, Function, and Genetics* **41**, 460–474 (2000).
- Schulz, G. E. & Schirmer, R. H.** in: *Principles of Protein Structure*, Springer-Verlag, Chapter 11 (1990).
- Smith, D.K.,** Radivojac, P., Obradovic, Z., Dunker, A. K., Zhu, G., Improved amino acid flexibility parameters, *Protein Science*, **12**, 1060–1072 (2003).
- Srinivasan, R. & Rose, G. D.,** LINUS: A Hierarchic Procedure to Predict the Fold of a Protein, *PROTEINS: Struct. Funct. and Genet.* **22**, 81–99 (1995).
- Taylor, W.R.,** Orengo, C.A. Protein structure alignment. *J Mol Biol* **208**, 1–22 (1989).
- Thomas, P.D.,** Dill, K.A., Statistical Potentials Extracted From Protein Structures: How Accurate Are They ?, *J. Mol. Biol.* **257**, 457–469 (1996).
- Vihinen, M.,** Torkkila, E., Riikonen, P., Accuracy of protein flexibility predictions. *Proteins* **19**: 141–149 (1994).
- Wright, P., & Dyson, H.,** Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).



3.1 Sistema proteína-solvente: Especificação do modelo



CAPÍTULO 3

Parametrização da cadeia protéica do modelo e o fator aleatoriedade no processo de *folding* de proteínas

O sistema empregado na simulação consiste de uma cadeia de tamanho fixo imersa no solvente, imitando uma solução com certa concentração de macromoléculas. No presente caso, a fração volumétrica ρ de ocupação do sistema, pelos monômeros da cadeia e moléculas do solvente, é $\rho = 0,4$. Em cada movimento das unidades, tanto do solvente como da cadeia, a amplitude máxima $\delta_{\max}$ de movimentos escolhida foi de $\delta_{\max} = 0,3$ (unidades arbitrárias; ver item abaixo), o que juntamente com o valor de ρ adotado garante que a relação entre movimentos aceitos e rejeitados seja estabelecida como sendo cerca de 50% para as partículas componentes do solvente, no regime de temperatura infinita, $T \rightarrow \infty$.

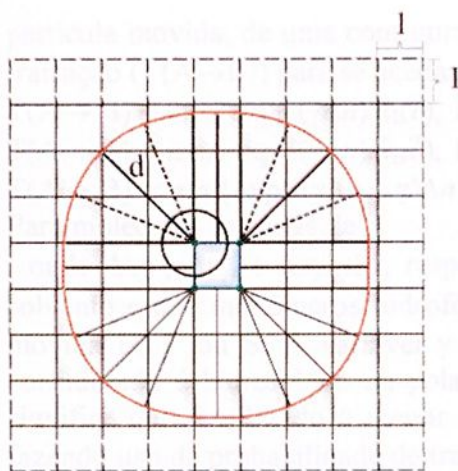
3.1 Sistema proteína-solvente: Especificação do modelo

O modelo tratado no presente trabalho consiste basicamente de um sistema formado por uma cadeia protéica de 35 unidades (também denominados aqui por contas, resíduos ou monômeros) imersa em solução aquosa contendo 12.565 moléculas do solvente, estando todas as partículas confinadas numa caixa cúbica de lado 49 (unidades arbitrárias) com condições periódicas de contorno (**Hansen, J. P. & McDonald, I. R. (1976); Metropolis, N. et al. (1953)**). A caixa cúbica é subdividida em pequenas células cúbicas unitárias $1 \times 1 \times 1$. Assim, toda medida de extensão usada neste trabalho guarda uma razão de proporcionalidade a esta unidade arbitrária.

Tanto as esferas do solvente quanto às contas da cadeia fitam a diagonal do cubo referente às células $1 \times 1 \times 1$, isto é, ambas são do mesmo tamanho, possuindo diâmetro $D = 3^{1/2}$. Note que o presente modelo é definido no espaço 3D contínuo (ou fora da rede), consequentemente a subdivisão do sistema em células unitária é feita somente com finalidade computacional, como contar exatamente o número total de vizinhos para cada partícula, ou seja, quais são os vizinhos e as distâncias destes a uma determinada partícula.

Duas partículas quaisquer do sistema são consideradas vizinhas, ou em contato topológico, quando a distância r entre seus centros forem $D < r \leq d$, com a distância de corte fixa em $d = 8^{1/2} \cong 1,63D$ (Figura 1). Para fins de simulação o sistema composto, proteína-solvente, é considerado sob o ponto de vista de um sistema microcanônico, portanto tem energia constante (detalhes na seção 4.4, do Capítulo 4).





Determinando o diâmetro D das partículas

$$D = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$$

$$D = 1,73$$

Determinando a distância máxima d

$$d^2 = 2^2 + 2^2$$

$$d^2 = 8$$

$$d = 2,83$$

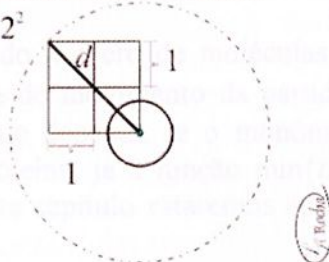


Figura 1 - Diâmetro dos monômeros D e distância de corte d - As partículas, com centro na região central (quadrado central em azul e branco), incluindo o contorno, está em contato topológico com qualquer outra dentro da "esfera" maior (em vermelho) até a distância máxima (d), com $d = 2,83$. Todas as partículas do sistema cadeia-solvente são esféricas e de mesmo tamanho, com diâmetro $D = 1,73$.

Cada unidade i da cadeia está conectada a outras duas unidades consecutivas $i+1$ e $i-1$ por ligações idealmente flexíveis de comprimento $\ell = D + \varepsilon$. A distância r entre os centros de duas unidades consecutivas ao longo da cadeia é governada por um potencial de pares $U_\ell(r)$, definida como $U_\ell(r) = 0$ para $D \leq r \leq \ell$ e $U_\ell(r) = \infty$ caso contrário (parte superior da Figura 2a). As moléculas S do solvente são também representadas por esferas rígidas (ou esferas duras) todas com diâmetro D , portanto para duas partículas i e j separadas por uma distância r_{ij} a interação solvente-solvente e solvente-cadeia é do tipo esfera rígida, sendo a energia potencial $U(r_{ij}) \rightarrow \infty$ para $r_{ij} < D$ e $U(r_{ij}) = 0$ caso contrário (parte inferior da Figura 2a). Cada unidade da cadeia é uma esfera rígida de diâmetro D podendo ter os atributos químicos de um elemento hidrofóbico H ou polar P (Figura 2b).

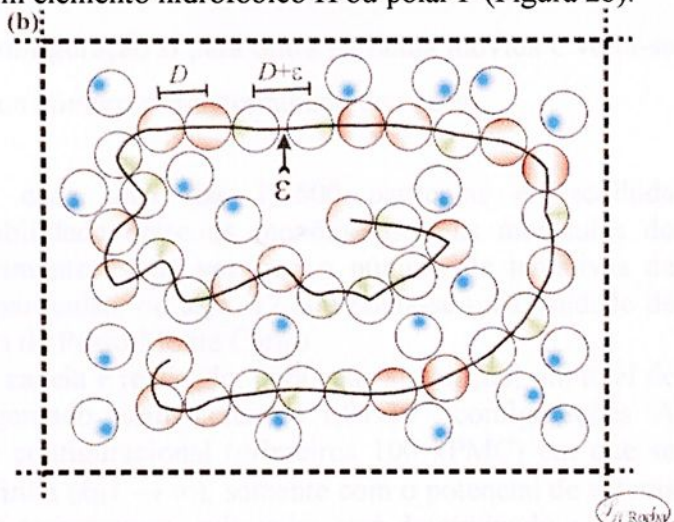
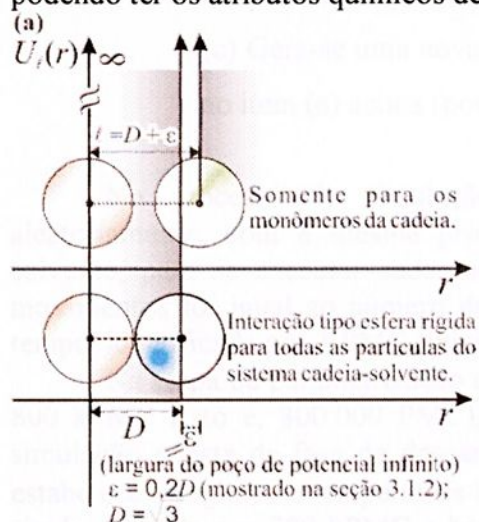


Figura 2- a) Potencial $U_\ell(r)$, entre pares, para partículas do sistema cadeia-solvente; b) Sistema composto pela cadeia (formada por monômeros H (em vermelho) e P (em verde)) e solvente (S (moléculas em azul)).

A interação entre os monômeros da cadeia e as moléculas do solvente depende da área superficial acessível de cada monômero às moléculas do solvente e dependendo da

partícula movida, de uma configuração A para outra configuração B, a probabilidade

transição ($T(A \rightarrow B)$) para se aceitar, ou não, a nova configuração B, será da seguinte forma

$$T(A \rightarrow B) = \min\{\exp(-(\gamma\Delta n)/k_B T), 1\}, \text{ com } \gamma=2k_B T \text{ e } \Delta n=\Delta n_S \quad \text{Para monômeros H} \quad (1)$$

$$T(A \rightarrow B) = \min\{\exp(-(\gamma\Delta n)/k_B T), 1\}, \text{ com } \gamma=-1k_B T \text{ e } \Delta n=\Delta n_S \quad \text{Para monômeros P} \quad (2)$$

$$T(A \rightarrow B) = \min\{\exp(-(\gamma\Delta n + \gamma'\Delta n')/k_B T), 1\}, \text{ com } \gamma=2k_B T, \Delta n=\Delta n_H \text{ e } \gamma'=-1k_B T \text{ e } \Delta n'=\Delta n_P \quad (3)$$

Para moléculas movidas de S
, onde Δn_S , Δn_H e Δn_P são, respectivamente, as variações do número de moléculas de solvente e dos monômeros hidrofóbicos e polares decorrente do movimento da partícula movida H, P ou S. A variável γ é um nível hidrofóbico que depende se o monômero considerado é hidrofóbico ou polar na sequência de uma proteína; já a função $\min\{i, j\}$ significa que é utilizado o menor dos dois valores i e j . Neste capítulo estaremos apenas fazendo uso da probabilidade de transição $T(A \rightarrow B)$.

Assim, a simulação Monte Carlo torna-se simples: para cada nova configuração B gerada, sucedendo uma determinada configuração A, a transição $A \rightarrow B$ é analisada da seguinte forma:

(a) A restrição de esfera rígida (sobreposição entre partículas) é verificada para a nova configuração B gerada; se esta condição for satisfeita (não superposição) passa-se para o item seguinte, ou seja, item (b); do contrário (superposição), passa-se para o item (c);

(b) A nova configuração é aceita de acordo com probabilidade de transição $T(A \rightarrow B)$ descrita pelas Equações 1-3, que é comparada com um número aleatório r , entre 0 e 1, caso $r \leq T(A \rightarrow B)$ aceita-se a nova configuração B, caso contrário permanece-se com a configuração A;

(c) Gera-se uma nova configuração B para outra partícula movida e volta-se ao item (a) acima (por um número pré-determinado de vezes).

No processo de simulação, cada uma das 12.600 partículas é escolhida aleatoriamente, com a mesma probabilidade entre os monômeros e as moléculas do solvente, para se executar cada movimento. Toda vez que o número de tentativas de movimentos for igual ao número de partículas, ou seja, 12.600 conta-se uma unidade de tempo, identificada por 1 PMC (iniciais de Passo Monte Carlo).

Na etapa de parametrização da cadeia é realizado, para cada simulação, um total de 800 kPMC (isto é, 800.000 PMC), gerando assim cerca de $1,01 \times 10^{10}$ configurações. A simulação consta da fase de descarte configuracional (primeiros 100 kPMC) em que se estabelece o regime de temperatura infinita ($k_B T \rightarrow \infty$), somente com o potencial de esferas rígida, seguida por 700 kPMC a baixa temperatura, cujo valor será determinado a seguir. Nesta segunda fase tanto o potencial de esfera rígida quanto o potencial hidrofóbico são considerados, para uma proteína pré-escolhida, no caso a ltsk.

3.1.1 Comportamento configuracional global: O fator temperatura

Na determinação da temperatura adequada para o processo de *folding* são escolhidos alguns valores estratégicos, para testes, desde baixa temperatura ($k_B T = 0,9$) em que a cadeia fica facilmente presa em mínimos locais de energia, até alta temperatura ($k_B T = 10,0$) na qual a evolução configuracional da cadeia é limitada basicamente pelos vínculos estéricos.

Para observar como o empacotamento global da cadeia se dá em função da temperatura, são utilizadas grandezas como o raio de giração (equação nº 1, Capítulo 1), cujas variações dos valores médios para diversas temperaturas registradas em três simulações independentes, rotuladas por "1", "2", "3", caracterizadas, cada uma, por diferentes condições iniciais através do emprego de três conjuntos diferentes de seqüências de números aleatórios r , para a escolha de partículas a serem movidas e para aceite configuracional (item (b) acima). Adicionalmente, também foi considerado o desvio padrão, Dp_G . Ambas as grandezas, Figura 3 a e b, são mostrados em função de $\log(k_B T)$, num intervalo de tempo afastado da fase de descarte, no caso de 401 ao 500 kPMC.

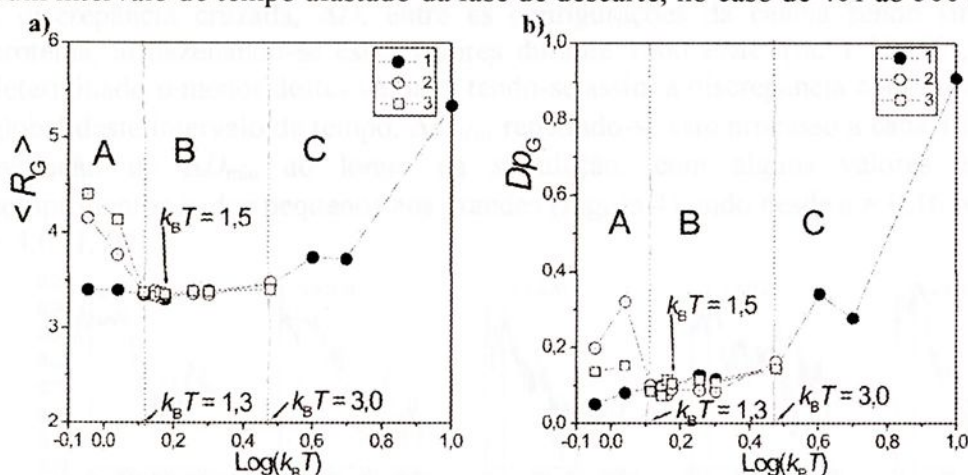


Figura 3- a) Raio de giração médio $\langle R_G \rangle$ versus $\log(k_B T)$ de 401 a 500 kPMC em três simulações independentes "1", "2", "3"; b) Desvio padrão Dp_G do raio de giração versus $\log(k_B T)$ nas mesmas três simulações para $\langle R_G \rangle$.

O comportamento de Dp_G em função da temperatura, $k_B T$, revela três distintas regiões:

- Para $k_B T < 1,3$ (Região A), Dp_G depende essencialmente das condições iniciais e neste caso a cadeia fica presa em mínimos energéticos locais.
- Em $1,3 \leq k_B T \leq 3,0$ (Região B), Dp_G é praticamente independente das condições iniciais e assume valores mínimos em $k_B T = 1,5$, o que significa conformação tipo glóbulo mais bem definida. Para $k_B T > 1,5$, Dp_G aumenta progressivamente seu valor até $k_B T = 3,0$.
- Para $k_B T > 3,0$ (Região C), Dp_G aumenta enormemente seu valor com o aumento da temperatura, devido a fortes flutuações no tamanho e forma do glóbulo, sendo que para altas temperaturas, $k_B T$ em torno de 10,0, já notamos uma saturação no valor de Dp_G .

Em $k_B T = 1,5$ a cadeia sempre, independente das condições iniciais, colapsa numa conformação tipo glóbulo compacto, com tamanho de glóbulo termodinamicamente mais

bem definido, pois as flutuações térmicas são suficientemente grandes para romper com contatos hidrofóbicos não otimizados, daí passamos a utilizar esta temperatura ($k_B T = 1,5$), inclusive, a seguir, para a determinação do valor mais propício do comprimento máximo, ϵ , entre os monômeros consecutivos na cadeia do modelo.

3.1.2 Parametrização da cadeia: A separação relativa entre monômeros consecutivos

Para verificar como e em quanto o valor do comprimento ϵ (Figura 2) afeta o empacotamento da cadeia, o grau de sucesso em encontrar configurações proximamente à estrutura nativa da proteína Itsk, e o desempenho computacional do programa, são utilizadas grandezas como o raio de giração, a percentagem de movimento aceito para os monômeros na cadeia e a função discrepância configuracional (Capítulo 2) entre as configurações da cadeia modelo simulada e a estrutura nativa da Itsk.

Na simulação (na condição inicial 1), após cada Passo Monte Carlo, verifica-se qual a discrepância cruzada, ΔD , entre as configurações da cadeia sendo simulada e a da proteína, armazenando-se estes valores durante 1000 PMC (ou 1 kPMC), em seguida é determinado o menor destes valores, tendo-se assim a discrepância configuracional mínima global deste intervalo de tempo, $\Delta D_{\min}$; repetindo-se este processo a cada 1 kPMC tem-se a evolução de $\Delta D_{\min}$ ao longo da simulação, com alguns valores estratégicos de comprimentos ϵ , dos pequenos aos grandes (Figura 4), indo desde $\epsilon = 0,16$ (ou $0,09D$) até $\epsilon = 3,0$ ($1,7D$).

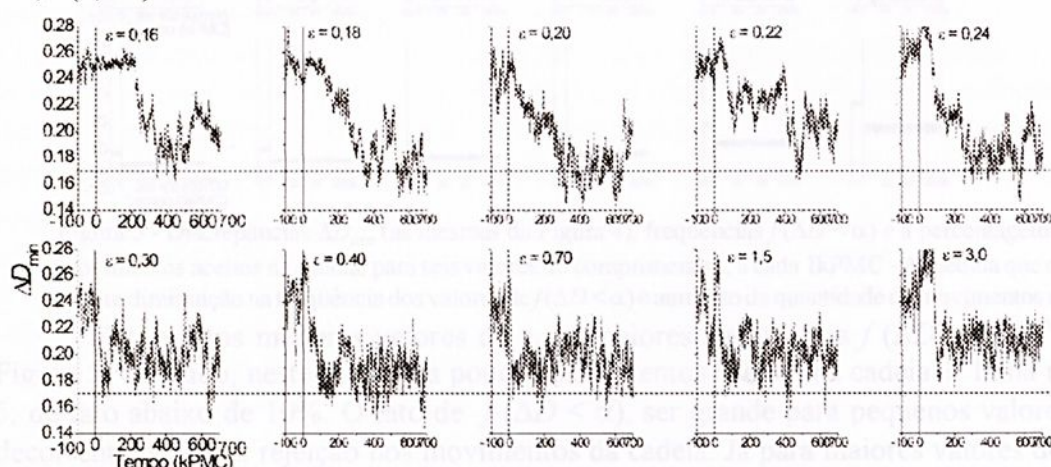


Figura 4 - Discrepância mínima, $\Delta D_{\min}$, a cada 1 kPMC, entre configurações modelo (da simulação) com a proteína Itsk, para vários comprimento ϵ , entre monômeros consecutivos no modelo. As melhores aproximações configuracionais, entre modelo e proteína, ocorrem com discrepâncias abaixo de 0,17 (o por que deste valor para a Itsk será mostrado na Tabela 1 deste capítulo), destacado por retas pretas horizontais, estando estas aproximações em ϵ de 0,18, 0,2, 0,24 e 0,4.

Obs. Tempos negativos (de -100 a 0 kPMC) correspondem a fase de descarte configuracional, a temperatura infinita.

Na Figura 4, observa-se que as melhores aproximações entre modelo e proteína, fornecidas por $\Delta D_{\min}$, no transcorrer do tempo de simulação, ocorrem para pequenos valores de ϵ , particularmente em 0,18, 0,2, 0,24 e 0,4. Em todos estes casos houveram acessos à discrepâncias ao redor de 0,14, quando ocorre a maior semelhança entre as configurações modelo com a proteína Itsk.

A partir dos resultados obtidos pela aproximação configuracional entre modelo e proteína, Figura 4, pode-se reduzir os casos a serem analisados mais minuciosamente, ficando com ϵ entre 0,18 e 0,4, onde ocorrem as menores e mais freqüentes discrepâncias mínimas.

Durante a simulação, para os seis diferentes valores de ϵ , entre 0,18 e 0,40, calcula-se ΔD entre a configuração nativa da proteína 1tsk e as configurações geradas e armazena-se o menor destes valores, representado por $\Delta D_{\min}$, dentre cada intervalo de tempo t de 1 kPMC ($\Delta D_{\min}(t)$, 1ª linha da Figura 5, com os seis melhores valores de ϵ , da Figura 4). Adicionalmente verifica-se a ocorrência de $\Delta D < \alpha$, acumulando toda vez que isto ocorrer na freqüência de discrepância $f(\Delta D < \alpha)$, sendo f um número inteiro que satisfaz $0 \leq f(\Delta D < \alpha) \leq 1000$, 2ª linha da Figura 5, e $\alpha = \Delta D_{\min}(t) + 0,05$. Também foi averiguada a percentagem de movimentos aceitos para os monômeros na cadeia protéica do modelo, 3ª linha da Figura 5.

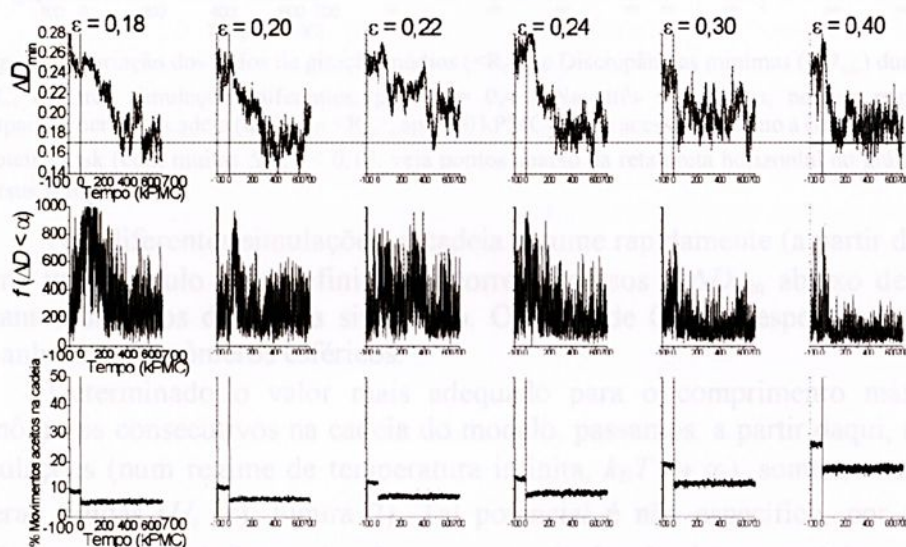


Figura 5 - Discrepâncias $\Delta D_{\min}$ (as mesmas da Figura 4), freqüências $f(\Delta D < \alpha)$ e a percentagem média de movimentos aceitos na cadeia, para seis valores do comprimento ϵ , a cada 1kPMC - A medida que ϵ aumenta ocorre diminuição na freqüência dos valores de $f(\Delta D < \alpha)$ e aumento da quantidade de movimentos na cadeia.

Nota-se nos menores valores de ϵ , as maiores freqüências $f(\Delta D < \alpha)$, 2ª linha na Figura 5. Contudo, nestes casos há poucos movimentos aceitos na cadeia 3ª linha na Figura 5, de fato abaixo de 10%. O fato de $f(\Delta D < \alpha)$, ser grande para pequenos valores de ϵ , é decorrente de muita rejeição nos movimentos da cadeia. Já para maiores valores de ϵ (0,3 e 0,4) as freqüências $f(\Delta D < \alpha)$ são mais reduzidas, mas os movimentos aceitos são significativamente maiores. Para $\epsilon = 0,4$, as frações de movimentos aceitos ficam próximas de 20%, sendo este valor quase a metade da freqüência dos movimentos aceitos para o solvente (50%, resultados estes não mostrados aqui), conseqüentemente, requerendo menos tempo de simulação para verificar as mesmas alterações na cadeia observadas em ϵ pequenos. Por exemplo, em $\epsilon = 0,4$ há cerca da metade do tempo de máquina requerido para $\epsilon = 0,2$. Portanto $\epsilon = 0,4$ foi o valor escolhido.

A escolha por $\epsilon = 0,4$ satisfaz dois critérios importantes: primeiro fornece boa aproximação configuracional entre o modelo e a proteína; segundo oferece melhor desempenho computacional.

Para completar a análise, com $\varepsilon = 0,4$, foram realizadas três simulações, com diferentes condições iniciais, Figura 6.

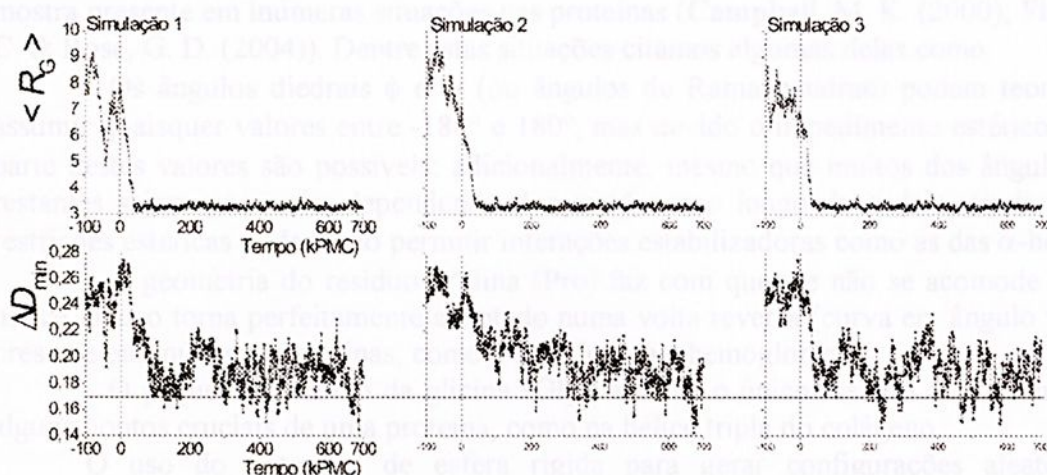


Figura 6 - Variação dos Raios de giração médios ($\langle R_G \rangle$) e Discrepâncias mínimas ($\Delta D_{\min}$) durante o tempo MC, em três simulações diferentes, para $\varepsilon = 0,4$. Nas três simulações, nota-se regularidade no empacotamento da cadeia (dada por $\langle R_G \rangle$, após 50 kPMC), e nos acessos próximo à configuração nativa da proteína 1tsk (com muitas $\Delta D_{\min} < 0,17$, veja pontos abaixo da reta preta horizontal no gráficos de $\Delta D_{\min}$ versus Tempo MC).

Nas diferentes simulações, a cadeia assume rapidamente (a partir de 50 kPMC) uma forma tipo glóbulo bem definido e ocorrem acessos a $\Delta D_{\min}$ abaixo de 0,17 em muitos instantes distintos durante a simulação. O valor de 0,4 corresponde à cerca de 23% do tamanho dos monômeros esféricos.

Determinado o valor mais adequado para o comprimento máximo ε entre os monômeros consecutivos na cadeia do modelo, passamos, a partir daqui, a utilizar este nas simulações (num regime de temperatura infinita, $k_B T \rightarrow \infty$), somente com o potencial de esferas rígidas ($U_i(r)$, Figura 2). Tal potencial é não específico, por não proporcionar nenhuma preferência interacional entre as partículas do sistema cadeia-solvente, e também não estabilizador, pelo fato da cadeia protéica do modelo não ter nenhuma configuração ou forma preferencial. Porém tal potencial se mostra propício para acessar grande parte do vasto espaço configuracional disponível para uma cadeia, pois pode se obter desvinculadamente, numa ampla faixa de compactação da cadeia, semelhanças entre estas configurações aleatórias “quaisquer” e estruturas de proteínas. Os resultados de tais semelhanças podem, entre outros aspectos, mostrar como se processa o acesso às configurações nativas das proteínas e também podem servir de base para outras situações, por exemplo, quando houver outros potenciais na simulação MC.

3.2 Configurações *random coil* de uma cadeia imersa no solvente

Dentre as várias interações energéticas encontradas nas proteínas (van der Waals, ligação de hidrogênio, interações hidrofóbicas e eletrostáticas, e outras como pontes dissulfeto), o de esfera rígida é tido, como uma representação interacional que se constitui em ingrediente fundamental e utilizando em praticamente todas as diferentes abordagens do

estudo protéico, inclusive no problema do *folding*. O impedimento estérico, dado pelo efeito de esfera rígida (representa a interação repulsiva do potencial de van der Waals), se mostra presente em inúmeras situações nas proteínas (Campbell, M. K. (2000); Fitzkee, N. C. & Rose, G. D. (2004)). Dentre estas situações citamos algumas delas como:

- Os ângulos diedrais ϕ e ψ (ou ângulos de Ramachandran) podem teoricamente assumir quaisquer valores entre -180° e 180° , mas devido o impedimento estérico somente parte destes valores são possíveis; adicionalmente, mesmo que muitos dos ângulos ϕ e ψ restantes sejam possíveis, dependendo dos resíduos ao longo da cadeia, ainda assim as restrições estéricas podem não permitir interações estabilizadoras como as das α -hélice.

- A geometria do resíduo prolina (Pro) faz com que ele não se acomode numa α -hélice, mas o torna perfeitamente adaptado numa volta reversa (curva em ângulo fechado), presente em inúmeras proteínas, como a mioglobina e hemoglobina.

- O pequeno tamanho da glicina (Gly), a torna o único resíduo a se acomodar em alguns pontos cruciais de uma proteína, como na hélice tripla do colágeno.

O uso do potencial de esfera rígida para gerar configurações aleatórias em simulação Monte Carlo, já foi vastamente empregado (Cantor, C. R. & Schimmel, P. R. (1980)), inclusive relacionando grandezas como os raios de giração e distância ponta a ponta (D_{pp} , distância euclidiana, entre o primeiro e o último resíduo da cadeia). Aqui, as configurações de uma cadeia imersa no solvente aquoso são abordadas em duas etapas:

- Na primeira (seção 3.2.1) será feita a caracterização destas configurações quanto a sua compactação e contatos entre seus monômeros;

- Na segunda etapa (seção 3.3) serão feitas comparações configuracionais destas configurações com outras configurações “alvos” tanto do modelo quanto de proteínas, por meio da função discrepância configuracional, ΔD e mapas de distância.

No presente contexto, entendemos por configurações alvos, configurações da cadeia do modelo, escolhidas estrategicamente durante uma simulação MC, para efeitos diversos de comparação configuracional. Tais comparações são feitas, durante a simulação, entre configurações da cadeia do modelo, ou entre estas e as estruturas nativas escolhidas para estudo (Apêndice, após o Capítulo 5) neste trabalho. Também as estruturas nativas são referidas como configurações alvos.

3.2.1 Caracterização quanto à compactação e contatos intra cadeia

A caracterização de configurações pode ser feita por grandezas que medem globalmente a compactação da cadeia e os contatos entre monômeros nestas cadeias. No presente caso tais grandezas devem tornar possível a comparação entre duas configurações quaisquer, como configurações alvos e *random coil* do modelo e entre estas e as estruturas das proteínas reais utilizadas aqui. De fato a capacidade de comparar configurações de diferentes resoluções espaciais é a principal característica que se espera de uma ferramenta, que mede semelhanças e diferenças configuracionais, como a função discrepância descrita no Capítulo 2.

As configurações alvos do modelo são escolhidas numa etapa inicial de simulação, que envolve 400 kPMC, em seguida são geradas as configurações *random coil*, após cada 1 PMC (de 1 a 600 kPMC), totalizando 600.000 destas configurações, num tempo total de simulação de 1000 kPMC.



Inicialmente procede-se a caracterizar estas configurações *random coil*, quanto ao nível de empacotamento e contatos entre seus resíduos, e só então se processa a análise do quanto tais configurações se assemelham às configurações alvo do modelo ou proteínas.

As configurações *random coil* podem ser também caracterizadas (Figura 7), quanto aos contatos entre seus resíduos, pelo número de contatos topológicos e pela ordem destes contatos. Já o seu empacotamento pode ser estabelecido pelo raio de giração e distância ponta a ponta. Para a contagem da frequência destas quatro grandezas, estas são arranjadas em intervalos, por exemplo, unitários, significando que seu valor estiver entre 0,0 e 1,0 ou 1,0 e 2,0 ou 2,0 e 3,0 estes passam a ser, respectivamente 0, 1, 2 e assim sucessivamente; sendo as frequências destas grandezas referentes a ocorrências destes valores.

Para se quantificar os contatos topológicos, ou seja, determinar o número η de resíduos que estão em contato (vizinhos) com o i -ésimo resíduo, são excluídos os dois primeiros resíduos consecutivos do resíduo i ao longo da cadeia (isto é, i com $i-1$, $i-2$, $i+1$ e $i+2$), como foi feito na função discrepância, ΔD (Capítulo 2). Além disso, os resíduos devem ter distância euclidiana com outros resíduos menor do que uma distância predefinida de corte, d_c .

A ordem de contato, χ , é definida como:

$$\chi = (N-1)^{-1} \sum_{i,j} v_{i,j} / \eta, \quad (4)$$

onde $(N-1)$ é o comprimento da cadeia de N resíduos; η é o número de contatos topológicos, podendo η variar idealmente desde 0 (zero) até $M = \binom{N-2}{2} = 528$ contatos

(Equação 3 do Capítulo 2), o máximo de contatos na cadeia; e $v_{i,j} = (j-i)$, com $j > i$ é o número de resíduos ao longo da cadeia entre os resíduos i e j em contato topológico.

O valor da ordem de contato χ (equação 4) é idealmente restrito ao intervalo $0 \leq \chi \leq 1$, mas para proteínas reais χ é tipicamente pequeno ((**Bonneau, R. et al. (2002), Grantcharova, V. et al. (2001))**), $\chi \leq 0,2$.

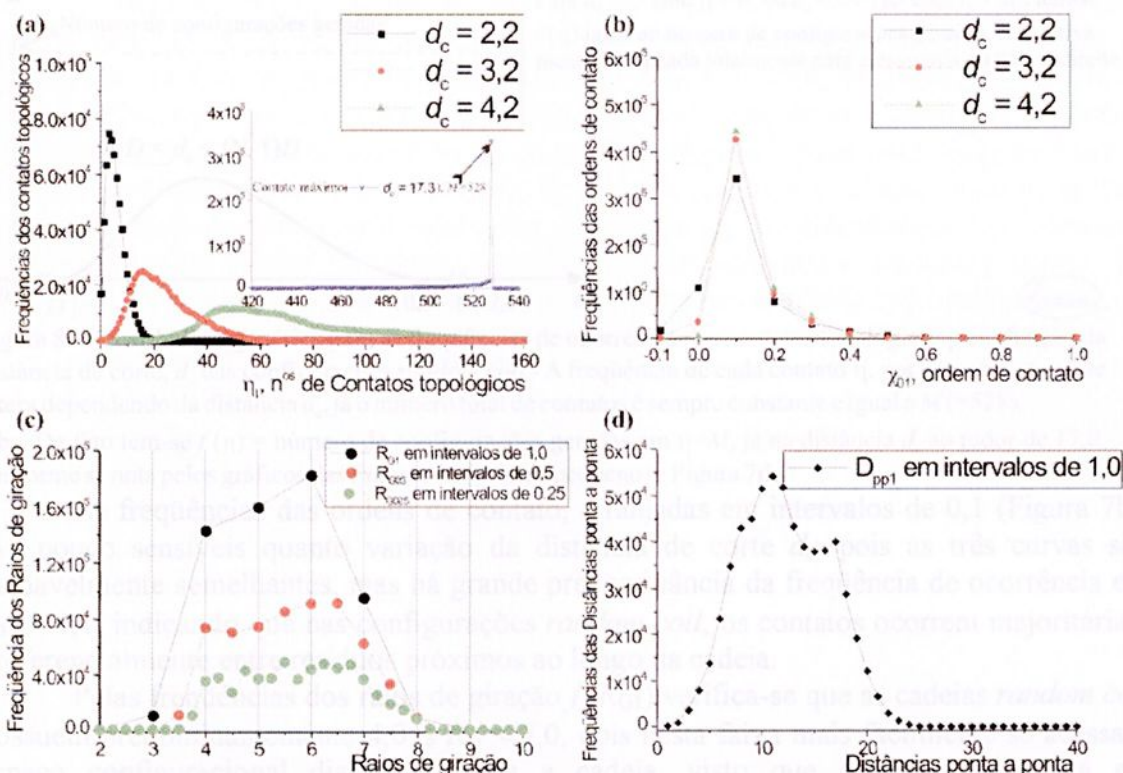


Figura 7 - Frequências dos contatos topológicos(a), das ordens de contato(b), dos raios de giração(c) e distâncias ponta a ponta(d) das configurações *random coil* geradas em 600 kPMC, para a primeira simulação, de um total de nove simulações, todas apresentando resultados quase idênticos aos mostrados acima. Se a distância de corte for pequena, como em $d_c = 2,2$, às vezes, há configurações sem contatos topológicos ($\eta=0$); para este caso foi estabelecido um valor fictício de $\chi = -0,1$ (Figura 7b). O critério para escolha das distâncias de corte, foram $d_c = 2,0 (\approx D)$; $3,2 \approx 2,0 \cdot D$; $4,2 \approx 2,5 \cdot D$; $17,3 = 10 \cdot D$.

Obs. - Cada distância D_{pp} é referente a apenas uma distância, para cada configuração, assim $f(D_{pp1})$, Figura 7d, fica mais dependente da simulação, por exemplo, o valor quase constante de $f(D_{pp1})$ em D_{pp1} ao redor de 15 não foi verificado nas outras oito simulações.

- Os valores de R_g são agrupados em três intervalos (1,0; 0,5; 0,25) devido a menor variação que D_{pp} .

A distribuição das frequências de ocorrências, $f(\eta_1)$ dos contatos topológicos, arranjados em intervalos unitários (η_1 , Figura 7a) depende da distância de corte, d_c , já que com o aumento de d_c (de 2,2 e 4,2) aumentam se os valores e as quantidades destes contatos. Assim, $f(\eta_1)$ muda drasticamente sua distribuição variando sua amplitude máxima e sua largura. As distâncias d_c grandes ($d_c \geq 4,2$) passam a ser basicamente ilustrativas do comportamento de grandezas como η e χ , pois não tornam possível identificar adequadamente diferenças estruturais nas *random coil*, daí nos próximos resultados serão empregados apenas d_c de 2,7 e 3,2. Nos casos extremos da distância d_c a frequência dos contatos topológicos (Figura 8) se torna única e tem valor igual ao número M de cadeias geradas, no caso $M = 600.000$.

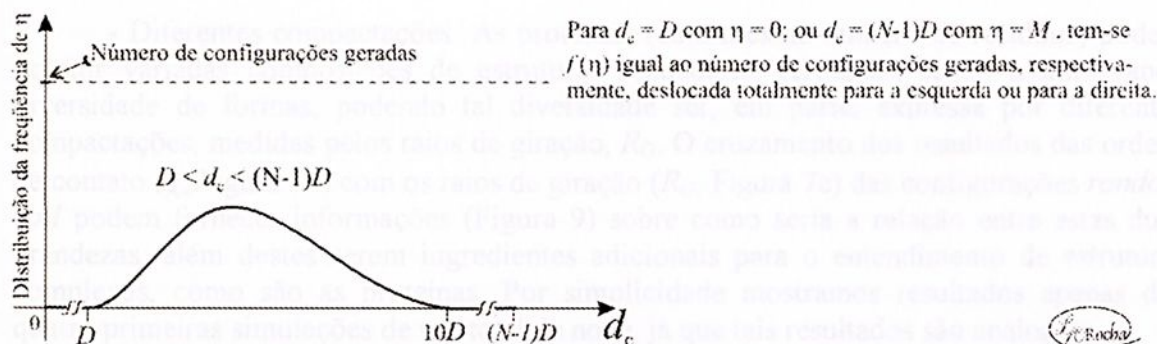


Figura 8 - Distribuição genérica para as frequências de ocorrência dos contatos topológico η , em função da distância de corte, d_c , das configurações *random coil* - A frequência de cada contato η , por exemplo, $\eta = 5$ se altera dependendo da distância d_c , já o número total de contatos é sempre constante e igual a $M (=528)$.

Obs. De fato tem-se $f(\eta) =$ número de configurações geradas em $\eta=M$, já na distância d_c ao redor de 17,0, conforme se nota pelos gráficos das Figuras 7a (gráfico pequeno) e Figura 7d.

As frequências das ordens de contato, arranjadas em intervalos de 0,1 (Figura 7b), são pouco sensíveis quanto variação da distância de corte d_c , pois as três curvas são razoavelmente semelhantes, mas há grande predominância da frequência de ocorrência em $\chi_{01} = 0,1$, indicando que nas configurações *random coil*, os contatos ocorrem majoritariamente e preferencialmente entre resíduos próximos ao longo da cadeia.

Pelas frequências dos raios de giração $f(R_G)$ verifica-se que as cadeias *random coil* possuem predominantemente $4,0 \leq R_G < 7,0$, pois nesta faixa mais facilmente se acessa o espaço configuracional disponível para a cadeia, visto que a cadeia não está em compactações extremadas (muito compactas ou muito distendidas). Para $4,0 \leq R_G < 7,0$, nota-se razoável uniformidade, em $f(R_{G025})$, como decorrência da utilização somente do potencial de esfera rígida na simulação; pode-se notar também as configurações muito compactas (abaixo de 4,0) sendo mais raras do que as muito distendidas (acima de 7,0), pois nestas últimas os raios de giração se reduzem mais lentamente (Cantor, C. R. & Schimmel, P. R. (1980)).

As distâncias entre as pontas, D_{pp1} , ocorrem majoritariamente entre 5,0 e 20,0, pois configurações extremadas (com pontas muito próximas ou muito distantes), com no caso de R_G , são mais difíceis de serem acessadas, embora estas distâncias possam hipoteticamente variar desde o seu diâmetro $D = 1,73$ até $\sqrt{3}L$, onde $L = 49$ é o lado da caixa cúbica contendo o sistema cadeia-solvente.

Pelos resultados mostrados na Figura 7, tem-se resumidamente que as configurações *random coil*, são, na maioria das vezes, encontradas com baixa ordem de contato, $\chi \approx 0,1$ e com compactação predominantemente de $R_G \approx 5,0$ e 6,0 e distância entre as pontas mais frequentes em $D_{pp} \approx 11,0$ (tais valores para R_G e D_{pp} estão próximos dos previstos teoricamente Doi, M. (1996)); dentre estes resultados, há dois ingredientes também encontrados em proteínas reais, são eles:

- Baixa ordem de contato, χ : Como nas configurações *random coil* (Figura 7b), nas proteínas nota-se preferencialmente χ com valores pequenos, pois as interações locais (com χ pequenos) ocorrem mais precocemente no processo de *folding* (Baldwin, R. L. & Rose, G.D. (1999); Lattman, E. E. & Rose, G.D. (1993)), assim restringindo subseqüentes eventos neste processo e influenciando decisivamente nas interações não locais (com χ grandes), daí determinando suas possibilidades de ocorrências.

- Diferentes compactações: As proteínas (com mesmo número de resíduos) podem possuir variadas composições de estruturas secundária, terciárias, tendo assim grande diversidade de formas, podendo tal diversidade ser, em parte, expressa por diferentes compactações, medidas pelos raios de giração, R_G . O cruzamento dos resultados das ordens de contato (χ , Figura 7b) com os raios de giração (R_G , Figura 7c) das configurações *random coil* podem fornecer informações (Figura 9) sobre como seria a relação entre estas duas grandezas, além destes serem ingredientes adicionais para o entendimento de estruturas complexas, como são as proteínas. Por simplicidade mostramos resultados apenas das quatro primeiras simulações de um total de nove, já que tais resultados são análogos.

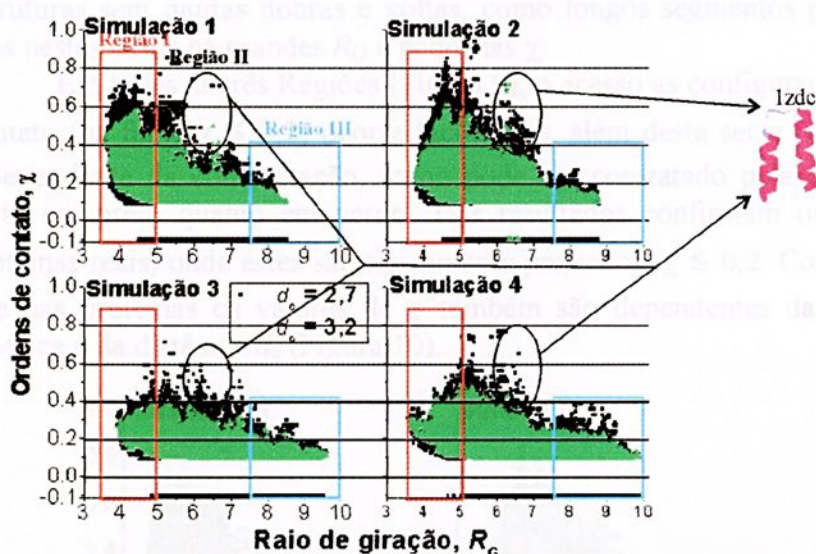


Figura 9 - Configurações *random coil* caracterizadas pelo Raio de giração (R_G) e ordens de contato (χ) para distância de corte $d_c = 2,7$ (em preto) e $3,2$ (em verde) em quatro simulações (cada uma com 60.000 pontos)- Nota-se, em todas as simulações, figuras com contornos diversificados e alguns locais isolados ("ilhados"). Na Região I ($R_G \leq 5,0$) a variação de χ além de ser maior tem sua distribuição mais dependente da simulação; Na Região II ($5,0 < R_G < 7,5$) o aumento do raio de giração levam a diminuição progressiva da orden de contato máxima, a exceção de configurações peculiares, em forma das letras U ou V, tipo proteína n° 14, 1zdc; Na Região III ($R_G > 7,5$) estruturas *random coil* mais distendidas, apresentam ordens de contato, normalmente abaixo de 0,2. Com o aumento das distâncias d_c , de 2,7 para 3,2, começa-se perder muitas informações sobre os contatos topológicos, representado pelos pontos pretos contornando verdes, por isso não utilizamos $d_c > 3,2$. Obs. - As frequências de R_G e χ se mostram idênticos aos da Figura 7, diferenciando-se por um fator 10, já que naquele caso foi utilizado 600.000 configurações ao invés das 60.000 daqui, obtidas após cada 10 PMC.

Observando as variações das ordens de contato, χ , com a compactação, R_G , nas nove simulações realizadas, é possível identificar as regiões I, II e III para R_G , tendo estas as seguintes características principais:

- Região I ($R_G \leq 5,0$), há proporcionalmente muitas configurações *random coil* com altas ordens de contato ($\chi > 0,4$), contudo os acessos a tais configurações são bastante dependentes das simulações, já que estas configurações, por serem muito compactas, nem sempre são facilmente acessíveis. A existência de grandes ordens de contato aliada a grande compactação torna esta região a preferida para a existência de estruturas mais complexas, como as compostas por diversificadas composições, semelhantes a *turns*, folhas β e α -hélices.

- Região II ($5,0 < R_G < 7,5$), contém a maior parte das configurações geradas nas simulações (como mostrado na Figura 7c) e com a diminuição progressiva da compactação

das cadeias, geralmente os contatos se tornam progressivamente mais locais, representado pela diminuição das ordens de contato máximas para os R_G . Nesta região ainda pode-se notar algumas configurações com disposições peculiares, por exemplo, na forma da letra U ou V (cujas estruturas são parecidas com a proteína nº 14, 1zdc ($\chi_{1zdc} = 0,53$ e $0,32$, respectivamente para distâncias $d_c = 4,2$ e $5,0$), mostrada no topo da Figura 9), sendo estas consideradas peculiares por serem menos compactas, mas ainda possuindo altas ordens de contato, característico da Região I.

- Região III ($R_G > 7,5$), ocorrem estruturas mais distendidas, ou menos compactas, com as menores ordens de contato (geralmente $\chi < 0,2$), sendo esta região a preferida para estruturas sem muitas dobras e voltas, como longos segmentos parecidos com α -hélices, pois nestes casos há grandes R_G e pequenas χ .

Em todas as três Regiões (Figura 9), o acesso as configurações com baixa ordem de contato (na faixa $\chi \leq 0,2$) ocorre facilmente, além desta ser a única faixa de χ existente independente da compactação, como pode ser constatado pela contínua área preenchida tanto em preto quanto em verde. Tais resultados confirmam os verificados em muitas proteínas reais, onde estes são tipicamente pequenos, $\chi \leq 0,2$. Contudo deve se considerar que nas proteínas os valores de χ também são dependentes da compactação da cadeia protéica e da distância d_c (Figura 10).

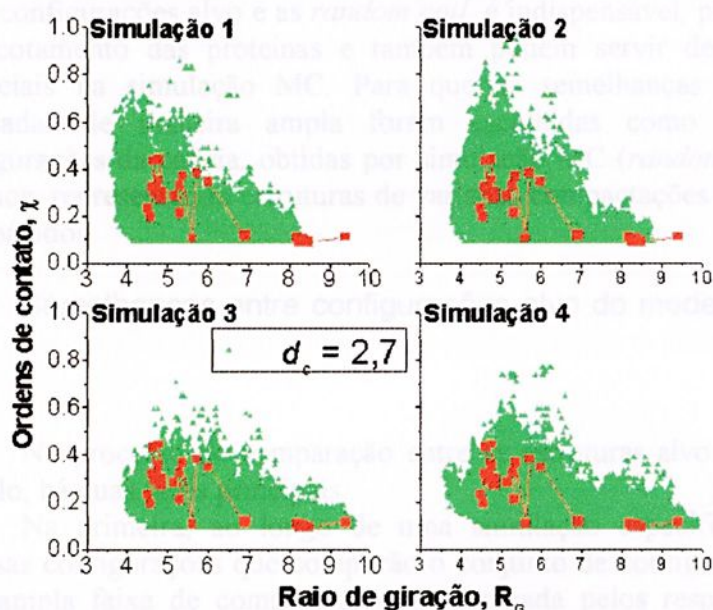


Figura 10 - Relação entre ordens de contato das proteínas e *random coil* - As ordens de contato, χ , das proteínas (pontos conectados por retas ambos em vermelhos e repetidos quatro vezes), quase sempre, apresentam-se como casos particulares das χ das configurações *random coil* (pontos em verde, com $d_c=2,7$), pois as primeiras geralmente estão contidas nas segundas. Nos gráficos desta figura as distâncias de corte, d_c , e os raios de giração, R_G , nas proteínas, são divididas por 1,8, ou seja, d_c é de 5,0 ao invés de 2,7 no modelo e R_G varia de 8,0 a 17,0 ao invés de 3,0 e 10,0 no modelo. Eventuais questões de escala, como o fator 1,8, não dificultam comparações entre estruturas proteína com *random coil*, pois a função discrepância, ΔD , é invariante quanto a escala (Capítulo 2).

Obs.- Pela disposição entre os pontos verdes e vermelhos acima constatamos que com as estruturas *random coil* é possível se ter estruturas mais compactas, mas não mais distendidas, que as das proteínas.

- Não mostraremos os resultados que revelam um fator 1,8 (mais especificamente este valor é de $1,77 \pm 0,15$) para a relação entre os raios de giração das proteínas com os raios de giração médios das cadeias do modelo, que mais se aproximam das proteínas, nas nove simulações.

Assim, proteínas com ordem de contato grande ($\chi > 0,2$), são possíveis, embora raras e sob este aspecto as estruturas nativas das proteínas se apresentam como casos particulares das *random coil*.

Após caracterizar globalmente as configurações *random coil* (Figuras 7, 9 e 10), inclusive identificando, nestas, padrões e características também existentes em proteínas reais, passamos a verificar mais peculiaridades destas estruturas aleatórias, ao medir suas semelhanças com outras configurações predeterminadas, denominadas alvo (do modelo e das proteínas). Para essa finalidade será empregada a função discrepância configuracional, ΔD , já descrita anteriormente.

3.3 Semelhanças configuracionais: As Discrepâncias cruzadas mínimas globais ($\Delta D_{\min G}$) e médias ($\langle \Delta D_{\min G} \rangle$)

O estabelecimento dos valores limiares das discrepâncias, refletindo as semelhanças entre configurações alvo e as *random coil*, é indispensável, para o correto entendimento do empacotamento das proteínas e também podem servir de base quando houver outros potenciais na simulação MC. Para que as semelhanças entre estruturas possam ser analisadas de maneira ampla foram escolhidas como configurações alvos, várias configurações da cadeia, obtidas por simulação MC (*random coil*) e 27 proteínas com 35 resíduos, representando estruturas de variadas compactações e composições de resíduos de aminoácidos.

3.3.1 Semelhanças entre configurações alvo do modelo e configurações *random coil*

No processo de comparação entre as estruturas alvo e *random coil* de cadeias do modelo, há duas fases principais.

Na primeira, ao longo de uma simulação específica são geradas e recolhidas diversas configurações que comporão o conjunto de configurações alvo. Eles representam uma ampla faixa de compactação caracterizada pelos respectivos raios de giração; são escolhidas ao final de cada 1.000 Passos Monte Carlo, sendo que, em geral, são separadas uma das outras por pelo menos 3 kPMC, o que garante muito baixa correlação entre elas.

Na segunda fase uma nova simulação (independente) da evolução configuracional é realizada, e a configuração obtida após cada PMC é usada para se determinar a sua correspondente discrepância, em relação a cada configuração do conjunto de configurações alvo, previamente selecionadas. Em cada simulação desta categoria são realizadas 600.000 comparações, Figura 11.



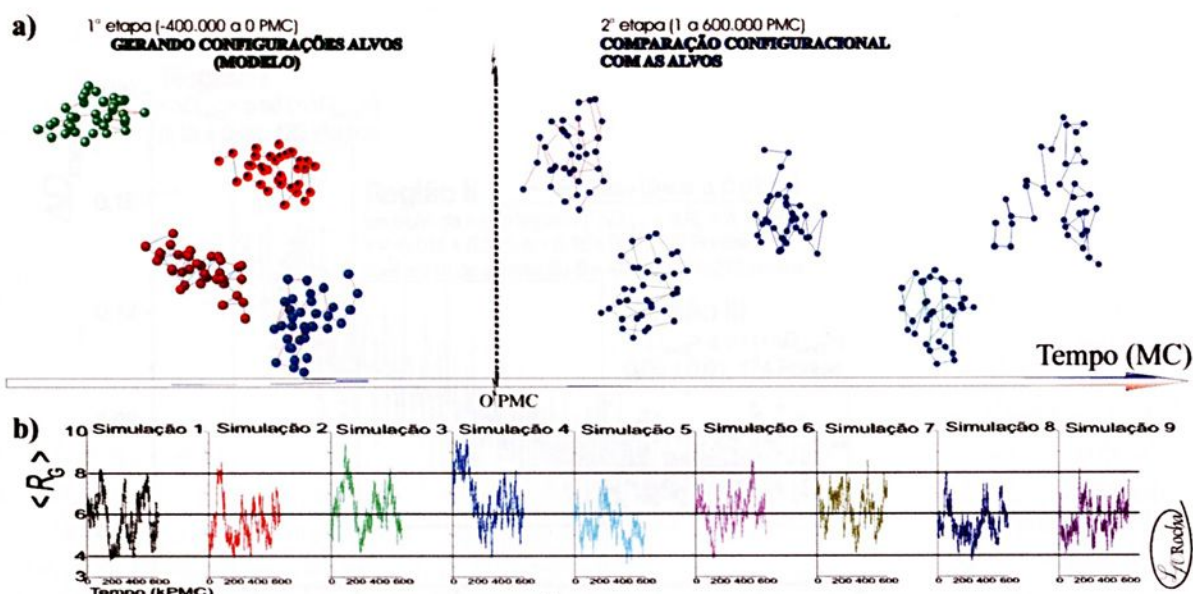


Figura 11 - Discrepâncias cruzadas: a) Numa 1ª fase são geradas e escolhidas as configurações alvos (do modelo); em seguida, numa 2ª fase, através de uma simulação independente da evolução configuracional das *random coil*, são feitas comparações com as configurações alvo através da função discrepância, ΔD ; b) A variação temporal dos Raios de giração médios, em 600 kPMC, de evolução configuracional das *random coil*, mostra valores desde muito compactos (R_G abaixo de 5,0) até os muito distendidos (R_G acima de 7,0).

A comparação configuracional com as alvos se processa de forma a se determinar a mínima discrepância alcançada possível, no tempo MC equivalente à 600.000 PMC, que é a janela de tempo utilizada. Assim, a configuração ao final de cada PMC é comparada, via discrepância ΔD , com cada uma das configurações alvo. Ao final da simulação para cada uma das estruturas alvo, de diferentes compactações, tem-se a maior semelhança possível alcançada com as configurações *random coil*; sendo a função discrepância denominada discrepância mínima global ($\Delta D_{\min G}$). Tal procedimento é adotado em nove simulações independentes (em diferentes condições iniciais de simulação). O conjunto de configurações alvo é composto por 569 configurações de diferentes R_G (Figura 12).

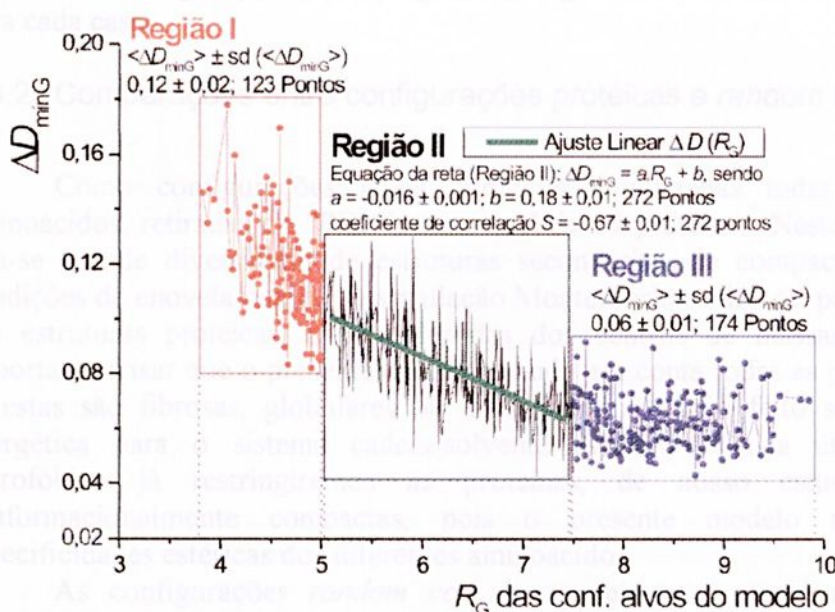


Figura 12 - Discrepâncias mínimas globais, $\Delta D_{\min G}$, obtidas após nove simulações independentes, em função dos raios de giração, R_G , das 569 estruturas alvos do modelo (total de pontos acima) - É possível identificar três Regiões:

Região I: Estruturas compactas ($R_G \leq 5,0$) com $\Delta D_{\min G}$, em geral de 0,10 a 0,14.

Região II: Estruturas com compactação intermediária ($5,0 < R_G < 7,5$), com $\Delta D_{\min G}$ decrescendo linearmente com o aumento do R_G , sendo S o fator de correlação linear para a equação da reta $\Delta D_{\min G}(R_G)$.

Região III: Estruturas distendidas ($R_G > 7,5$) com $\Delta D_{\min G}$ oscilando ao redor de 0,06.

Obs.: - As seguintes quantidades de estruturas alvos do modelo foram usadas, respectivamente, em cada uma das nove simulações independentes: 53, 68, 46, 52, 72, 59, 50, 67 e 36, totalizando as 569 configurações alvo.

- Por serem mais raras que as demais, para coletar mais configurações muito compactas ($R_G \leq 5,0$) foi executada adicionalmente a décima simulação, para mais 66 estruturas alvo.

- O termo sd (iniciais do Inglês, *square deviation*), em $sd(\langle \Delta D_{\min G} \rangle)$ é o desvio padrão de $\langle \Delta D_{\min G} \rangle$.

Pela Figura 12, de forma geral, temos que:

As discrepâncias mínimas globais, $\Delta D_{\min G}$, são menores para as cadeias mais distendidas, R_G grandes (Região III), pois é razoável admitir que para duas cadeias distendidas não há muitas formas de se diferenciarem entre si, e o valor de $\Delta D_{\min G}$, fica em $0,06 \pm 0,01$, como aliás foi demonstrado no Capítulo 2.

No outro extremo, isto é, para cadeias muito compactas (Região I), há muitas formas distintas das cadeias diferirem entre si, como manifestado por $\Delta D_{\min G}$, basicamente variando no intervalo compreendido por $0,12 \pm 0,02$. Na parte intermediária (Região II, de $5,0 < R_G < 7,5$) a $\Delta D_{\min G}$ se comporta linearmente, diminuindo com o aumento de R_G , e variando predominantemente entre 0,11 e 0,06.

A segmentação estabelecida pelas três regiões foram determinadas arbitrariamente, unicamente baseada na constatação visual do comportamento do valor de $\Delta D_{\min G}$ em função de R_G .

Uma vez determinada, em três regiões, a relação entre as discrepâncias configuracionais mínimas globais em função da compactação das estruturas alvos do modelo, passa a ser útil agora verificar, primeiro se estas regiões continuam existindo em proteínas reais, segundo, como se daria o acesso à estrutura nativa destas. Assim, agora as estruturas alvos passam a ser as proteínas e a natureza das *random coil* continua sendo a

mesma das configurações já empregadas até agora, totalizando nove diferentes simulações para cada caso.

3.3.2 Comparações entre configurações protéicas e *random coil*

Como configurações alvos, agora são utilizadas todas as proteínas de 35 aminoácidos, retiradas do “PDB”, num total de 27 proteínas. Neste conjunto de proteínas tem-se grande diversidade de estruturas secundárias, de compactações e de diferentes condições de enovelamento. Na simulação Monte Carlo utiliza-se, para a descrição espacial das estruturas protéicas, as coordenadas dos centros de massas dos aminoácidos. É importante frisar que o presente estudo levando em conta todas as proteínas (independente se estas são fibrosas, globulares ou de membrana) e foi feito sem qualquer interação energética para o sistema cadeia-solvente. Mesmo com a utilização do potencial hidrofóbico já restringiremos as proteínas, de nosso estudo (Capítulo 4), às conformacionalmente compactas, pois o presente modelo não inclui ainda as especificidades estéricas dos diferentes aminoácidos.

As configurações *random coil* do modelo, e a maneira como são feitas as comparações destas com as configurações alvo (agora, as 27 proteínas) são analisadas como antes (Figuras 11 e 12). Porém, embora as estruturas alvos permaneçam as mesmas, nas diferentes simulações, mas em número muito mais reduzido, agora passa a ser mais conveniente individualizar os valores de $\Delta D_{\min G}$ e destacando suas médias $\langle \Delta D_{\min G} \rangle$, das nove simulações independentes (Figura 13a). Também porque nestas simulações, para cada uma das 27 proteínas, verifica-se que as discrepâncias mínimas globais, $\Delta D_{\min G}$, no geral, estão variando dentro de uma faixa de até 0,03 (este resultado, será comentado adiante, Figura 16), daí é razoável, e facilita a análise, apresentar somente os valores médios das discrepâncias mínimas globais, $\langle \Delta D_{\min G} \rangle$ para os R_G (Figura 13b).

Os valores de $\langle \Delta D_{\min G} \rangle$ para as diversas R_G (Figura 13a) apresentam muitas equivalências com o observado nas estruturas alvo do modelo *random coil* (Figura 12). Entre essas equivalências destaca-se a existência de três regiões bem definidas de compactações, sendo as fronteiras entre as três regiões razoavelmente bem definidas. Região I formada por proteínas compactas, com $\langle \Delta D_{\min G} \rangle$ valores do que aqueles obtidos para o modelo, respectivamente com valores ao redor de 0,13 e 0,12. Na Região II nota-se um decréscimo linear de $\langle \Delta D_{\min G} \rangle$ com o aumento de R_G , com praticamente a mesma inclinação para o modelo e proteínas, com respectivamente $\langle \Delta D_{\min G} \rangle = -0,0165 R_G^{\text{alc}} + 0,12$ e $\langle \Delta D_{\min G} \rangle = -0,0204 R_G^{\text{alc}} + 0,13$. Já na Região III formada pelas proteínas “fibriladas”, com muitas α -hélices secundárias, as $\langle \Delta D_{\min G} \rangle$ mostram valores próximos ao do modelo, sendo estes 0,03 e 0,05. As mesmas três regiões observadas para $\langle \Delta D_{\min G} \rangle$ e R_G também são observadas em cada uma das nove diferentes simulações, inclusive com comportamento linear na Região II (Figura 13a).

O fato de configurações alvo mais dispersas na Região III, terem as menores discrepâncias $\langle \Delta D_{\min G} \rangle$, pode ser atribuído ao fato de que as estruturas mais dispersas são “contorno mais locais” ou mais próximas ao longo da sequência da cadeia (como em α -hélices). A medida que as estruturas proteicas e modelos se tornam mais compactas os



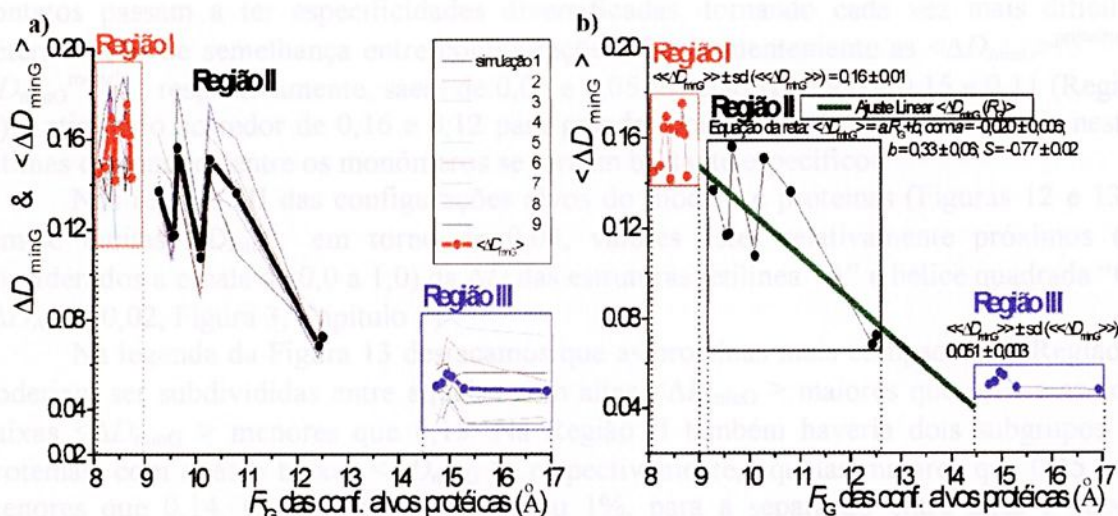


Figura 13 - a) Discrepâncias mínimas Globais ($\Delta D_{\min G}$), em 9 simulações independentes e suas médias ($\langle \Delta D_{\min G} \rangle$ destacada por pontos coloridos conectados por retas) em função dos raios de giração, R_G , das 27 estruturas alvos protéicas - Pela maneira como as discrepâncias variam para os R_G notar as regiões:

Região I ($R_G \leq 9,0$; 12 proteínas): Nesta região estão as proteínas de maiores complexidades estruturais.

Região II (com $9,0 < R_G < 14,5$; 9 proteínas): Nota-se um comportamento decrescente da $\langle \Delta D_{\min G} \rangle$ com o aumento do R_G proteico. Nas diferentes simulações o coeficiente linear S varia de 0,61 a 0,90.

Região III ($R_G > 14,5$; 6 proteínas): A $\langle \Delta D_{\min G} \rangle$ média das 6 proteínas fica em torno de 0,05 e nas diferentes simulações são geradas mais ou menos configurações muito distendidas.

b) São mostradas apenas as $\langle \Delta D_{\min G} \rangle$ de cada proteínas (pontos circulares coloridos, conectados por retas, o mesmo da parte "a" desta Figura) e a reta em verde interpolando os pontos na Região II. Na Região I (pontos vermelho) é possível identificar dois subgrupos de proteínas, com altas $\langle \Delta D_{\min G} \rangle$ maiores que 0,16 e baixas $\langle \Delta D_{\min G} \rangle$ menores que 0,15. Na Região II também há dois subgrupos de proteínas, com altas $\langle \Delta D_{\min G} \rangle$ aquelas maiores que 0,15 e as com baixas $\langle \Delta D_{\min G} \rangle$, as restantes, as menores que 0,14.

Obs. Note que os valores de R_G das proteínas são maiores, por um fator de 1,8 que os da *random coil* (Figura 12); a comparação só é possível graças às características da função discrepância introduzida, como descrito no Capítulo 2.

Os valores de $\langle \Delta D_{\min G} \rangle$ para os diversos $R_G^{\text{proteínas}}$ (Figura 13b), apresentam muitas equivalências com o observado nas estruturas alvos do modelo e *random coil* (Figura 12). Entre estas equivalências destaca-se a existência de três regiões de acordo com as faixas de compactações, sendo as fronteiras entre as três regiões razoavelmente bem definidas:

Região I formada por proteínas compactas, com $\langle \Delta D_{\min G} \rangle$ maiores do que aqueles obtidos para o modelo, respectivamente com valores ao redor de 0,16 e 0,12. Na Região II nota-se um decréscimo linear de $\langle \Delta D_{\min G} \rangle$ com o aumento de R_G^{alvos} , com praticamente a mesma inclinação para o modelo e proteínas, com respectivamente $\langle \Delta D_{\min G} \rangle = -0,016R_G^{\text{alvos}} + 0,18$ e $\langle \Delta D_{\min G} \rangle = -0,020R_G^{\text{alvos}} + 0,33$. Já na Região III formadas pelas proteínas distendidas, com muitas α -hélices sem dobras, as $\langle \Delta D_{\min G} \rangle$ assumem valores próximos, ao do modelo, sendo estes 0,05 e 0,06. As mesmas três regiões observadas para $\langle \Delta D_{\min G} \rangle \times R_G^{\text{proteína}}$, também são observados em cada uma das nove diferentes simulações, inclusive com comportamento linear na Região II (Figura 13a).

O fato de configurações alvos mais distendidas na Região III, terem as menores discrepâncias $\langle \Delta D_{\min G} \rangle$, pode ser atribuído aos contatos entre os monômeros nestes casos serem "contatos mais locais", ou mais próximos ao longo da sequência da cadeia (como em α -hélices). À medida que as estruturas protéicas e modelos se tornam mais compactas os

contatos passam a ter especificidades diversificadas, tornando cada vez mais difícil a determinação de semelhança entre configurações. Conseqüentemente as $\langle \Delta D_{\min G} \rangle^{\text{proteína}}$ e $\Delta D_{\min G}^{\text{modelo}}$, respectivamente, saem de 0,06 e 0,05 (Região III) indo a 0,15 e 0,11 (Região II) e atingindo ao redor de 0,16 e 0,12 para grandes compactações (Região I), pois nestas últimas os contatos entre os monômeros se tornam bastante específicos.

Nas regiões III das configurações alvos do modelo e proteínas (Figuras 12 e 13b) tem-se muitas $\Delta D_{\min G}$ em torno de 0,04, valores estes relativamente próximos (se considerados a escala de 0,0 a 1,0) da ΔD das estruturas retilínea “A” e hélice quadrada “G” ($\Delta D_{A,G} = 0,02$, Figura 3, Capítulo 2).

Na legenda da Figura 13 destacamos que as proteínas mais compactas na Região I, poderiam ser subdivididas entre aquelas com altas $\langle \Delta D_{\min G} \rangle$ maiores que 0,16 e as com baixas $\langle \Delta D_{\min G} \rangle$ menores que 0,15. Na Região II também haveria dois subgrupos de proteínas, com altas e baixas $\langle \Delta D_{\min G} \rangle$, respectivamente, aquelas maiores que 0,15 e as menores que 0,14. O intervalo de 0,01, ou 1%, para a separação entre altas e baixas $\langle \Delta D_{\min G} \rangle$ pode parecer precisão demasiada atribuída a um valor médio, mas há de se considerar dois aspectos: a) tais valores limiares são verificados, além de $\langle \Delta D_{\min G} \rangle$, também em muitas $\Delta D_{\min G}$ das proteínas nas nove simulações; b) a função discrepância $\Delta D_{\min G}$ é muito sensível, daí a necessidade de cautela em valores limiares fornecidos por ela. Estes limiares de altas e baixas $\langle \Delta D_{\min G} \rangle$ para as proteínas das Regiões I e II, indicam maiores ou menores dificuldades na semelhança entre proteínas e *random coil* e parecem não ser ao acaso, por isso serão analisados mais detalhadamente (Figura 15), ainda neste Capítulo.

Da caracterização das configurações *random coil*, quanto as suas compactações e contatos entre resíduos (seção 3.2), e também da comparação destas com as estruturas alvos protéicas (seção 3.3), já se pode verificar que tais configurações *random coil* ou “aleatórias” (ou pelo menos algumas destas) não são tão disformes, nem mesmo glóbulos fundidos (*molten globule*), pois estas configurações possuem muitas especificidades interacionais dos estados nativos protéicos, sendo tal suposição fortalecida tanto por algumas conclusões de resultados, já mostrados até aqui, quanto por outras expostas a seguir.

3.4 Resultados adicionais

As configurações *random coil* obtidas a partir do sistema cadeia-solvente apresentam características estruturais similares àquelas encontradas nas proteínas, ou mais provavelmente, as proteínas apresentam características conformacionais típicas das configurações *random coil*. A seguir destacamos estas características, ratificadas por outras evidências a fim de se entender de forma mais completa as relações entre estruturas protéicas e configurações *random coil*.

3.4.1 Os tamanhos dos resíduos influenciando nos valores de $\langle \Delta D_{\min G} \rangle$, obtidos na simulação



Os valores das $\langle \Delta D_{\min G} \rangle$, para as proteínas (Figura 13), além da especificidade dos contatos entre seus monômeros, também são consequências da diversidade de tamanhos e formas (especificidade estérica) dos resíduos, conforme estabelecida pela sequência dos mesmos nas correspondentes estruturas primárias. Este fator é provavelmente um dos principais responsáveis pela dificuldade “intrínseca” das configurações modelos conseguirem se aproximarem ainda mais das configurações nativas das proteínas, pois todos os monômeros são fixados como do mesmo tamanho. Para verificarmos tal suposição os aminoácidos são classificados simplificada e quanto aos seus tamanhos (Taylor, W. R. (1986), Zamyatin, A.A. (1972)), em pequenos e grandes (Figura 14), numa primeira aproximação.

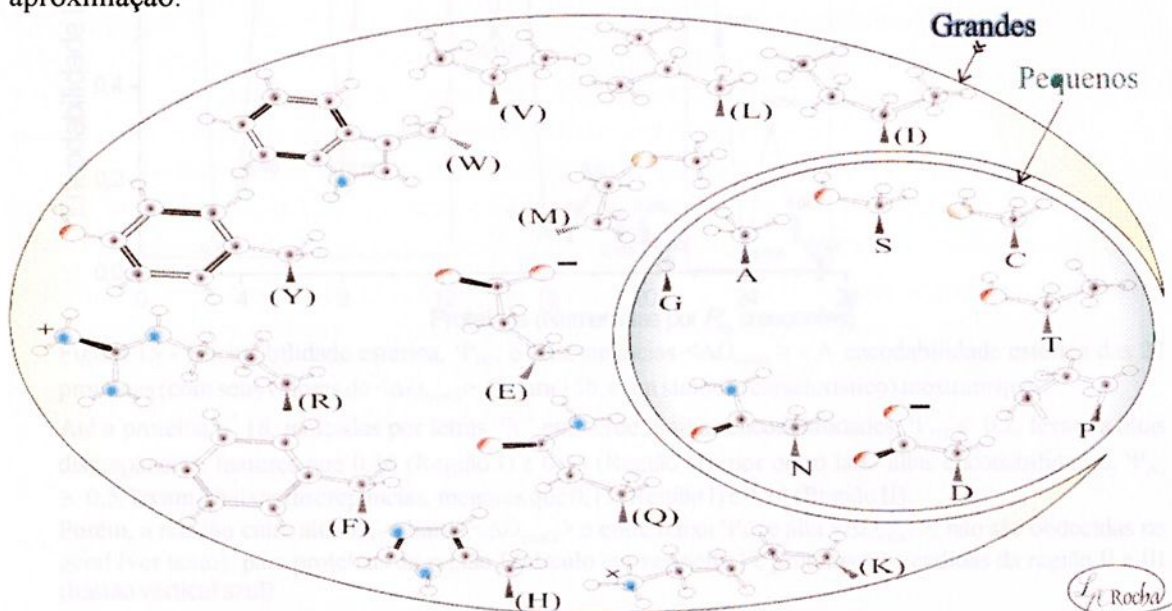


Figura 14 - Os aminoácidos, identificados pelo código de 1 letra (Capítulo 1) são representados aqui somente por suas cadeias laterais - Apesar da diversidade dos resíduos, quanto ao tamanho eles podem ser classificados apenas em pequenos P (num total de 9 resíduos, com volume de 60 (G) à 116 (T); elipse pequena) e grandes G (11 resíduos, com volume entre 138 (E) e 228 (W)). As cores dos átomos nos resíduos são as mesmas já utilizadas na Figura 4 do Capítulo 1.

Obs. Os resíduos não padrões (ou heterogrupos, ver Apêndice, após Capítulo 5), tiveram a seguinte classificação: Pequeno: Ace, NH₂, Dnp, Aba, Pca, Sin; Grande: lil.

Para se estabelecer correlação entre as variações dos valores de $\langle \Delta D_{\min G} \rangle$ para as proteínas nas diferentes Regiões I, II e III e a distribuição do tamanho de seus resíduos de aminoácidos (pequenos (P) e grandes (G)) em suas sequências, introduzimos a função encodabilidade (ou especificidade) estérica, Ψ_{PG} , para as 27 proteínas (Figura 15). A encodabilidade estérica, Ψ_{PG} , mede o nível de capacidade de registro de informação, dada pela distribuição dos resíduos pequenos N_P , e dos grandes N_G , na sequência de uma proteína (sendo $N_P + N_G = 35$ resíduos). A distribuição “equalizada” corresponde a 17 e 18 resíduos de cada tipo, se referindo à máxima capacidade de “codificação” da cadeia, quanto à diversidade de resíduos grandes e pequenos. Assim idealizamos a função Ψ_{PG} , dada por:

$$\Psi_{PG} = (17!18!)/(N_P!N_G!) \quad (6)$$

Quanto mais próximas os valores das quantidades N_P e N_G , na sequência protéica, mais Ψ_{PG} se aproxima de 1,0, no caso ideal, se $N_P = 17$ e $N_G = 18$, então $\Psi_{PG} = 1$. Baixas

encodabilidades, e.g., $\Psi_{PG} \leq 0,2$, indica o predomínio de um dos tipos de tamanho de resíduos sobre o outro tipo e conseqüente distanciamento da distribuição equalizada.

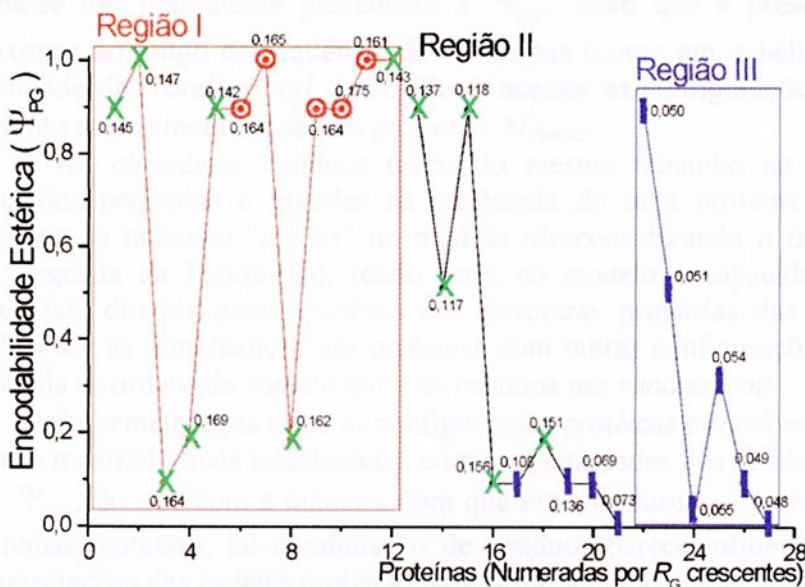


Figura 15 - Encodabilidade estérica, Ψ_{PG} , e Discrepâncias $\langle \Delta D_{minG} \rangle$ - A encodabilidade estérica das 27 proteínas (com seus valores de $\langle \Delta D_{minG} \rangle$, Figura 13b, e um símbolo característico) mostram que:

Até a proteína n° 18, realçadas por letras "X" em verde: baixas encodabilidades, $\Psi_{PG} \leq 0,2$, levam a altas discrepâncias, maiores que 0,16 (Região I) e 0,15 (Região II); por outro lado altas encodabilidades, $\Psi_{PG} \geq 0,5$, levam à baixas discrepâncias, menores que 0,15 (Região I) e 0,14 (Região II).

Porém, a relação entre alta Ψ_{PG} e baixa $\langle \Delta D_{minG} \rangle$ e entre baixa Ψ_{PG} e alta $\langle \Delta D_{minG} \rangle$, não são obedecidas no geral (ver texto): para proteínas da região I (círculo em vermelho) e proteínas distendidas da região II e III (bastão vertical azul).

Obs. - A proteína n° 14 (1zdc), cuja configuração nativa tem uma forma bastante peculiar, na forma de "U" também apresenta duas características, capturadas pelas grandezas utilizadas até aqui, sendo estas

- É a única proteína nas Regiões I e II que possui Ψ_{PG} (maior que 0,2 e menor que 0,8), ou seja, com composição de aminoácidos P e G, bastante diferente das demais.
- esta mesma proteína também tem ordem de contato bastante peculiar (Figura 9).

Como as $\langle \Delta D_{minG} \rangle$ das 27 proteínas apresentam comportamentos variados, dentro de cada uma das Regiões de compactações (I, II e III na Figura 13) é adequado avaliar a relação entre Ψ_{PG} e $\langle \Delta D_{minG} \rangle$ nestas regiões (Figura 15), sendo

- Para proteínas das Regiões I e II (até proteína n° 16 e a n° 18), baixas $\Psi_{PG} (\leq 0,2)$ levam a altas $\langle \Delta D_{minG} \rangle$ e altas $\Psi_{PG} (\geq 0,5)$ levam a baixas $\langle \Delta D_{minG} \rangle$ (letras "X" em verde), pois em baixa Ψ_{PG} a distribuição não equalizada dos N_P , N_G resíduos implicando numa dificuldade intrínseca das cadeias *random coil*, com seus resíduos todos de mesmo tamanho, conseguirem se aproximar demasiadamente das configurações nativas; por outro lado em altos valores de Ψ_{PG} a distribuição equalizada facilita o acesso as configurações nativas das proteínas, daí menores valores de $\langle \Delta D_{minG} \rangle$.

- A relação entre alta Ψ_{PG} e baixas ΔD_{minG} e entre baixas Ψ_{PG} e altas ΔD_{minG} não são verificadas, respectivamente para proteínas da Região I (círculos em vermelho) e proteínas muito distendidas das regiões II e III (bastões verticais em azul), pois no primeiro caso outros fatores mais influentes que os arranjos das quantidades de resíduos P e G

(medidos por Ψ_{PG}) podem estar presentes nestas proteínas, como a complexidade estrutural, característica das proteínas da Região I. No segundo caso a baixa compactação torna-se um ingrediente precedente à Ψ_{PG} , visto que a presença de muitos contatos próximos ao longo da sequência das proteínas (como em α -hélices), levam a uma maior facilidade das *random coil* conseguirem acessar as configurações nativas das proteínas e portanto naturalmente já se tem pequenas $\Delta D_{\min G}$.

Ao considerar resíduos todos do mesmo tamanho no modelo, os resíduos de tamanhos pequenos e grandes na sequência de uma proteína ficam “modulados” por resíduos de tamanho “médio” no modelo (desconsiderando o fator de proporcionalidade 1,8, legenda da Figura 10), tendo estes no modelo a capacidade de capturar aspectos essenciais ditados pelos resíduos nas estruturas primárias das proteínas. Tais aspectos podem ser as semelhanças das proteínas com outras configurações, como as *random coil*, ou ainda as ordens de contato entre os resíduos nas *random coil*.

As semelhanças entre as configurações protéicas e *random coil*, dada por $\langle \Delta D_{\min G} \rangle$, tem se mostrado mais relacionadas com as quantidades dos resíduos do tipo P e G, medida por Ψ_{PG} , do que com a maneira com que estes resíduos se combinam, nas sequências das proteínas; contudo, tal combinação de resíduos parece influenciar significativamente as compactações das cadeias protéicas, conforme será mostrado, em breve, na seção 3.4.4.

As conclusões tiradas a partir da Figura 15 reforçam a relação entre a estrutura terciária, medida por $\langle \Delta D_{\min G} \rangle$, e a estrutura primária, expressa por Ψ_{PG} .

3.4.2 Algumas implicações decorrentes de valores relativamente estáveis de $\langle \Delta D_{\min G} \rangle$

Os valores das discrepâncias mínimas globais, $\Delta D_{\min G}$ (Figura 13a), se mostrados individualmente para as 27 proteínas, nas nove simulações, podem evidenciar mais precisamente em quanto e como são as máximas semelhanças entre as proteínas e as *random coil* (Figura 16).

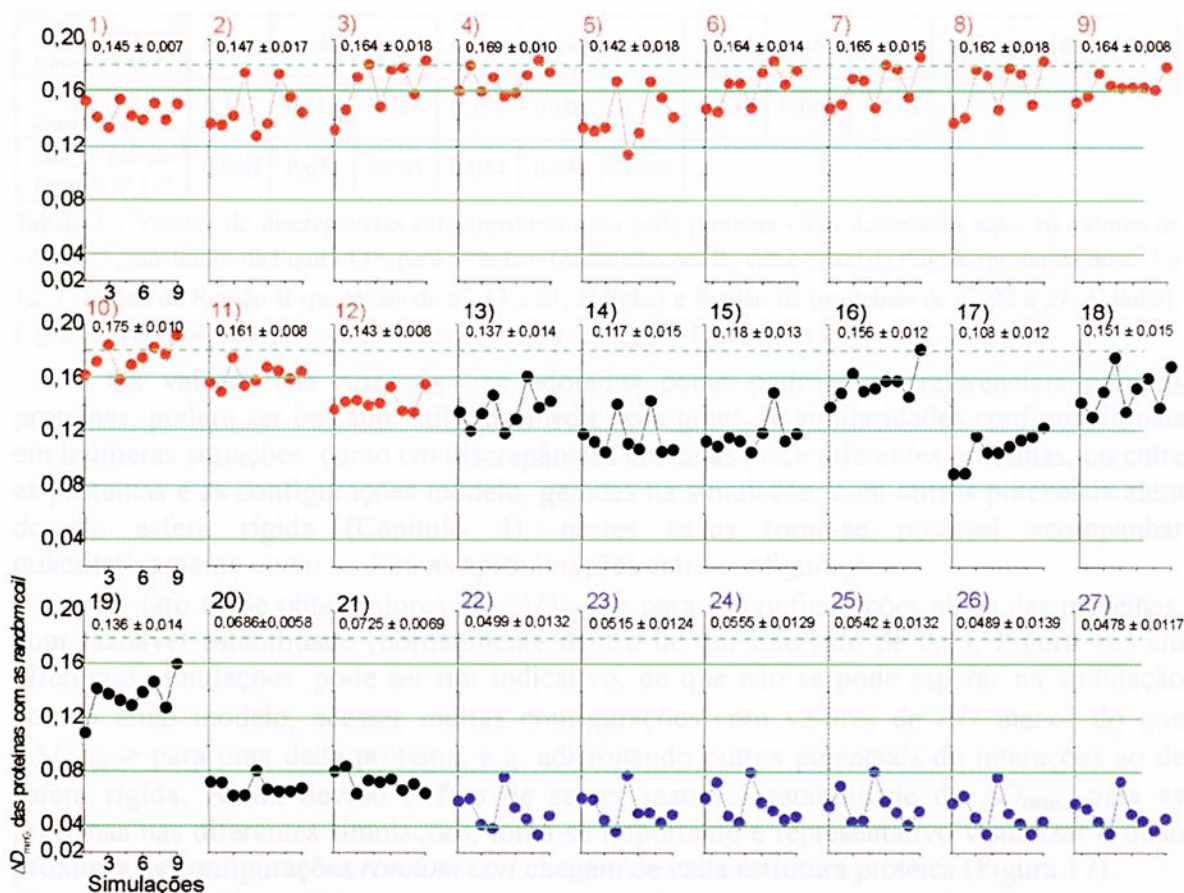


Figura 16 - Discrepâncias mínimas globais, $\Delta D_{\min G}$, para as 27 proteínas em nove simulações independentes - A sequência dos gráficos segue da proteína mais compacta (1c2u, n° 1) para a menos compacta (1peh, n° 27) nas Regiões I (círculos vermelhos), II (círculos pretos) e III (círculos azuis), Figura 13a. Acima de cada gráfico tem-se os n°s (de 1 a 27) das proteína e abaixo estão as médias ($\langle \Delta D_{\min G} \rangle$, Figura 13b) e os desvios padrões de $\langle \Delta D_{\min G} \rangle$ entre as proteínas e as *random coil* nas diferentes simulações.

Nota-se que os valores das discrepâncias mínimas globais, $\Delta D_{\min G}$, em geral, estão variando, para cada uma das três regiões, dentro de uma faixa (mínima e máxima) de até 0,03 e isso torna possível caracterizar cada proteína, pelo valor médio da discrepância mínima global, $\langle \Delta D_{\min G} \rangle$ com razoável confiabilidade.

Os valores das $\langle \Delta D_{\min G} \rangle$ obtidas para as configurações alvos protéicas (Figura 13b) podem ser consideradas referenciais para estas proteínas (Tabela 1), pois foram adquiridos dentre um vasto espaço configuracional, disponível para a cadeia do modelo, além de apresentarem pouca variações (normalmente dentro de 0,03) ao redor de cada um dos valores médios nas nove simulações realizadas. O mesmo poderia ser feito no caso das configurações alvos do modelo (Figura 12) bastando para isto se ter, o valor da compactação da cadeia.

$\Delta D_{\min G}$ para proteínas de 1 a 12	0,145	0,147	0,164	0,169	0,142	0,164	0,165	0,162	0,164	0,175	0,161	0,143
$\Delta D_{\min G}$ para proteínas de 13 a 21	0,137	0,117	0,118	0,156	0,108	0,151	0,136	0,069	0,073			
$\Delta D_{\min G}$ para proteínas de 22 a 27	0,050	0,051	0,055	0,054	0,049	0,048						

Tabela 1 - Valores de discrepâncias característicos para cada proteína - são destacados aqui, os valores de $\langle \Delta D_{\min G} \rangle$, mostrados da Figura 13b, para proteínas (ordenadas por R_G crescentes) da Região (proteínas de n^{os} 1 a 12, 1^a linha), da Região II (proteínas de n^{os} 13 a 21, 2^a linha) e Região III (proteínas de n^{os} 22 a 27, 3^a linha). E.g. a proteína n^o 4, ltsk, passa a ser caracterizada por $\langle \Delta D_{\min G} \rangle = 0,169 \cong 0,17$.

Os valores das $\langle \Delta D_{\min G} \rangle$ se adotados como padrões ou referenciais para as proteínas, poderá ser bastante útil para medir seus graus de similaridades configuracionais em inúmeras situações, como em discrepâncias cruzadas entre diferentes proteínas, ou entre as proteínas e as configurações modelo, geradas na simulação, com outros potenciais além do de esfera rígida (Capítulo 4), nestes casos torna-se possível acompanhar quantitativamente como se dará as aproximações entre configurações.

O fato de se obter valores de $\langle \Delta D_{\min G} \rangle$, para as configurações alvos das proteínas, com razoável estabilidade (normalmente dentro de um intervalo de 0,03, Figura 16) em diferentes simulações, pode ser um indicativo, de que não se pode esperar na simulação com o atual modelo, acessar muitas configurações com valores de ΔD menor do que $\langle \Delta D_{\min G} \rangle$ para uma dada proteína, e.g. adicionando outros potenciais de interações ao de esfera rígida. Ainda devido o fato de se ter razoável estabilidade de $\Delta D_{\min G}$ para as proteínas nas diferentes simulações, torna-se importante e representativo visualizar o quão próximas as configurações *random coil* chegam de cada estrutura protéica (Figura 17).

As discrepâncias de simulação ΔD entre as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida, foram calculadas para as proteínas da Região I (proteínas de n^{os} 1 a 12), da Região II (proteínas de n^{os} 13 a 21) e da Região III (proteínas de n^{os} 22 a 27). Os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida, e os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida. Os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida. Os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida.

Observando as semelhanças e diferenças entre as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida, podemos observar que as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida, apresentam valores de ΔD muito próximos, o que indica que as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida, são muito semelhantes.

Os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida, e os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida. Os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida. Os valores de ΔD foram calculados para as configurações alvo e as configurações modelo geradas com o potencial de esfera rígida.

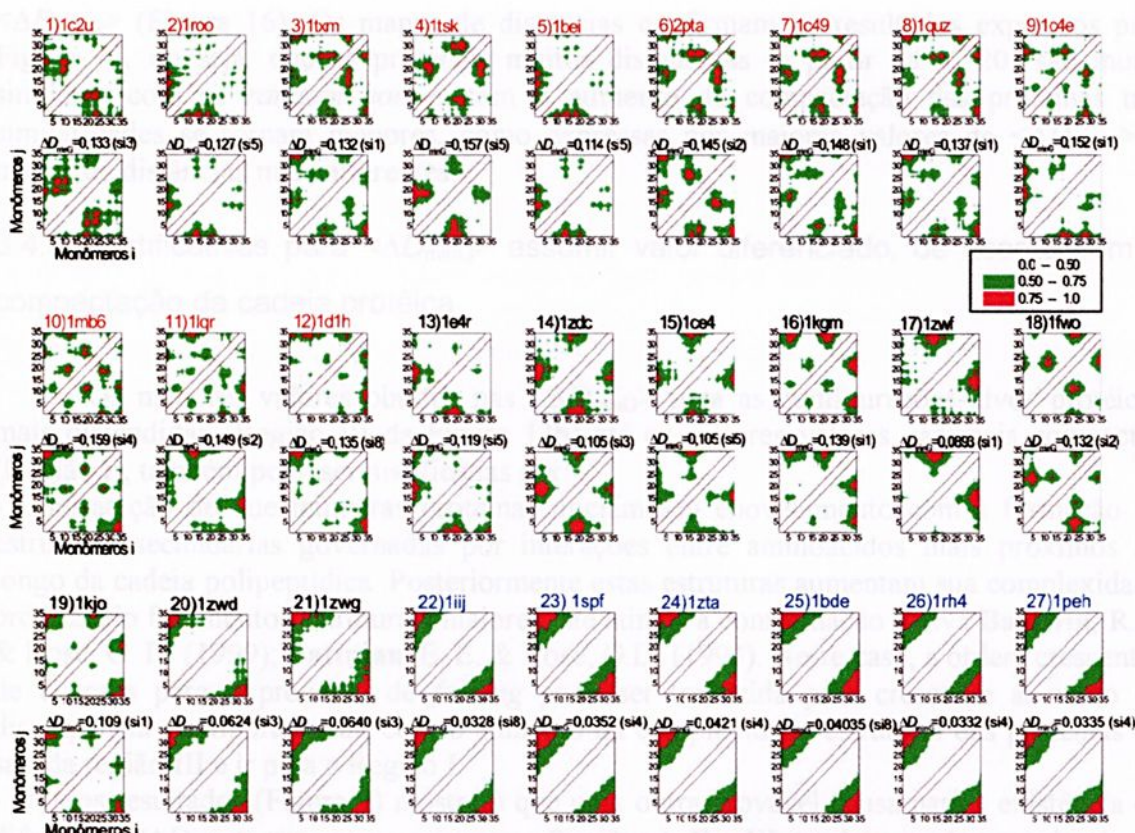


Figura 17 - Mapas de distâncias das 27 configurações proteicas e correspondentes *random coil*, com a menor das discrepâncias mínimas globais $\Delta D_{\min G}$, dentre as nove simulações - As distâncias entre os resíduos estão normalizadas (entre 0 e 1) e mostradas em cores branca, verde e vermelha, respectivamente para distâncias menores que 0,50; de 0,50 a 0,75 e maiores que 0,75 nas proteínas (acima, em linhas ímpares) e *random coil* (abaixo, nas linhas pares). A cor branca ao longo da diagonal principal é devido a exclusão das distâncias entre mesmos resíduos (i com i) e primeiros e segundo vizinhos (i com $i-1$, $i-2$, $i+1$, $i+2$). Acima de cada proteína consta seu número (por compactação decrescente), seu nome; acima de cada *random coil* consta a discrepância, $\Delta D_{\min G}$ com sua proteína e em qual simulação (de 1 a 9) ocorreu tal discrepância. Observando cada par de configuração proteína e *random coil*, por colunas, é possível identificar similaridades (mesma cores) e diferenças (cores diferentes) entre elas. Os mapas de distâncias, acima, confirmam que quanto menor o valor da discrepância maiores as semelhanças configuracionais.

Observando as semelhanças e diferenças entre as configurações proteicas e *random coil* pelos mapas de distâncias dos resíduos, em três cores, notam-se grandes semelhanças (representadas pelas mesmas cores), inclusive em resíduos distantes (pontos em vermelho) indicando que nas *random coil* existe muita especificidade interacional, sendo esta tão característica e fundamental nas estruturas nativas proteicas.

Nos mapas de distâncias da Figura 17 foi privilegiado, mesmo com um modelo tão simplificado, observar as distâncias grandes (acima de 0,5, dentre um intervalo de 0 a 1), as mais difíceis de serem idênticas entre duas configurações. É pouco freqüente (e indesejável) ocorrerem casos de pares de resíduos com longas distâncias nas proteínas (região em branco) estarem muito próximos (em vermelho) no modelo e vice-versa. Comumente se têm mesmas cores ou ainda (sendo tolerável) cores próximas (branco com verde ou verde com vermelho) nas duas configurações comparadas. Os mapas de distâncias das *random coil*, para cada proteína, costumam apresentar basicamente os mesmos comportamentos descritos aqui, nas nove simulações, devido a razoável estabilidade de

$\langle \Delta D_{\min G} \rangle$ (Figura 16). Os mapas de distâncias confirmam os resultados expressos pela Figura 13, ou seja, que as proteínas muito distendidas (a partir da n° 20) são muito similares com as *random coil* e com o aumento da compactação das proteínas tais similaridades se tornam menores, como expressas por maiores valores de $\langle \Delta D_{\min G} \rangle$ e mapas de distâncias mais diferentes.

3.4.3 Justificativas para $\langle \Delta D_{\min G} \rangle$ assumir valor diferenciado, de acordo com a compactação da cadeia protéica

Os menores valores obtidos nas $\langle \Delta D_{\min G} \rangle$ para as configurações alvos protéicas mais distendidas (Região III da Figura 13b) até os maiores valores nas mais compactas (Região I), também pode ser justificadas por:

- Constatação de que inúmeras proteínas iniciam seu enovelamento com a formação de estruturas secundárias governadas por interações entre aminoácidos mais próximos ao longo da cadeia polipeptídica. Posteriormente estas estruturas aumentam sua complexidade produzindo fragmentos estruturais maiores, até atingir a conformação nativa **Baldwin, R.L. & Rose, G. D. (1999); Lattman, E. E. & Rose, G.D. (1993)**. Neste caso, a ordem crescente de eventos para o processo de *folding* pode ser traduzida pelo crescente aumento da discrepância configuracional, com o aumento da complexidade estrutural das proteínas ao sair da região III e ir para a Região I.

- Nossos resultados (Figura 9) mostram que uma outra provável causa para a existência de diferentes $\langle \Delta D_{\min G} \rangle$ para as proteínas nas Regiões I, II e III, também pode ser oriunda da dependência da taxa de enovelamento das proteínas com parâmetros topológicos da estrutura nativa, como a ordem de contato, pois estruturas com baixas ordens de contato ($\chi \leq 0,2$), por serem muito freqüentes (existem em todos os valores de raios de giração), se enovelam mais facilmente do que aquelas com altas ordens de contato, por serem mais raras e dependentes das compactações e das simulações, tal hipótese foi comprovada ao se verificar as ordens de contato das estruturas *random coil*, em diferentes simulações (Figura 9).

3.4.4 Dependência da compactação de uma proteína com o tamanho de seus resíduos

O Paradoxo de Levinthal (**Honig, B. (1999); Levinthal, C. (1968)**) afirma ser altamente improvável, por busca aleatória de configurações, acessar o estado nativo de uma proteína num tempo finito, pois tal busca, mesmo para proteínas pequenas (como as com 35 resíduos, seção 1.6.3 do Capítulo 1), o tempo característico para se encontrar a nativa seria superior ao da idade do Universo. Nossos resultados têm mostrado que existem configurações *random coil* acessadas, num tempo finito e relativamente pequeno (abaixo de 600 kPMC, mais detalhes serão dados na seção 3.4.5 abaixo) com muitas características e semelhanças às das proteínas reais. É importante destacar que parte do tempo, em cada uma das nove simulações é gasto gerando estruturas com compactação distante da compactação de uma proteína específica; assim procedemos para cobrir a ampla faixa de compactações de todas as proteínas de 35 resíduos.



Se uma dada proteína, simplesmente pelo tamanho de seus resíduos em sua seqüência, consegue (ou ter influência suficiente para) acessar somente as configurações com compactações próximas da sua, na prática, a busca aleatória fica bastante facilitada, reduzindo em muito o tempo para ser ter boa semelhança entre a proteína com as *random coil*. Para verificar se uma proteína faz acesso às compactações próximas daquelas esperadas para ela, novamente aproveitaremos a simplificação de se ter os resíduos de aminoácidos, nos tamanhos pequenos P e grandes G (mostrados na Figura 15) e faremos a seguinte hipótese, um tanto quanto razoável (que poderá ou não ser confirmada pelas proteínas):

Se as proteínas tiverem muitos resíduos P e G em suas seqüências formando clusters, tendo estes três ou mais resíduos consecutivos (com dois resíduos os clusters podem ser casuais ou pouco influentes) do mesmo tipo P ou G, então se reduz em muito a possibilidade de conseguir acessar configurações compactas, pois tais proteínas provavelmente terão estruturas nativas mais distendidas, com baixa compactação. Se a existência de clusters de resíduos P e G for reduzida o acesso a estruturas compactas fica facilitado, pois tais estruturas podem ser traduzidas por grandes diversidades e combinações de tamanhos P, G, tendo assim preferencialmente poucos clusters, já que desta forma os resíduos encaixam-se adequadamente e formam configurações mais compactas (Schulz, G. E. & Schirmer, R. H. (1990)). O encaixe de muitos resíduos P ou G em clusters dificulta o acesso a estruturas compactas e muda a dinâmica de formação destas estruturas. Espera-se assim que com poucos clusters tenham-se estruturas mais compactas e com o aumento da quantidade de clusters tais estruturas fiquem mais distendidas. Pode-se supor também que resíduos P em clusters (a não ser em grandes quantidades) sejam menos influentes na compactação do que os de resíduos G em clusters, devido a maior facilidade de encaixe entre os primeiros.

Para comprovar a hipótese acima foi verificada a quantidade de clusters de resíduos grandes G (Figura 18a) e em seguida a quantidade de clusters de resíduos pequenos P (Figura 18b).

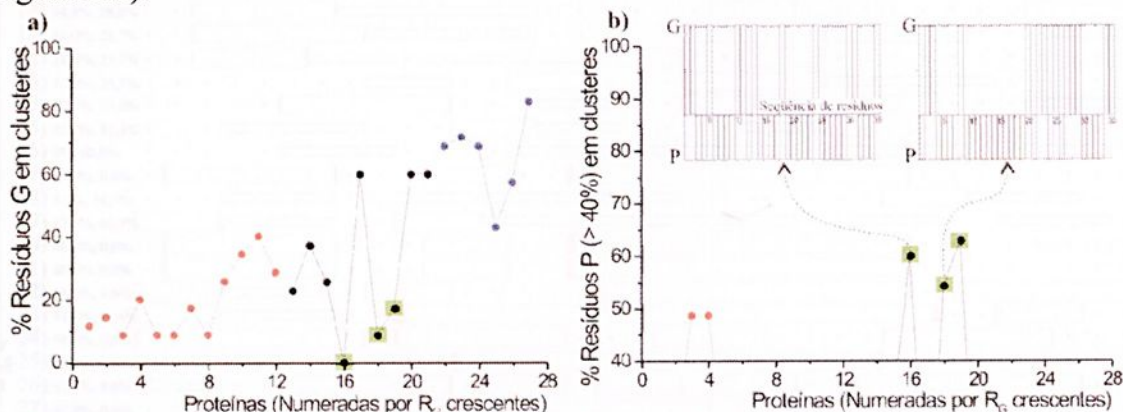


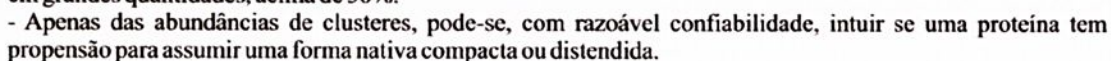
Figura 18 - Quantidades de resíduos G e P em Clusters e Compactações das cadeias proteicas:

a) Quantidade percentual de resíduos grandes (G) formando clusters nas 27 proteínas (nas Regiões I, II e III). São destacadas as três proteínas (quadrados verdes) com quantidades de clusters abaixo do esperado, em relação as outras proteínas de compactações similares.

b) Quantidades percentuais, acima de 40%, de resíduos pequenos (P) em clusters. As três proteínas (quadrados verdes, item a), apresentam grandes quantidades de resíduos P em clusters, assim gerando proteínas menos compactas. São destacadas a distribuição dos resíduos P e G para as proteínas de n^{os} 16 e 18.

Obs. - Nas proteínas, de n^{os} 1 a 12, Região I, a exceção das n^{os} 3 e 4, as quantidades de resíduos em clusters de resíduos P ficam ao redor de 20%, resultados estes não mostrados aqui.

As quantidades de resíduos de aminoácidos em clusters têm se mostrado significativamente influente para auxiliar as compactações adequadas das proteínas (Figura 19), pois nota-se com o aumento progressivo de resíduos em clusters (a partir da proteína n° 9) a diminuição da compactação. Por outro lado com a diminuição das quantidades de resíduos G, abaixo de 20%, em clusters, aumentam-se as compactações das proteínas, conforme notamos nas proteínas de de n°s 1 a 8.



Pode-se afirmar que a hipótese levantada acima da dependência da compactação, com as quantidades de resíduos em clusters foi confirmada, com algumas ressalvas e exceções (aceitáveis para um sistema complexo, como o proteína-solvente), por exemplo, não se explicam as baixas compactações das proteínas de n^{os} 3 e 4, mesmo com grandes quantidades (acima de 40%) de resíduos P em clusters, mas estas quantidades, podem confirmar as altas discrepâncias destas proteínas (Figura 15) e também ratificar que os clusters de resíduos P são menos influentes na determinação da compactação da proteína do que os resíduos G em clusters.

De maneira geral, pode-se afirmar que a combinação dos resíduos em clusters com os resíduos isolados parece atuar sincronizadamente para determinar as compactações das proteínas, tendo estruturas compactas, grande diversidade e combinações de tamanhos e com a diminuição destas diversidades e combinações as proteínas se tornam progressivamente mais distendidas.

A semelhança entre as estruturas protéicas e *random coil*, medidas por $\langle \Delta D_{\min G} \rangle$, parecem sofrer pouca influência da quantidade de resíduos em clusters, mas a quantidade de resíduos G em clusters, acima de 25%, nas proteínas de n^{os} 9, 10 e 11 talvez possa influenciar suas altas discrepâncias $\langle \Delta D_{\min G} \rangle$, o que não foi possível justificar somente pelos valores da encodabilidade estérica (Figura 15, destacada por círculos em vermelho). Já para as proteínas de n^{os} 6, 7 resta apenas justificar suas altas $\langle \Delta D_{\min G} \rangle$ como decorrência de suas altas complexidades estruturais (Figura 20).

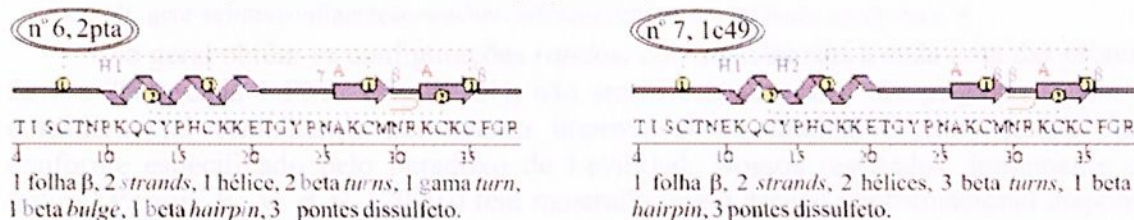


Figura 20 - Visualização das estruturas secundárias nas seqüências de resíduos das proteínas n^{os} 6, 7 e abaixo estão os resumos de sua quantidades e tipos de estruturas secundárias e pontes dissulfeto.

Obs. As proteínas n^{os} 6 e 7 são bastante similares, com 97,1% de homologia (Figura 8, do Capítulo 2).

As composições de estruturas secundárias nas configurações nativas das proteínas de n^{os} 6 e 7, podem gerar alta complexidade de estrutura e conseqüentemente influenciando significativamente nos altos valores de $\langle \Delta D_{\min G} \rangle$, sendo estes não justificados somente por suas encodabilidades estéricas.

3.4.5 Paradoxo de Levinthal versus Semelhanças entre estruturas protéicas e configurações *random coil*

Independente se as proteínas, por meio do tamanho de seus resíduos, conseguem acessar configurações cujas compactações são próximas de suas estruturas nativas (seção 3.4.4), podemos quantificar os acessos das *random coil* à estrutura nativa de cada proteína, verificando ao longo de cada uma das nove simulações quantas vezes (Figura 21) as estruturas *random coil* tiveram discrepâncias, ΔD , abaixo de α ($= \langle \Delta D_{\min G} \rangle + 0,02$).

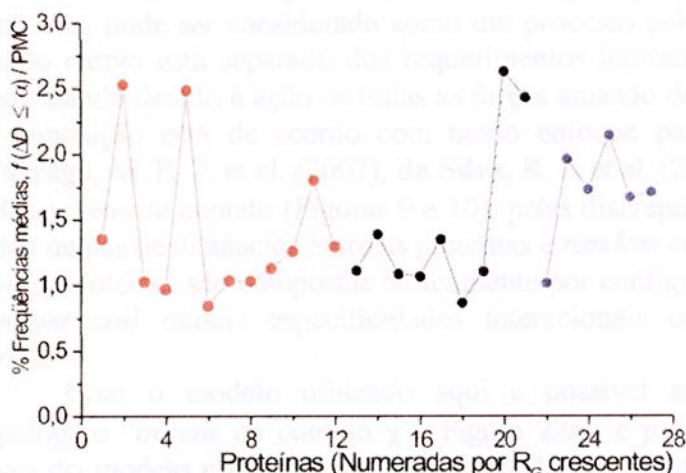


Figura 21 - Semelhanças entre estruturas alvo protéica e *random coil*, por meio das percentagens das frequências de ΔD abaixo de α ($=\langle D_{\text{média}} \rangle + 0,02$), para cada proteína e a cada 1 Passo Monte Carlo (PMC)- Com um pequeno incremento de 0.02 em $\langle \Delta D_{\text{média}} \rangle$ (Tabela 1) ainda é possível observar muitas semelhanças entre as estruturas em questão, próximas das observadas nos mapas de distâncias da Figura 17. Os valores de $f(\Delta D \leq \alpha)$ para as proteínas podem ter grandes variações nas diferentes simulações, dependendo se $\Delta D \leq \alpha$; por isso é apresentado um valor médio de frequências nas nove simulações.

E.g. no caso da proteína nº 4 (1tsk), há 1% de chance de ter, $\Delta D \leq \alpha$, a cada 1 PMC, ou seja, em média a cada 100 PMC gera-se uma configuração *random coil* razoavelmente semelhante a proteína nº 4.

No geral obtêm-se configurações *random coil* semelhantes a cada uma das proteínas de 1 a 2% a cada 1 PMC (Figura 21), não sendo este um valor tão pequeno a ponto de considerar tal semelhança um evento improvável ou inatingível num tempo finito, conforme especificado pelo paradoxo de Levinthal. Nossos resultados, juntamente com outros (Pappu, R. V. et al. (2000)) têm mostrado que o espaço conformacional disponível para a cadeia protéica é menor do que previamente se acreditava devido a influência das interações estéricas, como evidenciado em nosso caso, com a formação de clusters de resíduos grandes e pequenos, que podem facilitar os acessos a configurações nativas protéicas.

3.5 Comentários

Pelos resultados, deste Capítulo, é possível constatar que a interação de esfera rígida, mesmo sendo não específica e não estabilizadora, atua de forma a levar as configurações *random coil*¹ a acessarem conformações muito próximas das estruturas nativas das proteínas, isso para todas as proteínas estudadas neste trabalho.

¹...“no Universo descrito pelas ciências modernas, o aleatório tem um papel cada vez maior. Primeiramente apareceu a visão probabilística, nas tentativas de justificação microscópica da entropia: é a importante obra de Boltzmann. Depois, veio o golpe da mecânica quântica. Apesar das inúmeras tentativas de retornar à ortodoxia determinista, as probabilidades continuam a desempenhar nesta teoria um papel central. Em nossa escala de seres vivos agregados macroscópicos, parecia que a lei dos grandes números poderia restabelecer na prática o esquema determinista. Mas eis que as probabilidades fazem sua entrada a força, mesmo neste mundo: é esse um dos aspectos da descoberta da auto-estruturação dos sistemas macroscópicos longe do equilíbrio. O que os relaciona ao aleatório deve-se a variedades das formas que, de uma experiência a outra, esse mecanismos de estruturação podem desenvolver, a despeito do rigoroso controle das condições experimentais. Não se trata mais de fenômenos calculáveis por meios de leis gerais. Próximo ao equilíbrio as leis da natureza são universais; longe do equilíbrio, elas são específicas. Essas “instabilidades” exigem um fluxo de energia, elas dissipam energia”...Ilya Prigogine (Pessia-Paste, G. (1992))

Somente com o potencial de esfera rígida já é possível mostrar que o *folding* de proteínas, pode ser considerado como um processo pelo qual o mecanismo para acessar o estado nativo está separado dos requerimentos interacionais que garantem a estabilidade neste estado devido à ação de todas as forças atuando de forma efetiva e cooperativamente; tal suposição está de acordo com nosso enfoque para o processo de *folding* protéico (Tarragó, M. E. P. et al. (2003), da Silva, R. A et al. (2001)). Também é possível verificar, pelas ordens de contato (Figuras 9 e 10), pelas discrepâncias $\langle \Delta D_{\min G} \rangle$ (Figuras 12 e 13) e pelos mapas de distâncias entre as proteínas e *random coil* (Figura 17) que as configurações (*folds*) protéicas são compostas basicamente por configurações *random coil*, podendo ter as *random coil* muitas especificidades interacionais características dos estados nativos protéicos.

Com o modelo utilizado aqui é possível antecipar, por meio do parâmetro topológico “ordens de contato χ ” (Figura 22a), e pelas discrepâncias entre as estruturas alvos do modelo e *random coil* (Figura 22b) a existência de três Regiões com diferentes compactações e diferentes similitudes, $\langle \Delta D_{\min G} \rangle$, entre configurações das proteínas reais e configurações *random coil*. Por outro lado com as discrepâncias cruzadas $\langle \Delta D_{\min G} \rangle$ também é plausível fazer comparações (Figura 22c) com as auto discrepâncias máximas, $\Delta D_{\max}$ (Tabela 1 do Capítulo 2) entre as diversas estruturas de uma mesma proteína, depositadas no “PDB” e determinadas experimentalmente por ressonância magnética nuclear.

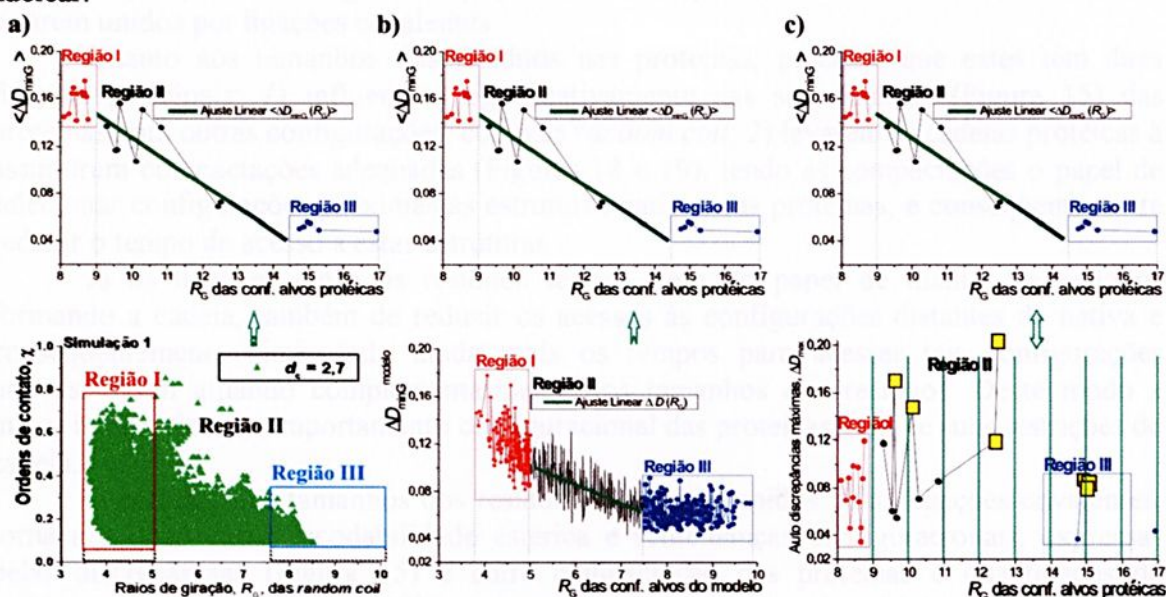


Figura 22 - As variações das discrepâncias cruzadas, $\langle \Delta D_{\min G} \rangle$, entre as configurações alvos protéicas e *random coil* (1ª linha dos gráficos acima) podem ser previstas qualitativamente, nas Regiões I, II e III, por

a) Ordens de contato em função da compactação (R_G) das estruturas *random coil* (Figura 10);

b) Discrepâncias, $\Delta D_{\min G}$, entre as configurações alvos do modelo e *random coil* (Figura 12).

Tais discrepâncias têm valores que

c) no geral, são superiores as auto discrepâncias máximas, $\Delta D_{\max G}$ (Tabela 1 do Capítulo II), mas algumas proteínas destacadas por um pequeno quadrado em amarelo, no gráfico das auto discrepâncias, tem $\Delta D_{\max}$ superiores as $\langle \Delta D_{\min G} \rangle$, sendo estas as proteínas de n° 15, 17, 20, 21, 23, 24, 25.

Obs. A proteína n° 26 ($R_G=15,31$) tem sua estrutura determinada por Raio X, daí sua ausência da Figura 22c.

As configurações alvos, tanto do modelo como das proteínas, têm suas semelhanças com as *random coil* (Figura 22b) dependentes da compactação das configurações alvos, nas Regiões I, II e III; tais resultados mostram que comparações configuracionais são mais significativas se realizada entre estruturas de compactações semelhantes ou próximas. Assim, uma configuração compacta deve ser comparada com outra pelo menos razoavelmente compacta, já uma configuração mais distendida deve ser comparada com outras também razoavelmente distendidas. Tal conclusão será empregada nas comparações configuracionais do Capítulo 4. Portanto, um mesmo grau (ou valor) de discrepância entre configurações assume diferente significado para proteínas de diferentes compactações.

Comparando os valores de $\langle \Delta D_{\min G} \rangle$ e $\Delta D_{\max}$ (Figura 22c) é possível verificar em metade das proteínas menos compactas (nas Regiões II e III, ou de 7 dentre 14), que suas $\langle \Delta D_{\min G} \rangle$ são menores que as auto discrepâncias $\Delta D_{\max}$, ou seja, as configurações *random coil* se mostram mais semelhante com estas proteínas, do que quando comparação entre algumas de suas estruturas obtidas experimentalmente são realizadas.

Os resultados mostrados, neste Capítulo, também evidenciam que o acesso às configurações nativas das proteínas, se dá, devido a dois ingredientes que a cadeia protéica carrega independente de quaisquer outras condições (de si ou dependências com o meio ao seu redor), sendo estes comum a todas as proteínas, independentes de seu tipo e função biológica, são eles: (a) o fato dela ter, em sua estrutura primária, resíduos de diversos tamanhos (ocupando um lugar no espaço com a condição de volume excluído) e (b) destes estarem unidos por ligações covalentes.

Quanto aos tamanhos dos resíduos nas proteínas, parecem que estes têm duas funções principais: 1) influenciar significativamente nas semelhanças (Figura 15) das proteínas com outras configurações, como as *random coil*; 2) levarem as cadeias protéicas a assumirem compactações adequadas (Figuras 18 e 19), tendo as compactações o papel de selecionar configurações próximas às estruturas nativas das proteínas, e conseqüentemente reduzir o tempo de acesso a estas estruturas.

Já as ligações entre os resíduos teriam além do papel de manter os resíduos formando a cadeia, também de reduzir os acessos às configurações distantes da nativa e conseqüentemente diminuindo ainda mais os tempos para acessar tais configurações nativas, assim atuando complementarmente aos tamanhos dos resíduos. Deste modo a maior influência no comportamento configuracional das proteínas vem de suas restrições de cadeia.

O enfoque nos tamanhos dos resíduos, que estão unidos pelas ligações covalentes, torna a relação entre encodabilidade estérica e semelhanças configuracionais, expressas pelas discrepâncias (Figura 15) e entre compactação das proteínas e quantidades de resíduos em clusters (Figura 18) em concordância com a premissa básica do processo de *folding* protéico, ou seja, que a seqüência dos resíduos unidos pelas ligações covalentes, determinam a estrutura tridimensional, ou forma, da proteína. Por exemplo, no caso das quantidades de resíduos em clusters (Figura 18) é possível diferenciar as proteínas com estruturas compactas das distendidas, pois todas as proteínas distendidas possuem altas quantidades de resíduos em clusters, em geral acima de 50%; por outro lado as estruturas compactas apresentam quantidades em torno de 10% e 20%, respectivamente para os resíduos G e P em clusters, salvo aquelas com estruturas muito peculiares, como a ltsk.



3.6 Bibliografia

- Baldwin**, R. L. & Rose, G. D. Is protein folding hierarchic? I. Local structure and peptide folding, *TIBS* **24**, 26-33 (1999).
- Bonneau**, R., I. Ruczinski, J. Tsai, and D. Baker, Contact order and ab initio protein structure prediction, *Protein Science* **11**, 1937-1944 (2002).
- Campbell**, M. K. em: *Bioquímica*, 3ª edição ARTMED, Capítulo 4: A Estrutura Tridimensional das Proteínas (2000).
- Cantor**, C. R. & Schimmel, P. R. in: *Biophysical Chemistry; Part III The Behavior of Biological Macromolecules*, W. H. Freeman and Company, New York, Chapter 18 (Configurational Statistics of Polymer Chains), 979-1018 (1980).
- da Silva**, R. A, Silva, M. A. A. da, Caliri, A., Deterministic folding: The role of entropic forces and steric specificities, *Journal of Chemical Physics* **114**, 4235 (2001).
- Doi**, M. in: *Introduction to polymer physics*, Oxford Science Publications, New York, Capítulo 1: Properties of an isolated polymer molecules 1-19 (1996).
- Fitzkee**, N. C. & Rose, G. D., Steric restrictions in protein folding: An α -helix cannot be followed by a contiguous β -strand, *Protein Science* **13**, 633 (2004).
- Grantcharova**, V., Alm, E. J., Baker, D., Horwich, A. L., Mechanisms of protein folding, *Current Opinion in Structural Biology* **11**, 70-82 (2001).
- Hansen**, J. P. & McDonald, I. R. in: *Theory of Simple Liquids*, Academic Press INC (London) LTD., New York, Chapter 3 (Computer Experiments on Liquids) (1976).
- Honig**, B., Protein Folding: From the Levinthal Paradox to Structure Prediction, *J. Mol. Biol.* **293**, 283 (1999).
- Lattman**, E. E. & Rose, G.D., Protein Folding – what's the question? *Proc. Natl. Acad. Sci. USA* **90**, 439 (1993).
- Levinthal**, C., Are there Pathways for Protein Foldign ?, *J. Chim. Phys* **65**, 44 (1968).
- Metropolis**, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., Equation of State Calculations by Fast Computing Machines, *J. Chem. Phys.* **21**, 1087 (1953).
- Pappu**, R. V., Srinivasan, R., Rose, G. D., The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding, *PNAS* **97**, 12565-12570 (2000).
- Pessia-Paste**, G. in: *Do caos à Inteligência Artificial*, Editora Unesp, 38 (1992).
- Rocha**, L.F.O., Tarragó, M.E.P., Caliri, A., The Water Factor in the Protein-Folding Problem, *Brazilian Journal of Physics* **34**, 90-101 (2004).
- Schulz**, G. E., Schirmer, R. H. in: *Principles of Protein Structure*, Springer-Verlag, Chapter 3 (Noncovalent Forces Determining Protein Structure) (1990).
- Taylor**, W. R., The Classification of Amino Acid Conservation, *J. Theor. Biol.* **119**, 205 (1986).
- Tarragó**, M. E. P., Rocha, L. F. O., daSilva, R. A, Caliri, A., Steric constraints as folding coadjuvant, *Physical Review E* **67**, 031901 (2003).
- Zamyatin**, A.A., Protein Volume in Solution, *Prog. Biophys. Mol. Biol.* **24**, 107 (1972).



CAPÍTULO 4

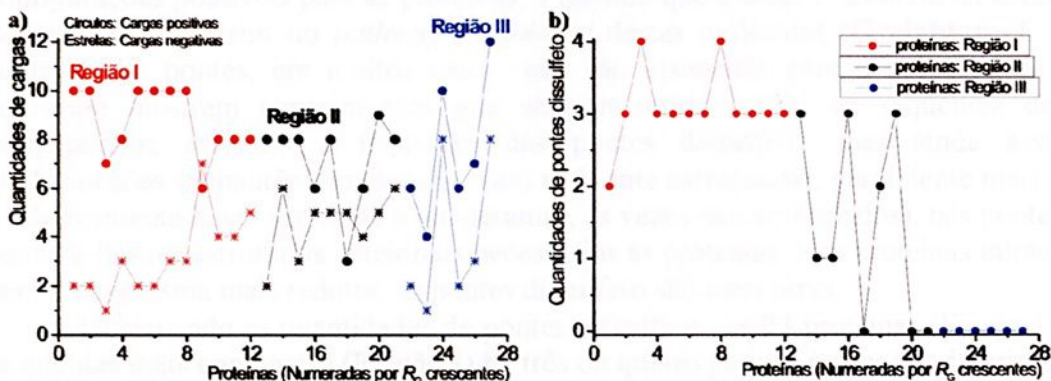
O Efeito Hidrofóbico e o *Folding* de Proteínas

Devido a várias questões ainda sem respostas para o processo de *folding* protéico (Pande, V.S. (2000)) é que há atualmente diversos modelos concorrentes, independentes ou complementares, para tal processo (seção 1.5.1 do Capítulo 1), além de vários outros modelos para classes específicas de proteínas.

Temos procurado, até aqui, neste trabalho, estudar aspectos do processo de *folding* sem distinção entre as possíveis topologias estruturais (globular e hélices). Porém neste Capítulo analisaremos preferencialmente as proteínas mais compactas (Região I, as doze primeiras, dentre as 27 proteínas), pois agora utilizaremos a interação hidrofóbica, no sistema proteína-solvente, como fator que leva as configurações protéicas do modelo a assumir uma forma compacta globular.

4.1 Forças que governam o processo de *folding* protéico

O processo que leva a cadeia polipeptídica à estrutura nativa de uma proteína e a se manter nela é dependente de um balanço delicado entre diversas forças (Schulz, G. E. & Schirmer, R. H. (1990)). Resultados teóricos e experimentais sugerem o efeito hidrofóbico (ou força entrópica) como a principal e mais geral força determinante da estrutura nativa (Dill, K. A. et al. (1995)), dentre todas as outras forças, como interações eletrostáticas, ligações de hidrogênio, van der Waals e pontes dissulfeto. Na Figura 1, mostramos como variam o número de aminoácidos carregados e pontes dissulfeto em função da compactação da proteína. As interações eletrostáticas e restrições configuracionais decorrentes possivelmente afetam a compactação do glóbulo protéico.



As interações eletrostáticas (**Honig, B. & Nicholls, A. (1995)**) entre grupos de cargas opostas (como entre os resíduos Lys e Asp), formando pares iônicos ou pontes salinas, estão presentes, em cerca de, 75% destes grupos nas proteínas (**Voet, D. et al., pg. 148-151 (2002)**). Contudo tais interações tendem a contribuir pouco para a estrutura e estabilidade das conformações nativas, exceto em soluções altamente ácidas ou básicas (**Dill, K.A. (1990)**), além disso, a energia livre das interações de carga do par iônico geralmente não é suficiente para compensar as perdas de entropia das cadeias laterais e da energia de solvatação quando se constitui um par iônico. Daí também decorre a pouca conservação destes pares iônicos em proteínas homólogas.

Observando as quantidades de cargas positivas e negativas nas 27 proteínas (Figura 1a), verifica-se que quase sempre há predomínio de resíduos com cargas positivas (Arg, Lys, His) sobre os resíduos com cargas negativas (Glu, Asp). Ainda pode-se observar variações destas cargas de acordo com as três regiões de compactações, sendo:

- Região I (em vermelho): Há um predomínio nítido das cargas positivas sobre as negativas (exceto na proteína nº 9), sendo tal predomínio menor para as proteínas menos compactas, a partir da proteína nº 10.

- Região II (em preto): Aqui ainda há, para algumas proteínas (nºs 13 e 15), predomínio de cargas positivas sobre as negativas; contudo nas demais proteínas tal predomínio fica reduzido e na proteína nº 18 as quantidades de cargas negativas superam as positivas.

- Região III (em azul): Ocorrem as maiores variações das quantidades de cargas positivas e negativas, sendo tais variações acopladas e dependentes da compactação, de uma para outra proteína. Não há grande predomínio das cargas positivas sobre as negativas, como nas proteínas da Região I.

Quando há predomínio, de um tipo de resíduo carregado sobre os de outro tipo, aumenta-se a restrição conformacional disponíveis para a cadeia protéica, como decorrência do volume excluído eletrostático (**Fixman, M. & Skolnick, J. (1978)**). Se tal predomínio for reduzido pode-se aumentar a influência eletrostática na proteína, com maior formação de pontes salinas.

A existência de pontes dissulfeto intra e inter cadeias, restringem muitas configurações possíveis para as proteínas, à medida que a estas se dobram na conformação nativa, atuando assim no *pathway* do *folding* destas moléculas (**Greighton, T. (1995)**); contudo tais pontes, em muitos casos, não são essenciais para o processo de *folding*, conforme mostram experimentos que alteram resíduos Cys na sequência de alguns polipeptídeos evitando a formação das pontes dissulfeto, mas ainda assim suas conformações se mantêm funcionais. Num ambiente extracelular, geralmente mais oxidante e relativamente hostil (com pH e temperatura, às vezes não controlados), tais pontes podem conferir limites estruturais adicionais necessários às proteínas. Nas proteínas intracelulares, com o citoplasma mais redutor, as pontes dissulfeto são mais raras.

Observando as quantidades de pontes dissulfeto nas 27 proteínas (Figura 1b), nota-se que nas mais compactas (Região I) há três ou quatro pontes, à exceção da primeira, com apenas duas pontes. À medida que a compactação diminui (Região II), há maiores variações, nas quantidades das pontes, desde zero a três. A partir da vigésima proteína, ainda na Região II, cujas conformações nativas são distendidas, não há pontes dissulfeto. Tais resultados mostram que proteínas compactas provavelmente fazem uso das pontes dissulfeto para ajudar na estabilização de suas conformações nativas.

As ligações de hidrogênio, mesmo exercendo um papel central nas estruturas das proteínas (seção 1.6, Capítulo 1), tendem a ter uma contribuição menor para a estabilidade das proteínas (**Kauzmann, W. (1959)**), quando comparadas ao número equivalente destas ligações para as proteínas nas formas desenovelada (com a água ao seu redor) e nativa (intra cadeia). Contudo estas podem constituir-se em importantes fatores em momentos estratégicos para a determinação e estabilização da estrutura nativa da proteína dentre um pequeno número de conformações hidrofobicamente estabilizadas. As quantidades de ligações de hidrogênio para as proteínas, consideradas neste trabalho, variam de 9 a 32 (Apêndice, após o Capítulo 5).

As forças de van der Waals são praticamente não seletivas e atuam somente a curta distância, e portanto são mais importantes para a estabilização da proteína, somente em conformações compactas, como no caso da conformação nativa.

Em solução aquosa a conformação de uma proteína é fortemente influenciada pela sequência de seus resíduos polares e apolares, pois contribuem diferentemente na variação da energia livre (**Rocha, L. F. O. et al. (2004)**). Os resíduos apolares da proteína tendem a se agregar minimizando sua exposição às moléculas de água. Este comportamento é conhecido como efeito hidrofóbico, sendo o responsável por uma das forças dominantes no processo de *folding* (**Southall, N. T. (2002)**).

4.2 Efeito hidrofóbico

O entendimento do efeito hidrofóbico a nível molecular está ainda em pleno andamento, por meio de diversos enfoques; dentre eles destacamos alguns estudos recentes (**Lum, K. et al (1999)**); **Chandler, D. (2002)** propondo que a segregação das substâncias apolares do *bulk* de água decorre ultimamente da versatilidade das moléculas de água rearranjar as interações energéticas, minimizando a redução da “quebra” de suas ligações de hidrogênio (ou ligações H). Para grandes extensões de interfaces água-substâncias apolares, é inevitável a quebra das ligações de hidrogênio e a fim de minimizar a energia potencial do sistema, estes se redistribuem no *bulk* da água, num processo análogo ao da interface ar – água, manifestado através da alta tensão superficial da água. Para pequenas moléculas apolares o número médio de ligações H praticamente fica inalterado, pois as ligações se contorcem e acabam por envolver as pequenas moléculas apolares. Conseqüentemente a energia de solvatação das moléculas apolares varia de acordo com o tamanho destas moléculas, sendo:

- Com pequenas moléculas, a quantidade de água afetada pela sua presença, e portanto a energia livre de solvatação, se torna proporcional ao volume das moléculas apolares (supondo moléculas esféricas de tamanho R , teremos o volume $(4/3)\pi R^3$), sendo a constante de proporcionalidade em torno de $1 \times 10^8 \text{ J/m}^3$ (ou $1 \times 10^{-22} \text{ J/\AA}^3$), para o etano, propano e outras moléculas similares.
- Com a presença de grandes moléculas apolares, a quebra das ligações H é inevitável e então a energia livre de solvatação fica proporcional a área superficial ($4\pi R^2$) da interface molécula apolar-água, sendo a constante de proporcionalidade estimada da tensão superficial da interface água – óleo, em torno de $6 \times 10^{-2} \text{ J/m}^2$ (ou $6 \times 10^{-22} \text{ J/\AA}^2$).



No cruzamento de domínio entre grandes e pequenas extensões de hidratação de moléculas apolares, ocorre a igualdade¹:

$$((4/3)\pi R^3) \times 1 \times 10^{-22} = (4\pi R^2) \times 6 \times 10^{-22}, \quad (1)$$

e assim obtêm-se que R ao redor de 1 nm é a dimensão que discrimina os diferentes comportamentos de hidratação de substâncias apolares.

Para grandes moléculas apolares (≥ 1 nm) o custo energético de agrupá-las é significativamente mais favorável que mantê-las separadas. Em pequenas moléculas (extensões $\lesssim 1$ nm), o número de ligações de hidrogênio, praticamente ficam inalterados, pois estas envolvem a molécula. Assim a energia hidrofóbica é insuficiente para superar a energia térmica, isso inclusive explica a solubilidade das pequenas moléculas apolares. No caso da mistura água – óleo o limiar de 1nm coincide com o raio crítico para a separação do óleo da água. Tal sistema é a base do que se entende pelo efeito hidrofóbico e embora para proteína não seja fácil e direto sua aplicação, são as ligações de hidrogênio entre as moléculas do solvente (água) a origem deste efeito em ambos os casos.

4.3 Termodinâmica do processo de *folding* de proteínas

As moléculas de água se ligam até a quatro outras (numa estrutura tetraédrica, aberta), formando uma extensa rede de ligações de hidrogênio no *bulk* da água (**Head-Gordon**, T. (1995); **Udachin**, K. & **Ripmeester**, J.A. (1999)); quando há grandes moléculas apolares no ambiente aquoso, observa-se uma certa “aversão” entre estas e a água (segregação das primeiras devido às interações preferenciais entre as próprias moléculas de água), e não como se imagina comumente, uma repulsão entre moléculas apolares e água. No caso de proteínas globulares imersas na água ocorre uma ação recíproca entre ambas, inclusive tornando tais proteínas solúveis em água, mas muitos modelos usados no estudo do *folding* de proteínas incluem o solvente implicitamente, integrando seus graus de liberdade (**Karplus**, M. & **Shakhnovich**, E. (1992)); tal tratamento pode ser melhor evidenciado utilizando a função partição (Z) do sistema proteína-solvente, sendo

$$Z = \sum_{\{R_i\}} \sum_{\{r_j\}} \exp[-E(R_1, \dots, R_N; r_1, \dots, r_m)/k_B T], \quad (2)$$

onde E é a função energia potencial do sistema, k_B é a constante de Boltzmann, T é a temperatura absoluta e as coordenadas da cadeia e solvente são, respectivamente, $\{R_i\}$ e $\{r_j\}$. A função Z pode ser simplificada como:

$$Z = \sum_{\{R_i\}} \exp[-\Phi(\{R_i\})/k_B T], \quad (3)$$

com

¹ O desenvolvimento matemático detalhado, feitos por nós, se torna necessário devido a importância da determinação do valor limite para o tamanho (R) das moléculas apolares e também por este valor não estar determinado explicitamente nos artigos de **Lum**, K. et al (1999) e **Chandler**, D. (2002).

$$\Phi(\{R_i\}) = -(k_B T) \ln \sum_{\{r_j\}} \exp[-E(R_1, \dots, R_N; r_1, \dots, r_m) / k_B T] \quad (4)$$

A função $\Phi(\{R_i\})$ é a energia livre do sistema quando a cadeia é considerada congelada numa dada configuração do conjunto $\{R_i\}$; neste ponto, utilizando a mecânica estatística, existem dois caminhos a serem seguidos para a abordagem do processo de *folding*.

(i) A energia livre Φ , é considerada como um potencial de força média sobre os graus de liberdade do solvente tratados implicitamente; utilizando-se desta abordagem muitos modelos simplificados do *folding* de proteínas reduzem-se à utilizar a interação intra cadeia (normalmente atrativa para contato entre resíduos de tipo iguais e repulsivo para tipo diferente) por meio do potencial de força média (Li, H. et al. (1996); Socci, N. D. et al. (1997); Garcia, L. G. et al. (2001)).

(ii) Uma outra alternativa é considerar o solvente explicitamente na dinâmica do processo de *folding*, enfatizando as interações cadeia-solvente (Levitt, M. & Sharon, R. (1988)). Antes de detalharmos implementações nesta hipótese alternativa, analisaremos algumas quantidades termodinâmicas no processo de *folding*, a fim de associar o efeito hidrofóbico com o estado de equilíbrio cadeia-solvente.

Um dos mais convincentes argumentos a favor do efeito hidrofóbico, como força dominante do processo de *folding* vêm da termodinâmica, considerando a reação de transição entre dois estados (Baldwin, R. L. & Rose, G. D. (1999)), verificada para muitas proteínas, por diferentes técnicas experimentais:

unfolding (desenovelado) $\xleftarrow{\quad}$ *folding* (enovelada ou nativa),

sendo a mudança na energia livre (ΔG) desta reação, nas duas conformações, separada nas diferentes contribuições do solvente e da cadeia. Assim

$$\Delta G = G_{\text{enovelada}} - G_{\text{desenovelada}} = \Delta H_{\text{cadeia}} + \Delta H_{\text{solvente}} - T\Delta S_{\text{cadeia}} - T\Delta S_{\text{solvente}}, \quad (5)$$

com ΔH representando as mudanças entálpicas e ΔS as mudanças entrópicas; similarmente $\Delta H = H_{\text{enovelada}} - H_{\text{desenovelada}}$, $\Delta S = S_{\text{enovelada}} - S_{\text{desenovelada}}$

Para analisarmos a contribuição de cada um dos termos mais detalhadamente, separamos as contribuições dos grupos polares e hidrofóbicos (apolares) da cadeia.

4.3.1 ΔG^P : Contribuições dos grupos polares (P) para ΔG

- O *backbone* da proteína forma muito mais ligações de hidrogênio e interações salinas quando exposto a água do que com ele próprio, assim $\Delta H_{\text{cadeia}}^P > 0$ (favorecendo o estado desenovelado); já no solvente, embora exista interação com os grupos polares, as moléculas de água preferem interagir mais entre elas, do que com o cadeia, daí $\Delta H_{\text{solvente}}^P < 0$ (favorecendo o estado nativo). Embora exista uma competição entre os dois termos, nas condições fisiológicas, a interação entre as moléculas de água prevalece, fornecendo:

$$\Delta H_{\text{cadeia}}^P + \Delta H_{\text{solvente}}^P \lesssim 0 \quad (6)$$

- A cadeia acessa muito mais configurações na conformação desenovelada do que na enovelada, assim a mudança entrópica da cadeia é negativa, gerando $-T\Delta S_{cadeia}^P > 0$ (contribuindo para o estado desenovelado); todavia no solvente as moléculas de água possuem muito mais opções configuracionais com a cadeia enovelada do que desenovelada, daí $-T\Delta S_{solvente}^P < 0$ (favorecendo o enovelamento). Na concorrência entre os termos entrópicos da cadeia e do solvente predomina, ligeiramente o positivo, deste modo

$$-T\Delta S_{cadeia}^P - T\Delta S_{solvente}^P \gtrsim 0 \quad (7)$$

Como saldos as contribuições dos grupos polares para a variação da energia livre favorecem ligeiramente a conformação desenovelada, isto é:

$$\Delta G^P = (\Delta H_{cadeia}^P + \Delta H_{solvente}^P) - T\Delta S_{cadeia}^P - T\Delta S_{solvente}^P \gtrsim 0 \quad (8)$$

4.3.2 $\Delta \tilde{G}^P$: Contribuições dos grupos apolares ($\tilde{P}$) para ΔG

- Na cadeia os grupos apolares interagem um pouco mais fortemente com a água do que com eles próprios; contudo quando a cadeia está desenovelada, prevalece nestes grupos a interação do *backbone* com o solvente tendo formação de mais ligações de hidrogênio com água do que com eles próprio, levando a preferência para o estado desnaturado com

$\Delta H_{cadeia}^{\tilde{P}} \gtrsim 0$ – apesar da preferência pelo estado desnaturado ser inferior a ΔH_{cadeia}^P – devido a interação atrativa de van der Waals ser mais forte entre estes grupos do que entre eles e a água. No solvente as moléculas de água preferem interagir mais entre si, do que com os grupos apolares, levando a preferência pelo estado enovelado, com $\Delta H_{solvente}^{\tilde{P}} < 0$. Na competição entre os dois termos entálpicos há tendência para o estado nativo, isto é,

$$\Delta H_{cadeia}^{\tilde{P}} + \Delta H_{solvente}^{\tilde{P}} \lesssim 0 \quad (9)$$

- Como no caso polar, na conformação desenovelada a entropia da cadeia é maior, assim os grupos apolares têm $-T\Delta S_{cadeia}^{\tilde{P}} > 0$. Já no solvente existe uma forte preferência pelo estado

enovelado, como no sistema água (solvente) – óleo (apolar), e $-T\Delta S_{solvente}^{\tilde{P}} \ll 0$; isso devido o rearranjo das moléculas de água, que preferem interagir entre si reduzindo ao mínimo a perda de suas ligações de hidrogênio e levando a diminuição da interface cadeia – solvente (de fato esta é considerada a principal força no *folding* de proteínas), daí

$$-T\Delta S_{cadeia}^{\tilde{P}} - T\Delta S_{solvente}^{\tilde{P}} \ll 0 \quad (10)$$

Assim, a contribuição dos termos entrópicos e entálpicos levam a uma variação de energia livre favorecendo, em muito o estado nativo, isto é

$$\Delta \tilde{G}^P = (\Delta H_{cadeia}^{\tilde{P}} + \Delta H_{solvente}^{\tilde{P}}) - T\Delta S_{cadeia}^{\tilde{P}} - T\Delta S_{solvente}^{\tilde{P}} < 0 \quad (11)$$

Em resumo, temos

Para a cadeia: $\Delta H_{cadeia}^P > 0$; $-T\Delta S_{cadeia}^P > 0$; $\Delta H_{cadeia}^{\tilde{P}} \gtrsim 0$; $-T\Delta S_{cadeia}^{\tilde{P}} > 0$ e

Para o solvente: $\Delta H_{solvente}^P < 0$; $-T\Delta S_{solvente}^P < 0$; $\Delta H_{solvente}^{\tilde{P}} < 0$; $-T\Delta S_{solvente}^{\tilde{P}} \ll 0$
 Portanto, graça a contribuição predominante do solvente para a conformação enovelada, o processo de *folding* é espontâneo, com

$$\Delta G = \Delta G^P + \Delta G^{\tilde{P}} \lesssim 0 \tag{12}$$

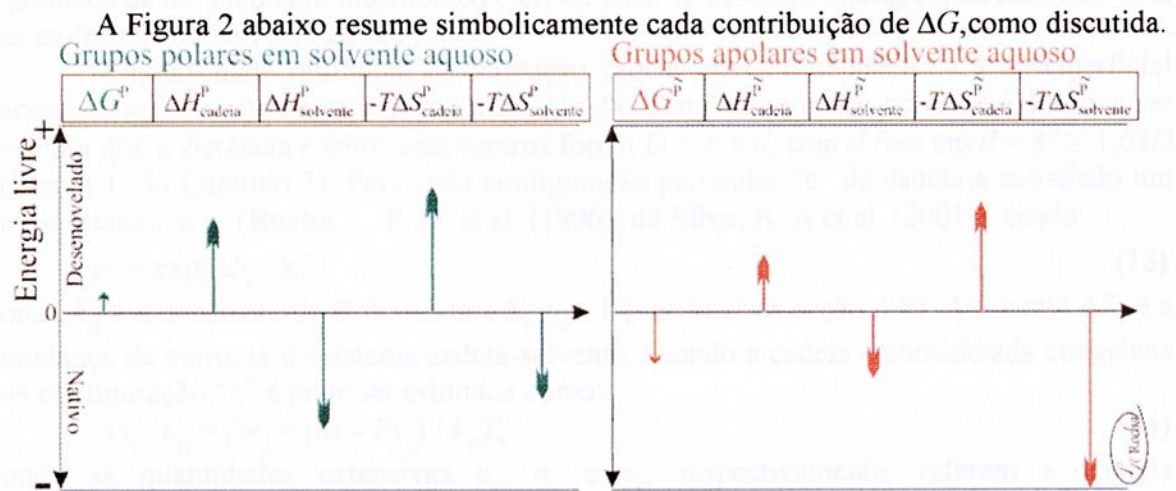


Figura 2 - Diagrama mostrando a contribuição dos termos entálpicos e entrópicos da cadeia e do solvente para a variação da energia livre (ΔG) dos resíduos polares (em verde) e apolares (em vermelho) de uma proteína num meio aquoso.
 Note a grande contribuição do solvente (última seta em vermelho) para a conformação enovelada.

Embora espontâneo – no sentido de não precisar do auxílio de nenhuma maquinaria biológica para sua realização (pelo menos para muitas proteínas pequenas) – o processo de *folding* protéico é marginalmente estável, pois para proteínas típicas ΔG é muito pequeno ($\lesssim 0$), entre -20 e -63 kJ/mol, correspondendo a poucas ligações de hidrogênio (cada uma tendo entre -21 e -8 kJ/mol).

Portanto após evidenciar detalhes termodinâmicos do solvente envolvidos no problema do enovelamento protéico, é fortemente sugestivo que o solvente seja considerado como protagonista do processo de enovelamento. Assim, passamos a utilizar esta idéia na abordagem via simulação Monte Carlo do modelo, particularmente no modelo simplificado no espaço contínuo (fora da rede).

4.4 Modelo simplificado: Potencial hidrofóbico e restrições estéricas

Na simulação Monte Carlo (MC) do sistema proteína – solvente (Figura 2 do Capítulo 3), a cadeia protéica é representada topologicamente, por um “colar de pérolas” e o solvente por “esferas”. Na cadeia com 35 monômeros, cada um aminoácido é representado por uma esfera rígida de diâmetro D conectados aos seus vizinhos por cordas idealmente flexíveis de comprimento ℓ , sendo $D \leq \ell \leq D + \varepsilon$ (com $\varepsilon = 0,4$, conforme

discutido na seção 3.1.2). As moléculas do solvente, num total de 12.565, também são representadas por esferas rígidas de diâmetro D .

Neste sistema proteína-solvente para quaisquer duas de suas partículas i e j separadas por uma distância $r_{i,j}$ é considerado o efeito de volume excluído através de interação do tipo esfera rígida, descrito pela energia potencial $U(r_{i,j}) \rightarrow \infty$ para $r_{i,j} < D$ e $U(r_{i,j}) = 0$ caso contrário. Adicionalmente os monômeros da cadeia possuem atributos químicos de um elemento hidrofóbico (2H) ou polar (P) e assim interagem diretamente com as moléculas do solvente ao seu redor.

A intensidade resultante da interação cadeia-solvente depende da área superficial acessível ao solvente de cada monômero. Um "contato" monômero-solvente é dito ocorrer sempre que a distância r entre seus centros forem $D < r \leq d$, com d fixo em $d = 8^{1/2} \cong 1,63D$ (Figura 1, do Capítulo 3). Para cada configuração particular "c" da cadeia é associado um peso estatístico p_c (Rocha, L. F. O. et al. (1996); da Silva, R. A et al. (2001)), sendo:

$$p_c = \exp(\Delta S_c / k_B), \quad (13)$$

onde k_B é a constante de Boltzmann e $\sum_c p_c = 1$ (ver final da seção 4.8). A quantia ΔS_c é a mudança de entropia do sistema cadeia-solvente, quando a cadeia é considerada congelada na configuração "c" e pode ser estimada como:

$$\Delta S_c / k_B = (-e_c + \mu n_c - P v_c) / k_B T, \quad (14)$$

onde as quantidades extensivas e_c , n_c e v_c , respectivamente, referem a energia configuracional da cadeia, ao número moléculas do solvente quimicamente ligada à cadeia e ao volume da cadeia; já as quantidades intensivas μ , P , T são, respectivamente, o custo entrópico de ligar o solvente à cadeia, a pressão e a temperatura do sistema.

No espaço de fase configuracional, ao ir de uma configuração congelada A para a configuração congelada B, tem-se uma probabilidade de transição $T(A \rightarrow B)$ e utilizando o peso estatístico p_c , como a probabilidade de equilíbrio $P(A)$ na relação de balanço detalhado, $P(A) T(A \rightarrow B) = P(B) T(B \rightarrow A)$, obtêm-se:

$$T(A \rightarrow B) = \min\{\exp(\delta S_{A,B} / k_B), 1\} \quad (15)$$

, onde $\delta S_{A,B} = \Delta S_B - \Delta S_A = (-\delta e_{A,B} + \mu \delta n_{A,B} - P \delta v_{A,B}) / T$ é a mudança de entropia do sistema cadeia-solvente com respeito a transição entre as configurações congeladas A e B da cadeia; a função $\min\{i, j\}$ significando o menor dos valores entre i e j .

$$T(A \rightarrow B) = \min\{\exp((- \delta e_{A,B} + \mu \delta n_{A,B} - P \delta v_{A,B}) / k_B T), 1\} \quad (16)$$

Assim, na aproximação de n e v constantes, temos que $\delta n_{A,B} = \delta v_{A,B} = 0$ e a variação $\delta S_{A,B}$ para o modelo considerado depende somente da mudança da energia hidrofóbica ($\delta S_{A,B} = -\delta e_{A,B} / k_B T$). Assim, estabelecendo uma relação linear entre a energia hidrofóbica e a área acessível ao solvente (Chothia, C. (1974)) podemos estimar $e_{A,B}$ associando-se um específico nível hidrofóbico (Figura 3) $\gamma = 2\gamma_0$ para monômeros hidrofóbicos e $\gamma = -\gamma_0$ (a quantia $\gamma_0 < 0$ é medida em unidades de $k_B T$) para monômeros polares.

Portanto no movimento $A \rightarrow B$, há três casos a serem considerados, para $T(A \rightarrow B)$, dependendo da partícula do sistema, sendo

$$T(A \rightarrow B) = \min\{\exp(2\gamma_0 \Delta n_S / k_B T), 1\} \text{ movendo um monômero H;} \quad (17)$$

$$T(A \rightarrow B) = \min\{\exp(-\gamma_0 \Delta n_S / k_B T), 1\} \text{ movendo um monômero P;} \quad (18)$$

² Passaremos a representar os monômeros hidrofóbicos pelo símbolo H, ao invés de $\tilde{P}$, da seção 4.3.

$T(A \rightarrow B) = \min\{\exp(\gamma_0 (2\Delta n_H - \Delta n_P)/k_B T), 1\}$ movendo uma molécula do solvente, (19) onde Δn_S é a mudança do número de moléculas de solvente em contato com o monômero movido; e Δn_H e Δn_P são, respectivamente, a mudança no número de monômeros polar e hidrofóbico ao redor da molécula de solvente movida; o valor de γ_0 foi fixado em $\gamma_0 = -1$.

Para cada nova configuração, gerada ao longo da simulação Monte Carlo, as condições $|\vec{r}_i - \vec{r}_j| \geq D$ (volume excluído) deve ser satisfeita para cada par (i, j) de esferas, e se “i” e “j” pertencerem à cadeia a condição adicional $|\vec{r}_i - \vec{r}_j| \leq D + \epsilon$ (preservação da cadeia, Figura 2 do Capítulo 3) também precisa ser observada. Estas condições sendo satisfeitas, então se utiliza o potencial hidrofóbico, passando a avaliar, de acordo com o tipo de partícula movida, a probabilidade de transição $T(A \rightarrow B)$ da configuração A para nova configuração B. De acordo com a probabilidade de transição (Equações 17-19) a nova configuração B pode ser aceita ou não, como já detalhado na seção 3.1. Desta forma são realizados 900.000 PMC (ou 900 kPMC, o que corresponde a cerca de $1,13 \times 10^{10}$

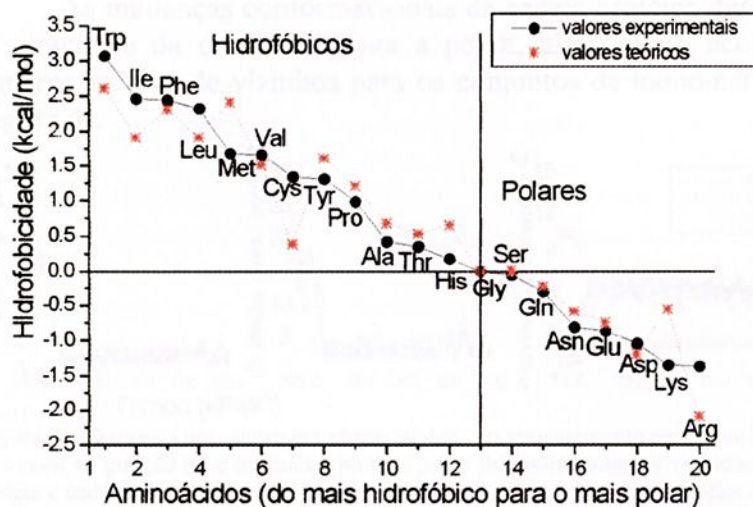


Figura 3 - Valores das hidrofobicidades (ΔG) dos 20 Aminoácidos padrões, segundo Eisenberg, D & McLachlan, A. D., pg. 200 (1986) - Para os monômeros Hidrofóbicos, $\Delta G > 0$, indicando que a transferência do aminoácido de um meio apolar para um polar é desfavorecida, ao contrário dos polares, daí para estes $\Delta G < 0$. Utilizamos a escala acima para determinar, de forma simplificada, os valores dos níveis hidrofóbicos, para os monômeros Hidrofóbicos ($\gamma = -2$, total de 12 Aminoácidos) e Polares ($\gamma = +1$, total de 8 Aminoácidos). Obs. $\Delta G^{\text{experimental}}$ dos aminoácidos são obtidos em relação a Glicina, da transferência do octanol para a água; $\Delta G^{\text{teórica}}$ é estimada pelo produto da acessibilidade de cada átomo para o solvente e seu parâmetro de solvatação atômica. configurações) em nove simulações independentes, conforme descrito no Capítulo 3.

No presente modelo, adota-se um alfabeto de duas letras, H e P com níveis hidrofóbicos $\gamma = +1$ e $\gamma = -2$, sendo cada monômero, na sequência HP do modelo, possuindo o atributo hidrofóbico de cada um dos 35 aminoácidos de uma proteína real, cuja estrutura é adotada como configuração alvo. Neste caso todas as restrições geométricas são praticamente eliminadas.

Em relação às escalas hidrofóbicas, como as da Figura 3, há diferenças significativas nas diversas escalas, quanto ao valor do nível hidrofóbico e até divergências quanto ao caráter hidrofóbico ou polar de alguns aminoácidos (Wang, Zi-Hao & Lee, H.C. (2000)). Tais diferenças são decorrentes dos métodos usados para construir tais escalas.

Algumas escalas (Janin, J. (1979); Rose, G. et al. (1985)) são construídas examinando estruturas nativas das proteínas e definindo o caráter hidrofóbico, como a tendência de resíduos serem encontrados no interior, em oposição à superfície protéica. Outras escalas (Wolfenden, R. (1981); Kyte, J. & Doolite, R. (1982)) são derivadas de propriedades físico-químicas das cadeias laterais dos aminoácidos. Fizemos a escolha pela escala de Eisenberg, D & McLachlan, A. D. (1986), pois os valores experimentais e teóricos de ΔG são consistentes (Figura 3), além de ser bastante adotada em diversos trabalhos recentes.

4.5 Propriedades configuracionais do modelo simplificado

Utilizando o modelo descrito acima a simulação é iniciada com o sistema (cadeia + solvente) em uma configuração típica do regime de alta temperatura ($k_B T \rightarrow \infty$), termalizada durante 100.000 PMC (ou 100 kPMC). Em seguida a temperatura é reduzida ($k_B T = 1,5$, conforme discutido na seção 3.1.1) e mantida constante nos próximos 800 kPMC da simulação.

As mudanças conformacionais da cadeia protéica durante a simulação são dos raios de giração e da distância ponta a ponta; além disso, acompanhamos as variações dos números médios de vizinhos para os conjuntos de monômeros hidrofóbicos H e polares P (Figura 4).

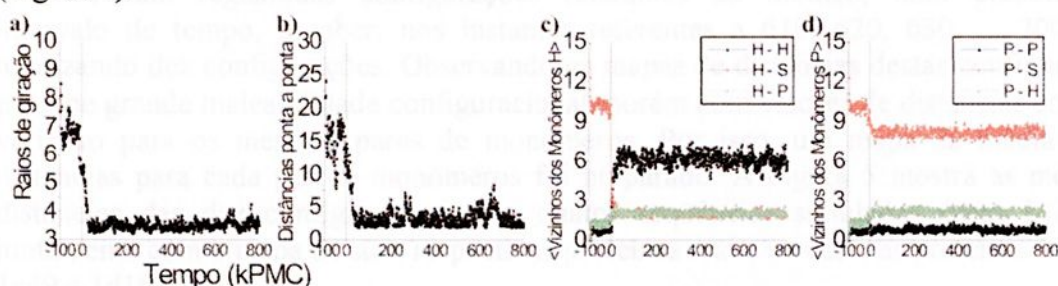


Figura 4 - Evolução das quatro grandezas globais do enovelamento para primeira simulação - As variações dos raios de giração (a) e distâncias ponta a ponta (b), assim como as variações médias dos vizinhos intra cadeia e cadeia-solvente (S) dos monômeros H (c) e P (d) mostram que, após a fase de descarte configuracional (com tempo positivo e temperatura $k_B T = 1,5$) as configurações do modelo (geradas na simulação) são, em média, compactas no transcorrer do tempo MC. Cada ponto nos gráficos acima, são correspondente a uma média de 1000 configurações.

Quando a cadeia colapsa em uma conformação tipo glóbulo (o que tipicamente ocorre após 50 kPMC), os raio de giração (Figura 4a) e a distância ponta a ponta (Figura 4b) atingem valores mínimos, pois os contatos entre os monômeros hidrofóbicos são otimizados com os monômeros hidrofóbicos ficando no interior do glóbulo (contatos H-H, Figura 4c), e protegidos pelos monômeros polares, preferencialmente distribuídos pela superfície do glóbulo em permanente contato com o solvente (contatos P-S, Figura 4d). Tais resultados confirmam (Figura 3, do Capítulo 3) que em $k_B T = 1,5$ a cadeia, independente das condições iniciais, sempre colapsa numa conformação tipo glóbulo compacto.

Utilizando os atributos químicos HP da sequência das quatro seguintes proteínas compactas: Iroo (nº 2), ltsk (nº 4), lc49 (nº 7) e ld1h (nº 12), como especificadores das sequências das cadeias do modelo, podemos verificar o quanto a disposição tridimensional dos aminoácidos nas configurações nativas destas proteínas conseguem ser preditas pelos

resultados obtidos por simulação, seção 4.6 abaixo. A escolha por estas proteínas se baseou em se utilizar proteínas de diferentes compactações e descorrelacionadas entre as mais compactas, da Região I (Figura 1).

4.6 Comparação entre as estruturas das proteínas alvos e conformações da cadeia preditas pelo modelo

Para fazer comparações de distâncias entre monômeros das proteínas reais e os do modelo simulado, as distâncias entre os monômeros, para cada um dos casos, separadamente, são apropriadamente re-escaladas e normalizadas, conforme equação 1, do Capítulo 2. As distâncias são medidas entre os centros de massa de cada resíduo de aminoácido das estruturas protéicas, e no caso da cadeia do modelo, entre os centros dos monômeros esféricos.

4.6.1 Mapas de Distância de configurações protéicas e do modelo simulado

Foram registradas configurações referentes ao modelo, num predeterminado intervalo de tempo, a saber, nos instantes referentes a 610, 620, 630, ... 700 kPMC, totalizando dez configurações. Observando os mapas de distâncias destas configurações se percebe grande maleabilidade configuracional, porém com valores de distâncias com pouca variação para os mesmos pares de monômeros. Por isso, um mapa da média das dez distâncias para cada par de monômeros foi preparado. A Figura 5 mostra as médias das distâncias das dez configurações para quatro seqüências simuladas (para $k_B T = 1,5$) juntamente com o mapa de suas respectivas proteínas reais, no caso, as proteínas 1roo, 1tsk, 1c49 e 1dlh.



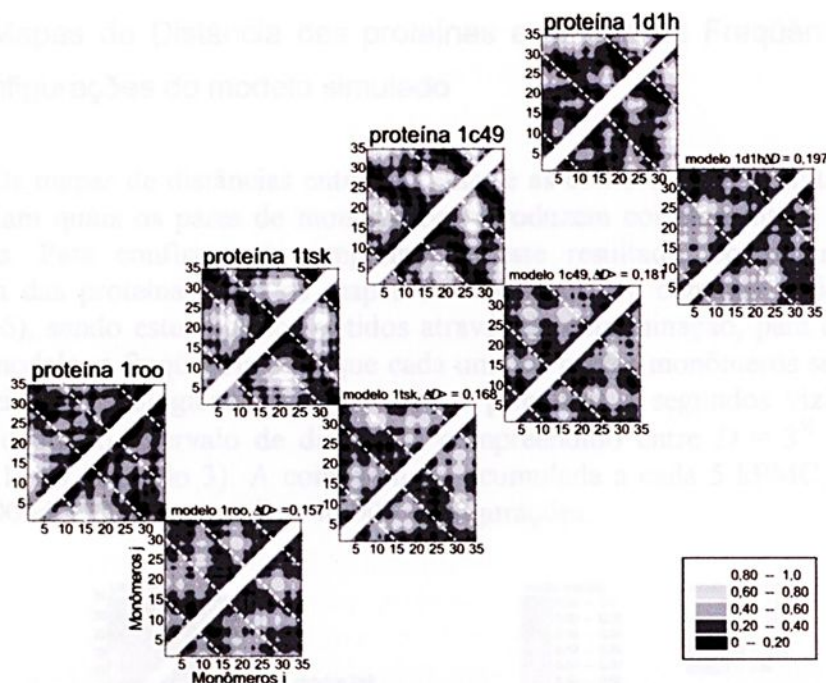


Figura 5- Mapas de distâncias normalizadas para proteínas e modelo, 1roo, 1tsk, 1c49 e 1d1h - Comparando os mapas das quatro proteínas alvos e da média das dez configurações do modelo é possível verificar visualmente suas semelhanças (mesmas cores ou cores parecidas) e dessemelhanças (cores muito diferentes). As retas paralelas pontilhadas brancas, nos gráficos, são utilizadas para guiar e facilitar comparações entre os mapas. Acima de cada mapa de distância do modelo, consta a discrepância, ΔD média das dez configurações do modelo, com sua respectiva proteína. Os primeiros e segundos vizinhos ao longo da cadeia são excluídos dos mapas de distâncias (regiões brancas na diagonal principal, como nos mapas de distâncias nos Capítulos 2 e 3).

Na figura 5 os mapas de distâncias tanto os referentes às proteínas reais, quanto aos resultados das simulações estão na mesma escala. Grandes distâncias normalizadas são representadas por cor clara, e pequenas distâncias por cor escura. Assim temos: (i) - regiões escuras próximas da diagonal principal, indicando contatos entre monômeros próximos na seqüência, como em estruturas tipo α -hélices; (ii)- regiões escuras na diagonal secundária, indicam contatos entre monômeros que lembram estruturas tipo folhas β .

Observando os gráficos da figura 5 destacamos alguns pontos:

- Os mapas da média das distâncias (organizadas em intervalos de distância de 0,2 com diferentes tons de cinza) eliminam coincidências eventuais, pois correspondem à contribuições de configurações descorrelacionadas.
- Com a ajuda de retas pontilhadas paralelas à diagonal secundária, é possível observar detalhes do quanto certas distâncias de pares de monômeros das estruturas das proteínas são preservadas e reproduzidas nos mapas dos modelos simulados.
- Para as proteínas simuladas, no intervalo de tempo considerado, coincidentemente observamos que as discrepâncias configuracional entre as proteínas reais e suas respectivas configurações do modelo simulado aumentam como a compactação, indo de 0,157, para a proteína mais compacta 1roo, à 0,197 para a proteína mais distendida 1d1h.

4.6.2 Mapas de Distância das proteínas e Mapas de Frequências de Contatos das configurações do modelo simulado

Os mapas de distâncias entre proteínas e as conformações resultantes da simulação evidenciam quais os pares de monômeros reproduzem corretamente o contato nativo das proteínas. Para confirmar e complementar este resultado, comparamos os mapas de distância das proteínas com os mapas de frequência de contatos obtidos por simulação (Figura 6), sendo estes últimos obtidos através da determinação, para cada monômero da cadeia modelo, a frequência com que cada um dos outros monômeros se configuram como seus vizinhos topológicos (excluindo-se os primeiros e segundos vizinhos ao longo da cadeia) dentre o intervalo de distâncias compreendido entre $D = 3\frac{1}{2}$ e $d = 8\frac{1}{2} \cong 2,828$ (Figura 1, do Capítulo 3). A contagem foi acumulada a cada 5 kPMC, no intervalo entre 501 e 700 kPMC, o que totaliza 40.000 configurações.

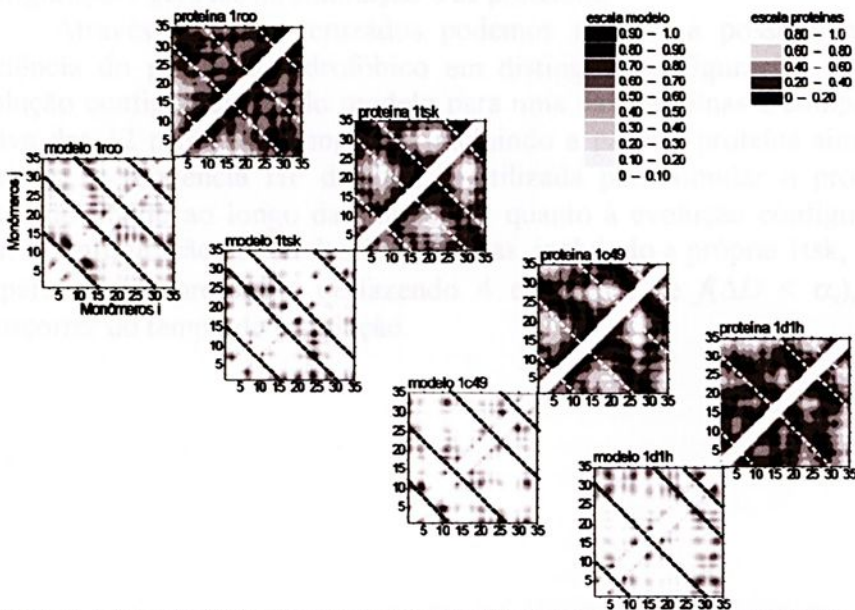


Figura 6- Mapas de distâncias para proteínas (gráficos superiores, os mesmos da Figura 5) e mapas de contato para modelos (gráficos inferiores) - Devidos tais mapas serem conceitualmente diferentes, são usadas diferentes escalas de cores, especificamente cinco cores para os mapas de distâncias e dez cores para os mapas de contato. Obs. Nos mapas de contato, quanto mais escuras as cores, maior a frequência de contato entre os pares de monômeros “i” e “j”. Para as proteínas e modelo são utilizadas, respectivamente, retas paralelas pontilhadas brancas e pretas.

Embora o mapa de frequência de contato e o mapa de distância forneçam medidas essencialmente distintas, pelo fato de podermos associar “altas frequências de contato” entre pares de monômeros à “curtas distâncias efetivas” entre eles, podemos esperar alguma semelhança entre tais mapas. De fato, percebemos para todos os casos uma estreita similaridade entre estes dois mapas. Portanto os mapas de frequência de contato reforçam os resultados e os comentários já feitos para os mapas de distâncias (Figura 5).

Agora passaremos a executar comparações de forma mais geral, empregando o potencial hidrofóbico, determinado pelos atributos químicos HP, das seqüências de quatro proteínas alvo. De fato, a comparação incluirá as discrepâncias cruzadas, como a descrita a seguir.



4.6.3 Discrepâncias cruzadas: comparações entre estruturas protéicas e configurações do modelo simulado

Ao longo de cada simulação de 800 kPMC, dentro de cada intervalo de tempo de 1 kPMC, calcula-se o valor da discrepância cruzada ΔD , entre as configurações do modelo e cada proteína, e verifica se a condição $\Delta D < \alpha$ é satisfeita, acumulando todo vez que tal condição ocorrer na frequência de discrepância $f(\Delta D < \alpha)$, para cada proteína, sendo f um número inteiro que satisfaz $0 \leq f(\Delta D < \alpha) \leq 1000$.

Sabemos que quanto menor ΔD , mais semelhantes serão duas configurações, assim foi estabelecido individualmente, para cada uma das quatro proteína simulada, o parâmetro limiar $\alpha = \langle \Delta D_{\min G} \rangle$ (a discrepância mínima global média dentre as nove simulações, Tabela 1, do Capítulo 3), que parece ser um valor adequado para comparação entre as configurações geradas na simulação e as proteínas.

Através de testes cruzados podemos ampliar a possibilidade de se verificar a eficiência do potencial hidrofóbico em distinguir configurações. Assim, o resultado da evolução configuracional do modelo para uma das proteínas é comparado com a estrutura nativa das 12 proteínas compactas, incluindo a própria proteína simulada (Figura 7). Por exemplo, a sequência HP da Itsk, é utilizada para simular o processo de *folding* e o resultado obtido ao longo da simulação, quanto à evolução configuracional é comparada com a configuração nativa das 12 proteínas, incluindo a própria Itsk, como mostra a Figura 7, para quatro proteínas, perfazendo 4 conjuntos de $f(\Delta D < \alpha_i)$, para cada uma, no transcorrer do tempo de simulação.

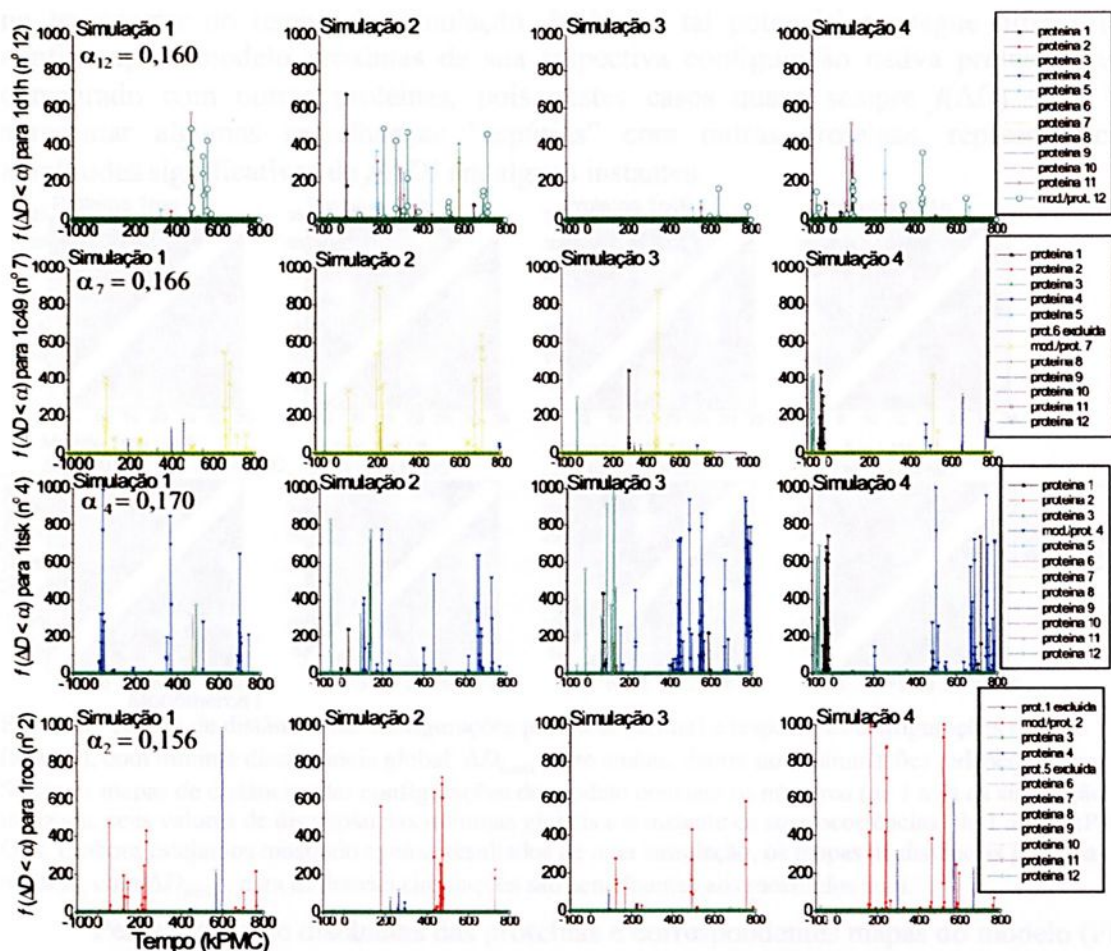


Figura 7 - Frequência da discrepância cruzada ΔD , ao longo de quatro simulações independentes (dentro um total de 9 realizadas) e para quatro proteínas “i” (o número da proteína de 1 a 27), em relação aos seus respectivos valores limiares α_i - Para todas as quatro proteínas “i” testadas, verificamos que o valor “pequeno” de α serve não apenas para compararmos configurações, mas também para distingui-las em testes cruzados modelos-proteínas, como mostrado nos gráficos. A amplitude de $f(\Delta D < \alpha)$ está relacionada à estabilidade configuracional residual, e que o número total de vezes com que a condição $\Delta D < \alpha_i$ é satisfeita está relacionado com a frequência com que a cadeia se aproxima da estrutura nativa de cada uma das proteínas. Note ainda que em todas os casos a máxima amplitude é referente a correta sequência modelo com sua proteína.

Obs. - Com 1roo fizemos $\alpha_2 = \langle \Delta D_{\min G} \rangle + 0,9$, e com 1d1h, $\alpha_{12} = \langle \Delta D_{\min G} \rangle + 1,7$, para melhor definição da aproximação modelo - proteína, já nas outras duas comparações mantivemos $\alpha = \langle \Delta D_{\min G} \rangle$.

- Devido as altas homologia entre proteínas (Figura 8) foram excluídas comparações entre a proteína 1roo com as proteínas n^{os} 1 e 5; e 1c49 com a proteínas n^o 6 (ver Figura 9).

- Tempo negativo se refere ao período de termalização (em $k_B T \rightarrow \infty$)

Observando a evolução da frequência de discrepância cruzada, ao longo da simulação (Figura 7), para o caso modelo da proteína simulada i , em relação à estrutura nativa da proteína i , notamos que o potencial tipo-hidrofóbico não proporciona estabilidade configuracional para o glóbulo, já que as grandes semelhanças (dada por $f(\Delta D < \alpha)$) não são persistentes no transcorrer do tempo de simulação, confirmando resultados das grandezas globais do enovelamento (Figura 4).

Contudo, o potencial hidrofóbico é capaz de induzir a cadeia através de conformações próximas ao estado nativo da proteína da qual se extrai o perfil HP (Figura 8), representada pelas maiores magnitudes de $f(\Delta D)$, que se apresentam sistematicamente,

no transcorrer do tempo de simulação. Inclusive tal potencial consegue diferenciar as configurações modelo próximas de sua respectiva configuração nativa protéica, quando comparado com outras proteínas, pois nestes casos quase sempre $f(\Delta D) \sim 0$, possa apresentar algumas semelhanças “espúrias” com outras proteínas, representado por amplitudes significativas de $f(\Delta D)$ em alguns instantes.

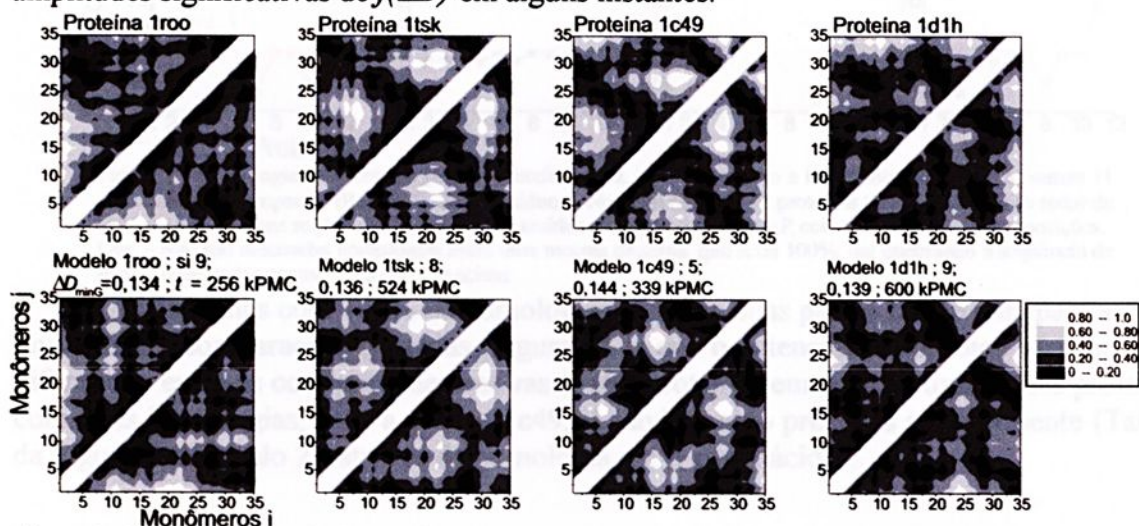


Figura 8 - Mapas de distância das configurações protéicas (acima) e respectivas configurações modelo (abaixo), com mínima discrepância global, $\Delta D_{\min G}$ entre ambas, dentre nove simulações independentes - Sobre os mapas de distâncias das configurações do modelo constam os números (de 1 a 9) da simulação utilizada, seus valores de discrepâncias mínimas globais e o instante de suas ocorrências (de 1 a 800 kPMC). Obs. Embora estejamos mostrando apenas resultados de uma simulação, os mapas de distâncias proteína-modelo, com $\Delta D_{\min G}$, para as demais simulações são semelhantes aos mostrados aqui.

Pelos mapas de distâncias das proteínas e correspondentes mapas do modelo (Figura 8), notamos que tais resultados confirmam visualmente a boa aproximação configuracional, para as proteínas, a partir das configurações do modelo simulado (Figura 7) e a eficiência do potencial hidrofóbico em induzir, a cadeia do modelo simulada, próxima à estrutura nativa protéica.

Notamos ainda que o modelo (com a sequência HP extraída de uma proteína real, Figura 7) comparado com outra proteína pode ter uma variação temporal de $f(\Delta D)$ bastante diferente, quando invertemos a proteína e o modelo. Por exemplo, a comparação 1tsk (modelo) com 1d1h (proteína) tem $f(\Delta D)$ não nula, porém a 1d1h (modelo) com 1tsk (proteína) tem $f(\Delta D)$ toda nula, ou seja, a função discrepância cruzada mínima não é invariante por inversão x e y , isto é: $\Delta D(x(t), y(p)) \neq \Delta D(y(t), x(p))$, sendo esta uma das características da função ΔD (seção 2.5 do Capítulo 2). Isso se deve às distintas rotas de enovelamento próprias de cada sequência HP de cada proteína e das peculiaridades da conformação nativa de cada proteína, como a presença de estruturas secundárias, determinando $\Delta D < \alpha$ distintos para cada uma delas.

Para a proteína 1roo (nº 2) foram excluídas comparações (Figura 7) com as proteínas nºs 1 (1c2u) e 5 (1bei) devido suas altas homologias, quanto a hidrofobicidade, (Figura 9), o mesmo ocorrendo com a proteína 1c49 (nº 7) e a 2pta (nº 6).

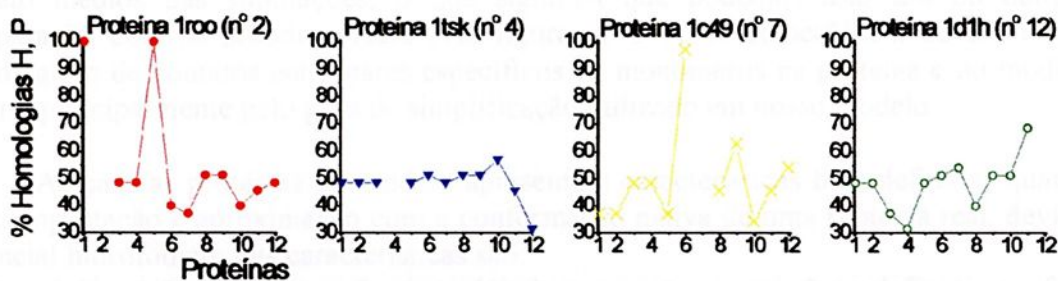


Figura 9 - Homologias percentuais para as proteínas nº 2, 4, 7, 12, quanto a hidrofobicidade, com as outras 11 proteínas mais compactas (Região I), de 35 resíduos - No geral cada par de proteínas tem homologias ao redor de 50%, ou seja, em suas seqüências há metade dos resíduos do mesmo tipo, H ou P, coincidentes nas mesmas posições. Obs. - Não são mostradas homologias entre uma mesma proteína, que seria 100%, daí quebrando a seqüência de pontos conectados por retas nos gráficos acima.

As proteínas com 100% de homologias, no caso duas para a 1roo e uma para a 1c49, inviabilizam comparações cruzadas (Figura 7), já que o potencial hidrofóbico não consegue diferenciar entre as configurações nativas destas proteínas em testes cruzados. As proteínas com altas homologias, para a 1roo e 1c49, já haviam sido previstas anteriormente (Tabela da Figura 8, Capítulo 2), através da homologia entre aminoácidos.

4.7 Comentários

Os mapas de distâncias médias das 10 estruturas, obtidas ao longo da simulação do modelo (Figura 5), ajudam a entender de que forma se dá a maleabilidade configuracional apresentada pelo modelo aqui utilizado. De fato, nas diferentes seqüências HP correspondentes às proteínas reais estudadas, uma fração significativa dos contatos entre pares de monômeros que aparecem nestas proteínas é preservada nos mapas obtidos por simulação, representadas por pontos escuros coincidentes, assim como pares de monômeros que se mantêm distantes, representados pelos pontos brancos coincidentes. Os pontos de tonalidade cinza revelam então situações não definidas estruturalmente. Apesar das características minimalistas do modelo aqui proposto, este resultado reforça nossa tese de que a seqüência específica de atributos HP da proteína é parte importante na determinação da estrutura nativa.

De fato, embora o presente modelo permita que o glóbulo apresente grande maleabilidade - com os monômeros vizinhos de um determinado monômero se aproximando, em seguida se afastando um pouco, e retornando a serem vizinhos novamente - a média, feita sobre 10 configurações independentes, tem a propriedade de realçar as condições estruturais mais permanentes e tornar mais difusos aqueles contatos intermitentes. Em outras palavras as distâncias corretas entre uma parte dos monômeros se mantêm em torno do observado nas proteínas.

É importante observar também que as estruturas das proteínas depositadas no PDB não são únicas. De fato, quando determinada por Ressonância Magnética Nuclear, existem várias estruturas nativas alternativas (Figura 6 do Capítulo 2). Utilizamos sempre a primeira destas estruturas para cada proteína, e assim seus mapas de distâncias devem ser tomados também com certa cautela nas comparações com os resultados da simulação.

Há grandes similaridades entre os mapas de distâncias e os mapas de frequências de contato médios das simulações, o que significa que podemos usar um ou outro na comparação com as proteínas reais (ver figuras 5 e 6). A coincidência de um número significativo de contatos entre pares específicos de monômeros na proteína e no modelo é notória principalmente pelo grau de simplificação utilizado em nosso modelo.

As cadeias protéicas do modelo apresentam características bem definidas quanto a sua compactação e aproximação com a conformação nativa de uma proteína real, devido o potencial hidrofóbico; tais características são:

(i) Em $k_B T = 1,5$, a conformação tipo glóbulo apresenta-se mais bem definida, conforme mostrado por grandezas como o raio de giração R_G e o desvio padrão do raio de giração Dp_G , para diversas temperaturas, Figura 3 do Capítulo 3;

(ii) O colapso da cadeia numa conformação tipo glóbulo, como expresso pela evolução temporal média do raio de giração e distância ponta a ponta, atinge valores próximos do mínimo possível, apresentando contatos energeticamente otimizados, entre monômeros hidrofóbicos e entre monômeros polares entre si e com as moléculas de solvente, conforme mostra a Figura 4;

(iii) E, como já comentado, uma fração significativa de contatos entre pares de monômeros que aparecem nas proteínas reais é reproduzida nas conformações obtidas pelo modelo ao longo da simulação, como revelado pelos resultados expostos nas Figuras 5 e 6.

As principais conclusões tiradas, como conseqüências dos itens (i), (ii) e (iii) acima, utilizando o atual modelo simplificado - com um alfabeto de duas letras - são que o potencial tipo hidrofóbico é eficiente em compactar a cadeia, mantendo o glóbulo maleável; é capaz de induzir a cadeia através de conformações próximas ao estado nativo, porém sem proporcionar definição configuracional para o glóbulo.

O teste cruzado (Figura 7), obtido utilizando-se o modelo com a seqüência HP de uma das proteínas e comparando o resultado da evolução configuracional com a estrutura nativa de todas as outras 12 proteínas mais compactas, dentre as 27 proteínas, através da frequência $f(\Delta D < \alpha)$, permitiu a constatação de que:

(1) A aproximação entre as configurações modelo (simulação) e sua respectiva estrutura nativa ocorre de forma intermitentemente, desaparecendo e voltando a surgir novamente, reproduzindo um processo recorrente;

(2) Praticamente somente para a proteína da qual extraímos a correta seqüência HP, a ser utilizada no modelo, é que tivemos $f(\Delta D < \alpha)$ significativamente diferente de zero. Assim o potencial hidrofóbico é eficaz em gerar configurações próximas da configuração nativa específica de uma proteína, ou seja, conseguindo discernir entre proteínas diferentes;

(3) Apesar da função $f(\Delta D < \alpha)$ precisar ser tratada com muito cuidado (seção 2.5 do Capítulo 2), pois seu poder de resolução é pequeno (normalmente dentro do intervalo 0,03, como mostrado na Figura 16 do Capítulo 3); através da definição do valor limiar α , variável para cada proteína, é possível diferenciar configurações de diferentes proteínas e melhorar adequadamente a comparação entre as configurações modelo com proteínas.

(4) A aproximação entre as configurações geradas pelo modelo (utilizando o potencial hidrofóbico) e sua configuração nativa da proteína correspondente (1roo, 1tsk, 1c49, ou 1d1h, Figura 8), pode ser entendida com uma predição (não muito rigorosa) das configurações nativas destas proteínas.



4.8 Bibliografia

- Baldwin**, R. L. & Rose, G. D. Is protein folding hierarchic? I. Local structure and peptide folding, *TIBS* **24**, 26-33 (1999).
- Chandler**, D., Two faces of water, *Nature* **417**, 491 (2002).
- da Silva**, R. A., Silva, M. A. A. da, Caliri, A., Deterministic folding: The role of entropic forces and steric specificities, *Journal of Chemical Physics* **114**, 4235-4242 (2001).
- Chothia**, C., Hydrophobic bonding and accessible surface area in proteins, *Nature* **248**, 338-339 (1974).
- Dill**, K. A., Dominant forces in protein folding, *Biochemistry* **29**, 7133-7155 (1990).
- Dill**, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P.D. and Chan, H. S., Principles of protein folding - A perspective from simple exact models, *Protein Science* **4**, 561-602 (1995).
- Eisenberg**, D & McLachlan, A. D., Solvation energy in protein folding and binding, *Nature* **319**, 199-203 (1986).
- Fixman**, M. & Skolnick, J., Polyelectrolyte Excluded Volume Paradox, *Macromolecules* **11**, 863-867 (1978).
- Garcia**, L. G., Treptow, W. L., and de Araújo, A. F. P., Folding simulations of a three-dimensional protein model with a nonspecific hydrophobic energy function, *Physical Rev.E* **64**, 011912-1 a 011912-5 (2001).
- Greighton**, T., Disulphide-coupled protein folding pathways, *Phil. Trans. R. Soc. Lond. B* **348**, 5-10 (1995).
- Head-Gordon**, T., Is water structure around hydrophobic groups clathrate-like?, *PNAS* **92**, 8308-8312 (1995).
- Honig**, B. & Nicholls, A., Classical Electrostatics in Biology and Chemistry, *Science* **268**, 1144-1149 (1995).
- Janin**, J., Surface and Inside Volumes in Globular Proteins, *Nature* **277**, 491-492 (1979).
- Karplus**, M. & Shakhnovich, E., Chapter 4: Protein Folding: Theoretical Studies of Thermodynamics and Dynamics; *Protein Folding*, T.E. Creighton Editor, W. H. Freeman and Company, New York, 127-193 (1992).
- Kauzmann**, W., Some factors in the interpretation of protein denaturation., *Adv. Prot. Chem.* **14**, 1-63 (1959).
- Kyte**, J. & Doolite, R., A Simple Method for Displaying the Hydropathic Character of a Protein, *J. Mol Biol.* **157**, 105-132 (1982).
- Levitt**, M. & Sharon, R., Accurate simulation of protein dynamics in solution, *PNAS* **85**, 7557-7561 (1988).
- Li**, H., Helling, R., Tang, C., Wingreen, N., Emergence of Preferred Structures in a Simple Model of Protein Folding, *Science* **273**, 666-669 (1996).
- Lum**, K., Chandler, D., Weeks, J. D., Hydrophobicity at Small and Large Length Scales, *J. Phys. Chem. B* **103**, 4570-4577 (1999).
- Pande**, V. S., Grosberg, A. Y., Tanaka, T., Heteropolymer freezing and design: Towards physical models of protein folding, *Rev. Mod. Phys.* **72**, 259-314 (2000).
- Rocha**, L. F. O., da Silva, M. A. A., Caliri A., Entropic force and folding of macromolecules, *Phys. Lett. A* **220**, 178-182 (1996).
- Rocha**, L. F. O., Tarragó, M. E. P., Caliri, A., The Water Factor in the Protein-Folding Problem, *Brazilian Journal of Physics*, **34**, 90-101 (2004).



- Socci**, N. D., Nymeyer, H., Onuchic, J. N., Exploring the protein folding funnel landscape, *Physica D* **1706**, 1-17 (1997).
- Rose**, G., Geselowitz, A., Lesser, G., Lee, R. and Zehfus, M., Hydrophobicity of Amino Acid Residues in Globular Proteins, *Science* **229**, 834-838 (1985).
- Schulz**, G. E., Schirmer, R. H. in: *Principles of Protein Structure*, Springer-Verlag, New York, Chapter 3: Noncovalent Forces Determining Protein Structure, 27-45 (1990).
- Southall**, N. T., Dill, K.A., Haymett, D.J., A View of the Hydrophobic Effect, *J. Phys. Chem. B* **106**, 521-533 (2002).
- Tarragó**, M. E. P., Rocha, L. F. O., daSilva, R. A, Caliri, A., Steric constraints as folding coadjuvant, *Physical Review E* **67**, 031901 (2003).
- Udachin**, K. & Ripmeester, J.A., A complex clathrate hydrate structure showing bimodal guest hydration, *Nature* **397**, 420-423 (1999).
- Voet**, D., Voet, J. G., Pratt, C.W., *Fundamentos de Bioquímica*, ArtMed Editora, 1ª reimpressão (2002).
- Wang**, Zi-Hao & Lee, H.C., Origin of the Native Driving Force for Protein Folding, *Phys. Rev. Lett.* **84**, 574-577 (2000).
- Wolfenden**, R., Andersson, L., Cullis, P. and Southgate, C., Affinities of Amino Acid Side Chains for Solvent Water, *Biochemistry* **20**, 849-855 (1981).

A probabilidade, p_c , da cadeia ser encontrada, numa configuração “c” particular, no ensemble microcanônico, é dada por

$$p_c = \frac{\Gamma_c}{\Gamma_o}, \quad (20)$$

onde Γ_o é o número total de estados do sistema heterogêneo, como um todo, cadeia e solvente; e Γ_c é o número de estados do sistema como um todo quando o subsistema (cadeia) é considerado congelado na configuração “c”, ou seja, é o número de estados possíveis do sistema complementar, quando a cadeia está na configuração “c”. Usando a definição

$$S = k_B \ln \Gamma, \quad (21)$$

tem-se que

$$p_c = \frac{e^{S_c/k_B}}{e^{S_o/k_B}}, \quad (22)$$

assim:

$$\sum_c p_c = \frac{\sum_c e^{S_c/k_B}}{e^{S_o/k_B}} = \frac{\sum_c \Gamma_c}{\Gamma_o} = \frac{\Gamma_o}{\Gamma_o} = 1 \quad (23)$$



CAPÍTULO 5

Conclusões

Os resultados apresentados ao longo deste trabalho são decorrentes das três premissas estabelecidas no Capítulo 1 (seção 1.1). Todos os detalhes mais relevantes da investigação levada a cabo estão explicitados nos capítulos precedentes, e assim neste Capítulo nossa intenção é sistematizar de forma resumida as nossas conclusões, face aos conceitos e idéias a respeito do fenômeno do *folding* adotadas como verdadeiras neste trabalho. É importante frisar que, a exceção do potencial hidrofóbico, todos os demais ingredientes possivelmente influenciáveis no processo do *folding* tiveram seus efeitos anulados ou reduzidos a uma condição minimalista de influência, como foi o caso das interações estéricas. Tudo isso para se analisar de forma isolada a importância do efeito hidrofóbico no processo de *folding*.

A metodologia de estudo adotada foi definida de acordo com a primeira hipótese de trabalho, a saber, que *toda informação necessária para que uma cadeia polipeptídica encontre a sua estrutura nativa está "codificada" em sua seqüência de aminoácidos*. Assim, somente esta informação foi empregada nas simulações, via método Monte Carlo. De fato, o conhecimento *a priori* da estrutura nativa somente foi utilizado como recurso de verificação dos resultados obtidos.

5.1 As proteínas adotam como estruturas nativas configurações dentre aquelas mais prováveis para uma simples cadeia em solução.

A segunda premissa estabelece os ingredientes básicos que governam o processo de *folding*: o efeito hidrofóbico e as especificidades estéricas dos aminoácidos. No presente trabalho, as especificidades estéricas dos componentes de uma cadeia polipeptídica foram minimizadas ao extremo, por meio do modelo empregado, para que as interações hidrofóbicas fossem estudadas de forma isolada. De fato, as características estéricas da macromolécula foram reduzidas às restrições que mantêm os monômeros interligados numa constituição de cadeia e o monômeros, por sua vez, todos considerados como tendo o mesmo volume, interagindo entre si através de um potencial de esfera dura, com diâmetro D . A cadeia assim obtida (colar de pérolas), foi imersa no solvente, também constituído de esferas duras de diâmetro D .

Dentre os experimentos realizados com o sistema descrito no parágrafo acima, com relação ao espaço de fase explorado ao longo da simulação, o resultado expresso pela figura 17 do Capítulo 3 é um dos principais achados deste trabalho, principalmente por sua significância, mas também por ter sido inicialmente totalmente inesperado. Fundamentalmente, todas as estruturas das 27 proteínas consideradas foram muito proximamente alcançadas (baixa discrepância estrutural) ao longo da simulação; lembrando que somente as restrições que definem a constituição de cadeia, juntamente com o efeito de volume excluído foram considerados nesta etapa. Em suma, este resultado nos permite afirmar que os padrões estruturais adotados pelas proteínas, em geral, se compõem



basicamente das conformações mais prováveis de uma cadeia livre de restrições energéticas (*random coil*) embebida no solvente, semelhante, aliás, à condição de uma cadeia polipeptídica no estado desenovelado (*unfolded*).

É importante ressaltar que este resultado está de acordo também com o fato das estruturas protéicas conhecidas serem “montadas” a partir de um número relativamente pequeno ($\sim 10^3$) de *folds* ricos em contatos próximos ao longo da cadeia (pequeno valor da ordem de contato).

5.2 O potencial hidrofóbico é eficiente para encontrar a estrutura nativa, mas não confere à mesma a necessária estabilidade energética.

A outra parte da segunda premissa adotada neste trabalho assume que o efeito hidrofóbico é um dos principais agentes que direcionam a cadeia para a estrutura nativa. De fato, de acordo com os resultados referentes às propriedades configuracionais do modelo simplificado, conforme revelados nos itens 4.5 a 4.7 do Capítulo 4, o potencial hidrofóbico apresenta duas características principais:

A primeira delas refere-se ao fato de a cadeia permanecer maleável durante toda a simulação, o que significa que o glóbulo se forma e funde intermitentemente, vários contatos (região escura indicando curtas distâncias topológicas intermonômeros) mantidos fixos (alguns corretos e outros não), e da mesma forma algumas grandes distâncias (regiões brancas) se mantêm constantes, como ilustrado pelos mapas de distância, Figura 5, do Capítulo 4. Como este mapa referente ao modelo é o resultado da média de 10 configurações independentes, os diferentes tons significam: (i)- região negra: contatos permanentes; (ii)- brancas distanciamento permanente; e (iii)- tons de cinzas: intermitência de distâncias intermonômeros.

Evidentemente não se espera identidade configuracional definitiva entre os mapas de distância das proteínas reais e as correspondentes médias dos mapas de distâncias para o modelo, pois o modelo não incorpora a necessária especificidade estérica. Contudo é notável que, embora distorcido, os correspondentes mapas revelam a existência de contatos corretos perenes, melhor revelado pelo mapa de frequência de contato médio, Figura 6, do Capítulo 4.

A segunda característica notável do potencial hidrofóbico refere-se à sua capacidade de induzir a cadeia para configurações muito próximas da nativa correspondente. Contudo, as vizinhanças das estruturas das outras nativas também podem ser eventualmente visitadas, como ilustra a Figura 7 para a comparação cruzada, entre as estruturas protéicas e as configurações do processo de *folding* simulado para as quatro proteínas reais - discrepância cruzada. Porém, a conformação com menor discrepância em relação à estrutura da nativa correta é que é acessada mais frequentemente (como se a cadeia fosse atraída para a nativa), e embora marginalmente, é a mais estável (mais alta frequência). A Figura 8, do Capítulo 4, fornece, finalmente, uma imagem de quanto próximo da nativa o simples colar de pérola pode chegar, induzida pelo potencial hidrofóbico.

5.3 No processo de *folding*, a busca pela nativa e a estabilidade da cadeia podem ser considerados como processos desacoplados.



A diferença entre a energia livre do estado nativo e do desenovelado de uma proteína globular típica é notavelmente pequena. Há razões e explicações que ajudam a entender este dado experimental, mas a observação que nos parece mais importante é aquela relacionada às peculiaridades do estado nativo, quanto à forma com que diversos fatores envolvidos no processo de *folding* são otimizados, como:

- os resíduos (ou as partes dos resíduos) polares ocupam principalmente a parte exterior do glóbulo, com o oposto ocorrendo com os apolares, conferindo solubilidade e maleabilidade à proteína;
- as ligações de hidrogênio intra cadeia são protegidas da competição energética com o meio solvente, e somente desta forma elas contribuem energeticamente para a estabilidade do sistema;
- a compactação do glóbulo é maximizada através das especificidades estéricas dos resíduos, e a ponto de tornar efetivas energeticamente as interações de van der Waals.

Assim, somente no estado nativo todos os fatores se alinham no sentido de conferir estabilidade ao glóbulo. Longe da estrutura nativa, a cadeia peptídica seria mais livre energeticamente, visitando configurações que lembram *random coil* imersa no solvente.

Neste trabalho enfocamos a questão de como se dá o acesso ao estado nativo, até mesmo porque os potenciais de esfera rígida e hidrofóbico não trazem a especificidade interacional necessária para a estabilização do glóbulo protéico.

5.4 Trabalhos futuros.

Vários tópicos devem ser ainda desenvolvidos, para que possamos aprimorar o entendimento do mecanismo de *folding*. Porém a tarefa que se segue imediatamente, é determinar todas as implicações das especificidades estéricas no processo de *folding*.

De fato simulações na rede cúbica, com 27 resíduos, utilizando-se um potencial hidrofóbico em associação com restrições estéricas (da Silva, R. A. et al (2001), Tarragó, M. E. P. et al (2003); Rocha, L. F. O. et al (2004)¹) tem mostrado que restrições estéricas adequadas podem mudar drasticamente a atividade configuracional do sistema, tendo como consequência a mudança na forma da curva da capacidade térmica e um aumento acima de cinco ordens de magnitude na frequência na qual a cadeia permanece no estado nativo.

Porém, no espaço contínuo a tarefa é mais trabalhosa, e o caminho que escolhemos tomar é o da representação da molécula em nível atômico. Numa primeira aproximação, os átomos de hidrogênio serão ignorados; todos os demais átomos assim como as moléculas de água serão considerados com uma esfera de mesmo diâmetro D . Evidentemente, o grande teste que permitira medir o nosso entendimento dos mecanismos envolvidos no processo de *folding* será prever a estrutura desconhecida de proteínas da qual só se conhece a estrutura primária.

¹ As informações completas, destas três referências, estão na seção 4.8, do Capítulo 4.



APÊNDICE

Características gerais das proteínas analisadas neste trabalho

As 27 proteínas analisadas neste trabalho, todas com 35 resíduos de aminoácidos e detalhadas aqui, têm-se grande diversidade de estruturas secundárias (como folhas β , voltas β e γ , *hairpins*, *bulges*, α -hélices) e diferentes condições de enovelamento, pois algumas possuem hetero grupos (ou resíduos além dos 20 aminoácidos padrões), e outras são domínios ou regiões de proteínas maiores, além de muitas terem diversas pontes de dissulfeto e outras proteínas possuírem apenas aminoácidos padrões.

Devido as grandes diversidades estruturais das 27 proteínas, estas são dispostas em ordem da mais compacta (n° 1c2u, com menor raio de giração), para a mais distendida (n° 27 lpeh, com maior raio de giração), sendo estas classificadas em três tipos:

Tipo I) Proteínas globulares somente com resíduos padrões e não sendo domínio de outra proteína maior, contendo as proteínas n°s

2) 1roo; 4) 1tsk; 6) 2pta; 7) 1c49; 10) 1mb6; 11) 1lqr; 12) 1dlh; 13) 1e4r; 15) 1ce4.

Tipo II) Proteínas, algumas de membrana ou são domínios, porém todas possuindo somente resíduos padrões, sendo estas as proteínas n°s

16) 1kgm; 18) 1fwo; 19) 1kjo; 20) 1zwd; 22) 1ijj; 23) 1spf; 24) 1zta.

Especificamente tem-se: n° 16) Domínio com resíduos na seqüência de 57-91; n° 18) Domínio: 382-416; n° 19) Domínio: 20-54; n° 20) Domínio: 3-37; n° 22) Domínio: 650-684; n° 23) não globular (lipoproteína); n° 24) não globular (Motivo Ligante-DNA).

Tipo III) Proteínas com pelo menos um resíduos não padrão e algumas também são domínios e outras não são globulares, com as proteínas n°s

1) 1c2u; 3) 1txm; 5) 1bei; 8) 1quz; 9) 1c4e; 14) 1zdc; 17) 1zwf; 21) 1zwg; 25) 1bde; 26) 1rh4; 27) 1peh.

Tem-se os seguintes resíduos não padrões

1) ACE, na primeira posição da seqüência de aminoácidos das proteínas 1zwf (n° 17), 1bde (n° 25), 1rh4 (n° 26), 1peh (n° 27);

2) NH₂, na última posição (ou 35^a) das proteínas 1txm (n° 3), 1quz (n° 8), 1zdc (n° 14), 1bde (n° 25), 1rh4 (n° 26), 1peh (n° 27);

3) DNP, na vigésima segunda posição da proteína 1bei (n° 5);

4) ABA, nas posições três e trinta e cinco da proteína 1c2u (n° 1);

5) PCA, na primeira posição da proteína 1c4e (n° 9);

6) ILL, nas posições sete, dezoito e vinte e nove da proteína 1rh4 (n° 26).

7) SIN, na primeira posição da seqüência de aminoácidos da proteína 1zwg (n° 21).

Tem-se ainda as proteínas n°s

17 e 21) Domínio: 4 - 37;

26) Tetrâmero;

27) Domínio: 236 - 268.



Quanto aos heterogrupos, presentes nas proteínas de Tipo III, tem-se:

Hetero grupos	Nome	Fórmula
ACE	Grupo Acetil	C_2H_3O
NH ₂	Grupo Amino	H_2N
DNP	Dinitrofenilene	$C_6H_4N_2O_4$
ABA	Ácido Alfa-aminobutírico	$2(C_4H_9NO_2)$
PCA	Ácido Piroglutamico	$C_5H_9NO_2$
IIL	Iso-isoleucina	$3(C_6H_{13}NO_2)$
SIN	Ácido Sucínico	$C_4H_6O_4$

As conformações nativas das 27 proteínas, utilizadas no atual trabalho e visualizadas pelo programa gráfico RasMol, são mostradas a seguir, dentro de um retângulo, em forma de fitas coloridas, com as seguinte cores para suas estruturas secundária: magenta (α -hélices), amarelo (folha β), azuis (volta β).

Acima de cada proteína consta seu código de um número seguido por 3 letras utilizado pelo Banco de Dados de Proteínas ("PDB"). As proteínas tem suas estruturas nativas determinadas por RMN (Ressonância magnética nuclear), a exceção da 1rh4 (n° 26) que é determinada por Raio X. As proteínas são mostradas segundo seus valores de raios de giração, estando numa seqüência que vai da proteína mais compacta n° 1, 1c2u, para a mais distendida n° 27, 1peh.

Para cada proteína além de sua estrutura nativa disposta em fitas coloridas, há ao lado desta, um resumo de

- As estruturas secundárias: hélices (indicadas também por hélices), voltas β (representada pela letra β), folha β (representadas por setas), γ turn (indicadas pela letra γ);
- Pontes dissulfeto, representada por retas verde pontilhadas, conectando círculos amarelos, representando os resíduos de cisteínas;
- A seqüência de resíduos, inclusive de aminoácidos não padrão (quando presentes nas proteínas), num código de uma letra e abaixo seu número na seqüência da proteína;
- Informações sobre as estruturas secundárias, como suas localizações e tipos;

Há ainda um resumo das principais informações de cada proteína, em geral tendo:

Classificação e o nome (fornecido pela base de dados "SCOP") da proteína;

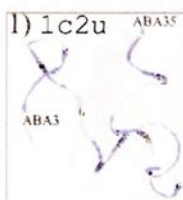
Título (ou *Title*), normalmente a referência completa (e.g. Revista publicada) não está disponível, daí inserimos apenas seu título em inglês;

Estrutura (molécula ou o nome da proteína);

Fonte (Origem, seguida pelo nome da Espécie, contendo seu nome científico e comum);

Outros detalhes revelantes sobre a proteína.

As vezes algumas proteínas podem possuir mais informações que outras, como disponibilizado pelo "PDB" (**Berman**, H.M. et al.), como seu Componente Celular, Processo Biológico, Função Molecular.



β -turns

Localização	Tipo	... continuação	Localização	Tipo
4 - 7	IV	21 - 24	I	
13 - 16	I	22 - 25	IV	
14 - 17	IV	23 - 26	IV	
15 - 18	IV	25 - 28	IV	
16 - 19	IV	29 - 32	I	
18 - 21	IV	30 - 33	IV	
19 - 22	IV			

γ -turns

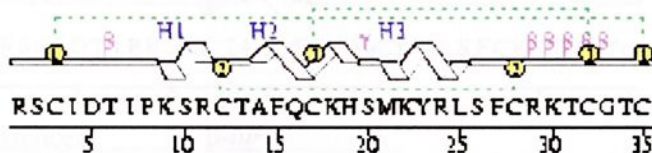
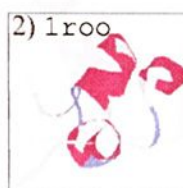
Localização	Tipo
7 - 9	Inversa
19 - 21	Inversa

Classificação: Toxina ; Proteína: Toxina k anemone do mar.

Title: *Solution structure of [abu3,35]shk12-28,17-32.*

Estrutura (Proteína): Peptídeo sintético análogo da toxina *shk*.

Fonte: Sintética. Mutação: sim. Outros detalhes: O peptídeo foi quimicamente sintetizado. A sequência deste peptídeo é naturalmente encontrada na *Stichodactyla helianthus* (*Sun anemone*).



Hélices

Localização	Tipo
9 - 11	G
14 - 19	H
21 - 25	H

β -turns

Localização	Tipo
5 - 8	VIII
28 - 31	IV
29 - 32	IV
30 - 33	IV
31 - 34	IV
32 - 35	IV

γ -turn

Localização	Tipo
19 - 21	Inversa

Classificação: Inibidor de canal de Potássio ; Proteína: Toxina k anemone do mar.

Title: *Nmr solution structure of shk toxin, nmr, 20 structures*

Estrutura: Toxina *Shk*. Sinônimo: *shktx*.

Fonte: Sintética. *Stichodactyla helianthus* (*Sun anemone*).

3) 1txm (MauroToXin)



Strands

Localização	Sheet
23 - 24	A
29 - 30	A

Hélice

Localização	Tipo
6 - 16	H

β -turns

Localização	Tipo
19 - 22	IV
25 - 28	IV
31 - 34	IV

β -hairpin

Localização	Classe
23 - 30	2: 4

Classificação: Neurotoxina ; Proteína: Maurotoxina

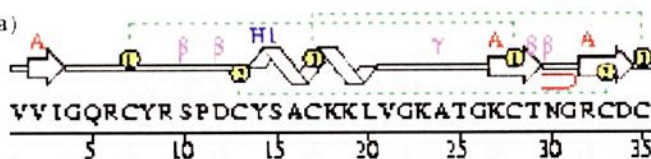
Title: *Scorpion toxin (maurotoxin) from scorpio maurus, nmr, 35 structures*

Estrutura: Maurotoxina.

Fonte: Sintética. *Scorpio maurus* (Escorpião). Organismo: *telson*. Secreção: veneno

Componente Celular: região extracelular ; Processo Biológico: patogênese; Função Molecular: inibidor da atividade de canal iônico.

4) 1tsk (TS Kapa)



Strands		Hélice		β-turns		β-hairpin	
Localização	Sheet	Localização	Tipo	Localização	Tipo	Localização	Classe
2 - 3	A	14 - 20	H	9 - 12	IV	27 - 34	2: 2 IP
27 - 29	A			11 - 14	IV		
32 - 34	A			28 - 31	IV		
				29 - 32	I		

γ-turn	
Localização	Tipo
23 - 25	Inversa

Classificação: Neurotoxina ; Proteína: Toxina ts kappa

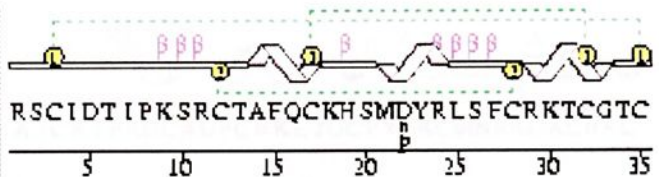
Title: *Scorpion toxin (ts kappa) from tityus serrulatus active on small conductance potassium channel, nmr, 30 structures*

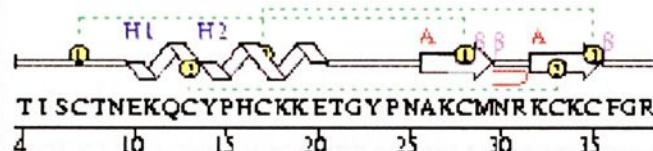
Estrutura: Ts kapa.

Fonte: Sintética. *Tityus serrulatus* (Escorpião Brasileiro). Secreção: veneno

Componente Celular: região extracelular; Processo Biológico: patogênese; Função Molecular: inibidor da atividade de canal iônico.

5) 1bei





Strands		Hélices		β-turns		β-hairpin	
Localização	Sheet	Localização	Tipo	Localização	Tipo	Localização	Classe
26-29	A	10-13	G	28-31	IV	26-35	2:2 IP
32-35	A	14-20	H	29-32	V		
				35-38	IV		

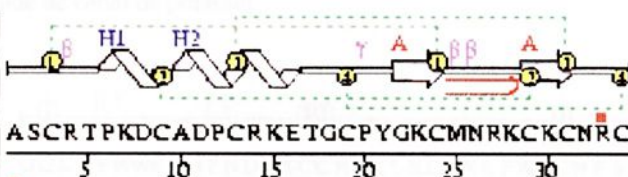
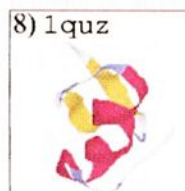
Classificação: Toxina ; Proteína: Toxina *Pandinus*

Title: *Structural and functional differences of two toxins from the scorpion pandinus imperator*

Estrutura: Toxina k-beta. Fragmento: peptídeo completo. Sinônimo: *pitx-kb*, *alfa-ktx5.2*.

Fonte: *Pandinus imperator* (Escorpião Imperador). Secreção: veneno. Expresso em: *escherichia coli*.

Componente Celular: região extracelular; Processo Biológico: patogênese; Função Bioquímica: inibidor da atividade de canal iônico.



■ Resíduos em interação com ligante

Strands		Hélices		β-turns		β-hairpin	
Localização	Sheet	Localização	Tipo	Localização	Tipo	Localização	Classe
22-24	A	6-9	G	3-6	IV	22-31	2:4
29-31	A	10-16	H	24-27	IV		
				25-28	V		

γ-turn

Localização	Tipo
19 - 21	Inversa

1 Interação hélice-hélice

Nº resíduos interagindo	
Hélice1	Hélice2
4	5

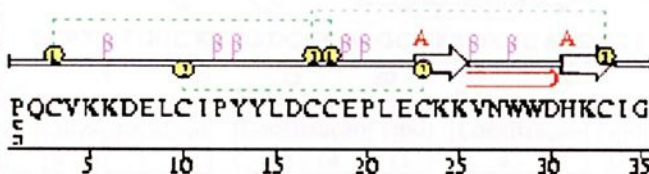
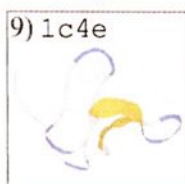
Classificação: Toxina ; Proteína: Toxina *hstx1* atuando em canal de potássio

Title: *Solution structure of the potassium channel scorpion toxin hstx1*

Estrutura: Toxina *Hstx1*. Outros detalhes: quimicamente sintetizado.

Fonte: Sintética. Outros detalhes: sintetise química em fase sólida usando a metodologia *fmoc*. A sequência deste peptídeo ocorre naturalmente em escorpiões (*scorpionidae*).

Componente Celular: região extracelular ; Processo Biológico: patogênese; Função Molecular: inibidor da atividade de canal iônico



Strands		β-turns		β-hairpin	
Localização	Sheet	Localização	Tipo	Localização	Classe
23 - 25	A	5 - 8	II	23 - 33	3 : 5
31 - 33	A	11 - 14	II		
		12 - 15	IV		
		18 - 21	IV		

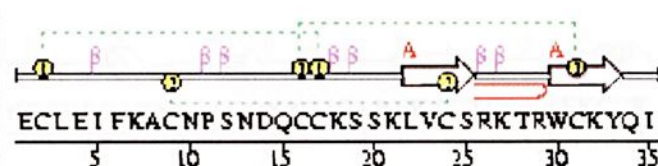
... continuação	
Localização	Tipo
19 - 22	IV
25 - 28	IV
27 - 30	IV

Classificação: Proteína de Planta ; Proteína: Polipeptídeo *Gurmarin*, um supressor do gosto doce.

Title: *Gurmarin from gymnema sylvestre*

Estrutura: *Gurmarin*. Outros detalhes: sintético; quimicamente sintetizado

Fonte (Espécie): *Gymnema sylvestre*. Localização (tecido): Folhas.



Localização	Sheet
22 - 25	A
30 - 33	A

Localização	Tipo
4 - 7	IV
10 - 13	I
11 - 14	VIII
17 - 20	I

Localização	Tipo
18 - 21	I
25 - 28	IV
26 - 29	I

Localização	Classe
22 - 33	4:4 I

<i>β-bulge</i>	
Localização	Tipo
25A 29A 30A	Antiparalela GI

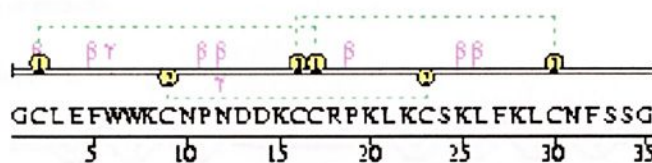
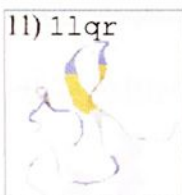
Classificação: Toxina ; Proteína: *Huwen* toxina-IV

Title: *Three dimensional solution structure of huwentoxin-iv by 2d 1h-nmr*

Estrutura: *Huwentoxin-iv*.

Fonte: *Selenocosmia huwena* (Chinese bird spider). Secreção: veneno

Componente Celular: região extracelular; Processo Biológico: patogênese; Função Molecular: inibidor de atividade de canal de potássio.



Localização	Tipo
1 - 4	VIII
4 - 7	IV
10 - 13	IV
11 - 14	IV

Localização	Tipo
18 - 21	I
24 - 27	IV
25 - 28	IV

Localização	Tipo
5 - 7	Clássica
11 - 13	Clássica

Classificação: Toxina ; Proteína: *GSmtx4*

Title: *Solution structure of a toxin from the tarantula, grammostola spatulata, which inhibits mechanosensitive ion channels*

Estrutura: *Gsmtx4*.

Fonte: *Grammostola spatulata* (Tarântula rosa Chilena). Outros detalhes: Veneno



Localização	Sheet
19 - 21	A
28 - 30	A

Localização	Tipo
12 - 14	G

Localização	Tipo
4 - 7	II
16 - 19	IV
22 - 25	IV
23 - 26	IV

Localização	Classe
19 - 30	4: 6

Localização	Tipo
2 - 4	Inversa
22 - 24	Inversa

β -bulge	23 - 26	I
Localização	Tipo	
22A 26A 27A	Antiparalela GI	

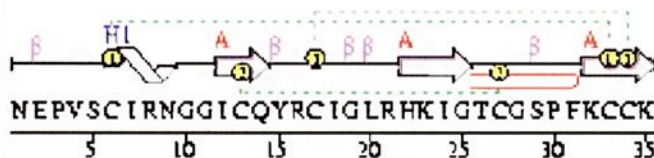
Classificação: Toxina ; Proteína: *Hana* toxina 1

Title: *Solution structure of hanatoxin 1*

Estrutura: Hanatoxina tipo 1.

Fonte: *Grammostola spatulata* (Tarântula). Secreção: Veneno

Componente Celular: região extracelular; Processo Biológico: patogênese; Função Molecular: inibidor da atividade de canal de potássio.



Strands		Hélice		β-turns		... continuação	
Localização	Sheet	Localização	Tipo	Localização	Tipo	Localização	Tipo
12 - 14	A	6 - 9	G	1 - 4	I	19 - 22	IV
22 - 25	A			14 - 17	VIII	28 - 31	I
32 - 35	A			18 - 21	IV		

β-bulge		β-hairpin	
Localização	Tipo	Localização	Classe
34A 23A 24A	Antiparalela larga	22 - 35	7: 7

BETA ALPHA BETA UNITS	
Start	End
12	25

Classificação: De defesa ; Proteína: Beta-defensin, BD

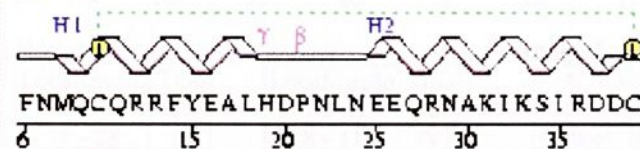
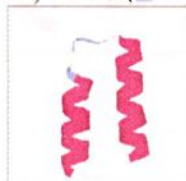
Title: *Solution structure of the mouse defensin mbd-8*

Estrutura: Beta-defensin (BD) 8. Sinônimo: mbd-8

Fonte: Sintético. Espécie: *Mus musculus* (Rato).

Componente Celular: região extracelular ; Processo Biológico: reação (resposta) de defesa e reação a peste, patogênese or parasita.

14) 1zdc (Domain, Z34C)



Hélices		β-turn		γ-turn	
Localização	Tipo	Localização	Tipo	Localização	Tipo
8-18	H	20 - 23	IV	18 - 20	Inversa
25-38	H				

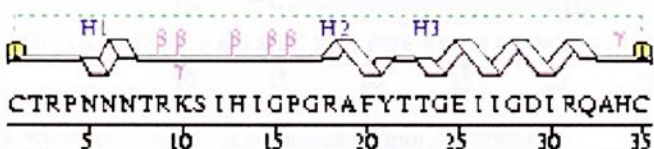
Classificação: Domínio de ligação Igg; Proteína: Imita domínio de ligação da proteína A Ig(Fc).

Title: *Disulfide-stabilized mini protein a domain, z34c, nmr, 24 structures*

Estrutura: Um Domínio, z34c, Mini proteína estável.

Outros detalhes: estável, duas-hélices, variante de ligação igg de uma proteína.

Fonte: Sintética: sequência difere daquela da z38 (entrada 1zda) tendo cinco resíduos removidos do N-terminal, uma cisteína adicionada no C-terminal, e gln 10 substituída por cisteína para forma uma ponte dissulfeto.



Hélices		β-turn		... continuação		γ-turn	
Localização	Tipo	Localização	Tipo	Localização	Tipo	Localização	Tipo
5 - 7	G	8 - 11	IV	14 - 17	IV	9 - 11	Inversa
18 - 21	G	9 - 12	IV	15 - 18	IV	33 - 35	Inversa
23 - 32	H	12 - 15	IV				

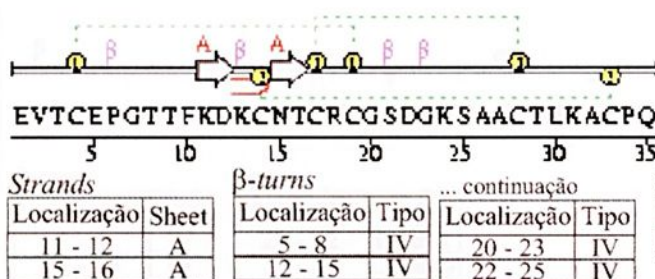
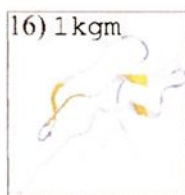
Classificação: Hélice Anfipática ; Proteína: Fragmentos loop V3

Title: *Conformational model for the consensus v3 loop of the envelope protein gp120 of HIV-1*

Estrutura: Loop V3 da proteína envelope de HIV-1. Unidade Biológica: monomero.

Outros detalhes: Este peptídeo, loop v3, é um fragmento da proteína envelope gp120 de HIV-1.

Fonte: Sintética. Outros detalhes: sequência tomada do vírus da imunodeficiência humana tipo 1 (clone 12), gene env.



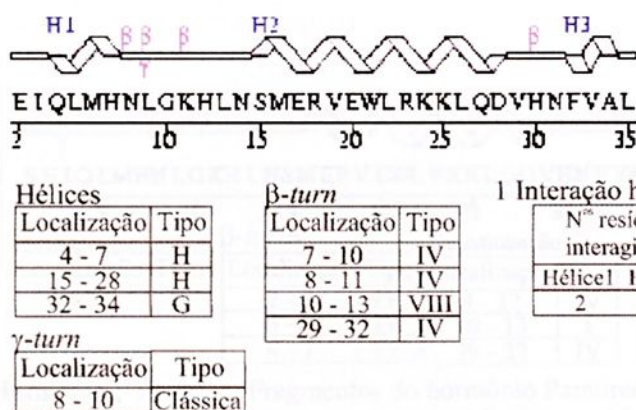
Classificação: Hidrolase ; Proteína: Inibidor Protease SGCI

Title: *Solution structure of the small serine protease inhibitor sgci.*

Estrutura: Inibidor i Serina protease. Fragmento: resíduos entre 57-91. Sinônimo: inibidor quimotripsina (*sgci*) em *schistocerca gregaria*.

Fonte: Sintética. Outros detalhes: síntese em fase sólida, estratégia *fmoc*, sequência naturalmente encontrada em *desert locust* (*schistocerca gregaria*).

Função Molecular: Atividade inibidor protease.



Classificação: Hormônio ; Proteína: Fragmentos do hormônio Paratireóide

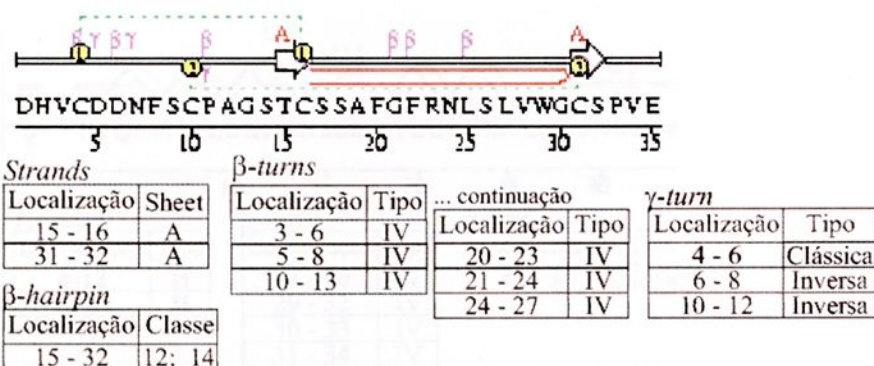
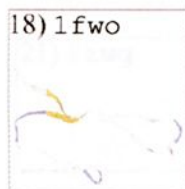
Title: *Structure of N-terminal acetylated human parathyroid hormone, nmr, 10 structures*

Estrutura: Fragmento do hormônio paratireóide. Fragmento: resíduos entre 4 - 37.

Sinônimo: N-acetil-*hpth* (4-37). Mutação: N-terminal acetilado.

Fonte: Sintética. *Homo sapiens* (Humano).

Componente Celular: região extracelular ; Função Molecular: atividade hormonal.

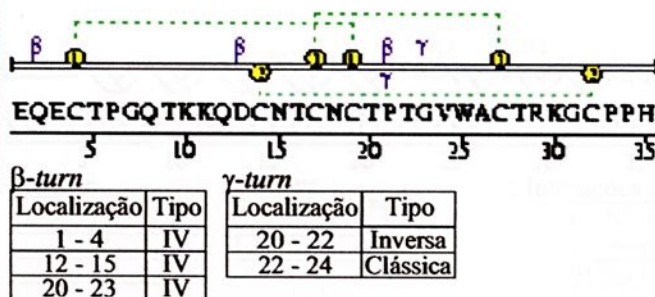
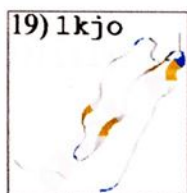


Classificação: Hidrolase ; Proteína: Cadeia beta *Oryzain*

Title: *The solution structure of a 35-residue fragment from the granulin/epithelin-like subdomain of rice oryzain beta (rob 382-416 (c398s,c399s,c407s,c413s))*

Estrutura: Cadeia beta *Oryzain*. Fragmento: C-terminal tipo *granulin/epithelin* extensão (resíduos entre 382-416). Mutação: sim

Fonte: Sintética. *Oryza sativa* (Arroz). Outros detalhes: o peptídeo foi quimicamente sintetizado. A sequência do peptídeo naturalmente ocorre em *oryza sativa* (arroz).

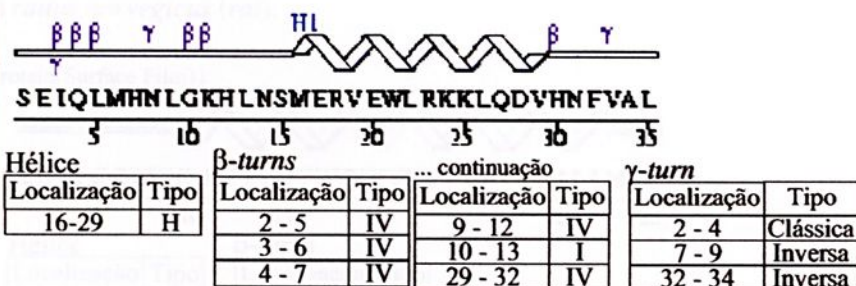
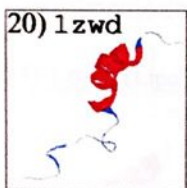


Classificação: Hidrolase ; Proteína: Inibidor Protease SGCI

Title: *Solution structure of the small serine protease inhibitor sgci.*

Estrutura: Inibidor i Serina protease. Fragmento: resíduos 20-54. Sinônimo: inibidor tripisina (sgti) em *schistocerca gregaria*.

Fonte: Sintética. Outros detalhes: síntese do peptídeo em fase sólida, estratégia *fmoc*, sequência naturalmente encontrada em *schistocerca gregaria* (*desert locust*).



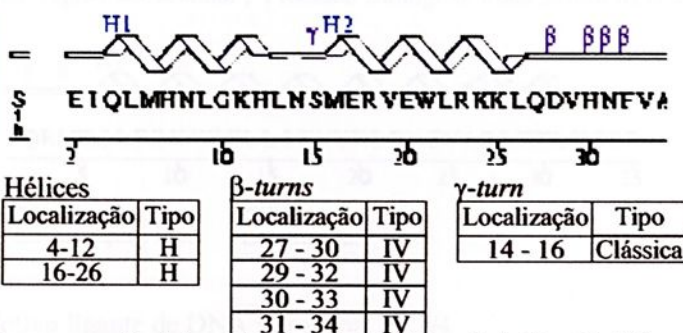
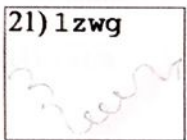
Classificação: Hormônio ; Proteína: Fragmentos do hormônio Paratireóide

Title: *Structure of human parathyroid hormone fragment 3-37, nmr, 10 structures*

Estrutura: Fragmento do Hormônio paratireóide. Fragmento: 3 - 37. Sinônimo: *hpth*(3-37).

Fonte: Sintética. *Homo sapiens* (Humano).

Componente Celular: região extracelular; Função Molecular: atividade hormonal.



Classificação: Hormone ; Proteína: Fragmentos do hormônio Paratireóide

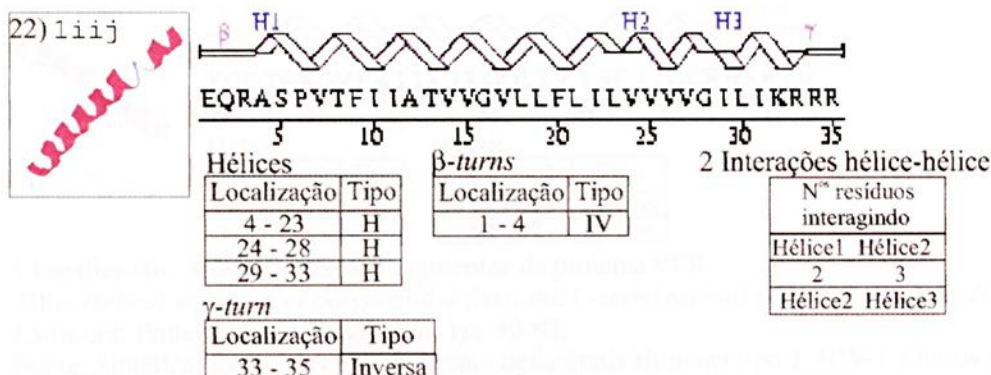
Title: *Succinyl human parathyroid hormone 4-37, nmr, 10 structures*

Estrutura: Fragmento do Hormônio Paratireóide. Fragmento: 4 - 37. Sinônimo: *n-succinyl-hpth*(4-37). Mutation: N-terminal Succinilado.

Fonte: Sintética. *Homo sapiens*. Humano.

Componente Celular: região extracelular; Função Molecular: atividade hormonal.

Obs.: Para esta proteína o programa RasMol, não destinou tonalidade de cores às suas estruturas secundárias.



Classificação: Proteína de sinalização; Proteína: Segmento Neu/erbB-2 atravessando membrana

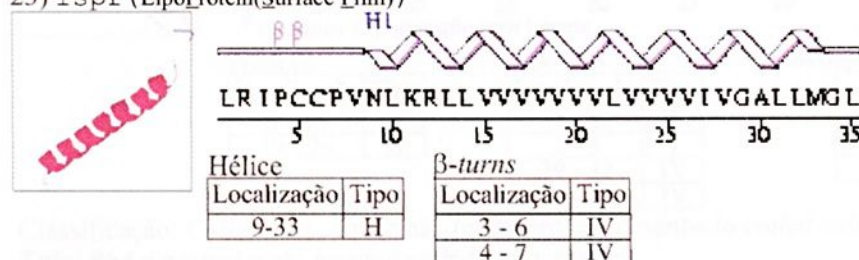
Title: *Solution structure of the neu/erbB-2 membrane spanning segment*

Estrutura: Proteína tyrosine kinase (1iij) do Receptor *ErbB-2*. Fragmento: Domínio transmembrana, resíduos 650-684.

Fonte: Sintética, baseada na sequência de *rattus norvegicus* (rat). *Rattus norvegicus*.

Outros detalhes: Este peptídeo foi quimicamente sintetizado. Esta sequência ocorre naturalmente em *rattus norvegicus* (rat).

23) 1spf (LipoProtein(Surface Film))



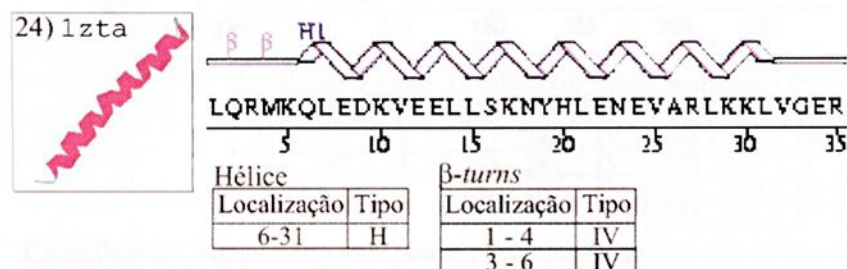
Classificação: Lipo proteína (*surface film*) ; Proteína: Polipeptídeo C associado a surfactante pulmonário

Title: *The NMR structure of the pulmonary surfactant-associated polypeptide SP-C in an apolar solvent contains a valyl-rich -helix.*

Estrutura: Surfactante Pulmonário associado ao polipeptídeo C(sp-c).

Fonte: *sus scrofa* (Porco).

Componente Celular: região extracelular ; Processo Biológico: troca gasosa na respiração.



Classificação: Motivo ligante de DNA ; Proteína: GCN4

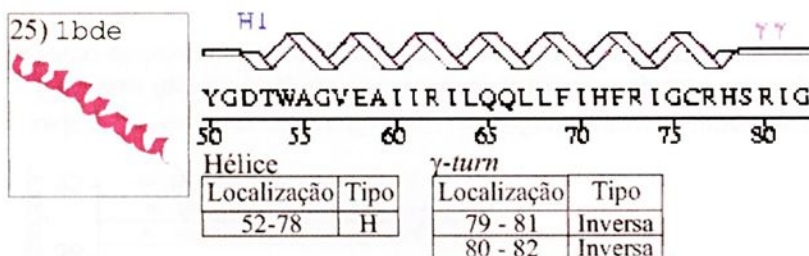
Title: *The solution structure of a leucine-zipper motif peptide.*

Estrutura: Monomero zíper de Leucina

Fonte: Polipeptídeo sintético atravessando a sequência do ativador transcricional gen4 proteína zíper leucine. *Saccharomyces cerevisiae* (Baker's yeast).

Componente Celular: núcleo; Processo Biológico: regulação de transcrição, DNA-dependente;

Função Molecular: ligante de DNA.



Classificação: Aids ; Proteína: Fragmentos da proteína VPR

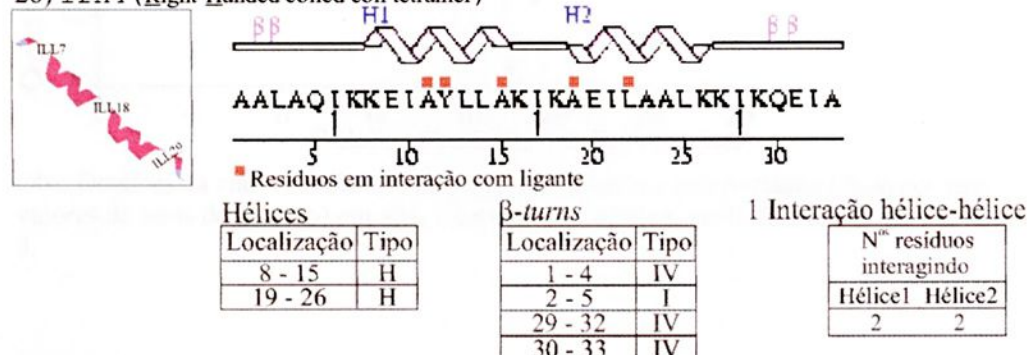
Title: *Helical structure of polypeptides from the C-terminal half of HIV-1 vpr, nmr, 20 structures*

Estrutura: Proteína Vpr . Fragmento: vpr 50-82.

Fonte: Sintética. Espécie: Vírus da Imuno deficiência Humana tipo 1. HIV-1. Outros detalhes: resíduos 50 - 82 da vpr HIV.

Função Molecular: Desconhecida.

26) 1rh4 (Right-Handed coiled coil tetramer)



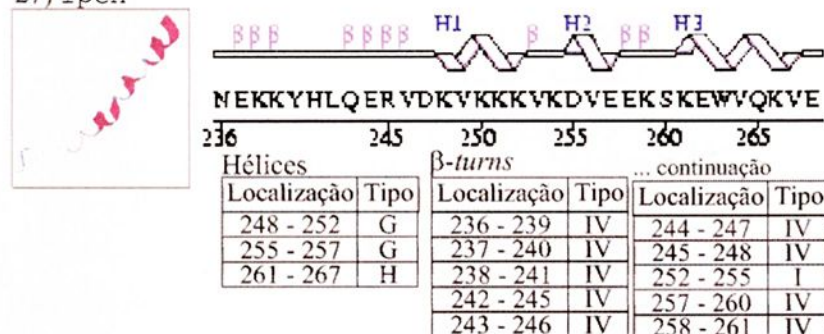
Classificação: Coiled coil ; Proteína: Tetrâmero rh4 desenhado coiled coil com volta à direita

Title: *Rh4 designed right-handed coiled coil tetramer*

Estrutura: Tetrâmero coiled coil com "volta" a direita. Unidade Biológica: tetrâmero

Fonte: Sintética: Polímero construído sinteticamente; Espécie: Sequência sintética não biológica.

27) 1peh



Classificação: Nucleotídeo transferase ; Proteína: Peptídeo alfa-hélice de ligação à membrana da transferase *Cytidyl transferase*

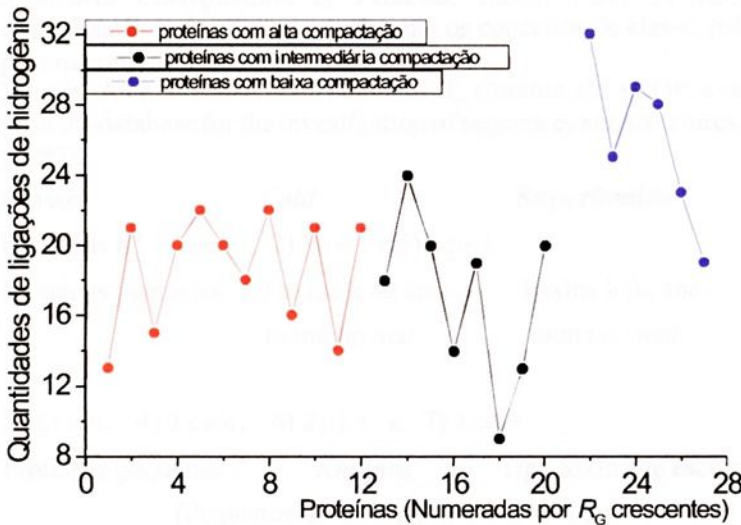
Title: *Nmr structure of the membrane-binding domain of ctp phosphocholine cytidyltransferase, 10 structures*

Estrutura: *Pepnh1*. Fragmento: resíduos 236 - 268. Sinônimo: Domínio peptídeo que liga a membrana à *cytidyl transferase*. Mutação: I263w

Fonte: Sintética. *Rattus norvegicus* (Rat.). Órgão: fígado

Processo Biológico: biosíntese; Função Molecular: atividade de nucleotídeo transferase.

Quanto as quantidades de ligações de hidrogênio intra-cadeia referentes a cada proteína, o programa gráfico RasMol, fornece seus prováveis valores (a exceção da proteína 1zwg, n° 21, onde tal informação não foi disponibilizada), mostrados no gráfico abaixo.



Obs. Detalhes da classificação das 27 proteínas, quanto a compactação (dada por seus valores de raios de giração) em alta, intermediária e baixa, serão fornecidas no Capítulo 3.

Para identificar relacionamentos de sequência e estrutura, entre as proteínas deste trabalho e também para auxiliar o entendimento de cada uma destas proteínas, utilizamos uma classificação estrutural e evolutiva, da base de dados SCOP (iniciais do inglês *Structural Classification of Proteins*, versão 1.69). A SCOP está baseada numa classificação que utiliza gradualmente os conceitos de classe, *fold*, superfamília, família, para maiores detalhes ver

Murzin, A.G, Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**:536-540 (1995).

Classe	Fold	Superfamília	Família
Proteínas nº 1) 1c2u, 2) 1roo e 5) 1bei			
Proteínas pequenas	Toxina k da anemona do mar	Toxina k da anemona do mar	Toxina k da anemona do mar
3) 1txm, 4) 1tsk, 6) 2pta e 7) 1c49			
Proteínas pequenas	<i>Knottins</i> (Pequenos inibidores, toxinas, lectinas)	Tipo toxina de escorpião	Tipo toxina de escorpião
8) 1quz			
Proteínas desenhadas	Desenhos baseados em toxina de escorpião	Desenhos baseados em toxina de escorpião	Toxina hstx1 em canal de Potássio
9) 1c4e			
Proteínas pequenas	<i>Knottins</i> (Pequenos inibidores, toxinas, lectinas)	Tipo gurmarina	Gurmarina, um polipeptídeo que elimina o gosto de doce
10) 1mb6, 11) 1lqr e 12) 1dlh			
Proteínas pequenas	<i>Knottins</i> (Pequenos inibidores, toxinas, lectinas)	Tipo toxina omega	Toxinas de aranhas
13) 1e4r			
Proteínas pequenas	Tipo defensiva	Tipo defensiva	Defensiva
14) 1zdc			
Proteínas desenhadas	Imita domínio de ligação da proteína A Ig(Fc)	Imita domínio de ligação da proteína A Ig(Fc)	Imita domínio de ligação da proteína A Ig(Fc)
15) 1ce4			
Peptídeos	Fragmentos de proteína do envelope gp120 do HIV-1	Fragmentos de proteína do envelope gp120 do HIV-1	Fragmentos de proteína do envelope gp120 do HIV-1

Classe	Fold	Superfamília	Família
16) 1kgm e 19) 1kjo			
Proteínas pequenas	Inibidores PMP	Inibidores PMP	Inibidores PMP
Proteínas nºs 17) 1zwf, 20) 1zwd e 21) 1zwg			
Peptídeos	Fragmentos do hormônio da paratireóide	Fragmentos do hormônio da paratireóide	Fragmentos do hormônio da paratireóide
18) 1fwo			
Proteínas pequenas	<i>Knottins</i> (Pequenos inibidores, toxinas, lectinas)	Repete a Granulina	Repete a Granulina
22) 1iiij e 23) 1spj			
Peptídeos	Fragmentos helicoidal Transmembrana	Fragmentos helicoidal Transmembrana	Fragmentos helicoidal Transmembrana
24) 1zta			
Proteínas <i>coiled-coil</i>	<i>Coiled-coil</i> paralela	Domínio <i>zipper</i> Leucine	Domínio <i>zipper</i> Leucine
25) 1bde			
Peptídeos	Fragmentos da proteína VPR	Fragmentos da proteína VPR	Fragmentos da proteína VPR
26) 1rh4			
Proteínas desenhadas	Tetrâmero rh4 desenhado <i>coiled coil</i> com volta à direita	Tetrâmero rh4 desenhado <i>coiled coil</i> com volta à direita	Tetrâmero rh4 desenhado <i>coiled coil</i> com volta à direita
27) 1peh			
Peptídeos	Peptídeo de ligação à membrana da transferase <i>Cytidylyl phosphocholine</i>	Peptídeo de ligação à membrana da transferase <i>Cytidylyl phosphocholine</i>	Peptídeo de ligação à membrana da transferase <i>Cytidylyl phosphocholine</i>

Quanto ao itens Classe, *Fold* e Família temos as seguintes observações

Classe

Proteínas pequenas - usualmente com predomínio de ligantes metálicos, heme, e/ou pontes dissulfeto.

Proteínas desenhadas - estruturas experimental de proteínas com seqüências essencialmente não natural. Não é uma classe verdadeira.



Peptídeos - peptídeos e fragmentos. Não é uma classe verdadeira.

Proteínas *coiled-coil* - Não é uma classe verdadeira.

Fold

Toxina k da anemona do mar - *fold* todo alfa, abundante em pontes dissulfeto.

Knottins (Pequenos inibidores, toxinas, lectinas) - *fold* ligado por pontes dissulfeto, contendo beta-hairpin com duas pontes dissulfeto adjacentes.

Tipo defensiva - *fold* abundante em pontes dissulfeto, quase todo beta.

Inibidores PMP - *fold* abundante em pontes dissulfeto; todo beta:3 *strands* anti paralelas.

Coiled-coil paralela - este não é um *fold* verdadeiro; inclui oligômeros de mais curtas hélices idênticas.

Família

Fragmentos helicoidal transmembrana - os membros desta família podem não estar relacionados.



