**UNIVERSIDADE DE SÃO PAULO**

**FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO**

# O papel do código estereoquímico e das flutuações térmicas locais no processo de *folding* de proteínas

João Paulo Dal Molin

Ribeirão Preto
2011

**UNIVERSIDADE DE SÃO PAULO**

**FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO**

# O papel do código estereoquímico e das flutuações térmicas locais no processo de *folding* de proteínas

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas para obtenção do Título de Doutor em Ciências

Área de Concentração: Ciências Farmacêuticas – Física Biológica.

**Orientado:** João Paulo Dal Molin

**Orientador:** Antonio Caliri

Ribeirão Preto
2011

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Dal Molin, João Paulo

O papel do código estereoquímico e das flutuações térmicas locais no processo de *folding* de proteínas. Ribeirão Preto, 2011.
92 p.:Il; 30cm

Tese de Doutorado, apresentada à Faculdade de Ciências Farmacêuticas de Ribeirão Preto/USP – Área de concentração: Ciências Farmacêuticas – Física Biológica.

Orientador: Caliri, Antonio.

1. *Folding* de proteínas. 2. Código estereoquímico. 3. Flutuação térmica local.

**FOLHA DE APROVAÇÃO**

Nome do aluno: João Paulo Dal Molin
Título do trabalho: O papel do código estereoquímico e das flutuações térmicas locais no processo de *folding* de proteínas

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas para obtenção do Título de Doutor em Ciências

Área de Concentração: Física Biológica.

Orientador: Antonio Caliri

Aprovado em:

Banca Examinadora

Prof. Dr. _____

Instituição: _____ Assinatura:_____

Prof. Dr. _____

Instituição: _____ Assinatura:_____

Prof. Dr. _____

Instituição: _____ Assinatura:_____

Prof. Dr. _____

Instituição: _____ Assinatura:_____

Prof. Dr. _____

Instituição: _____ Assinatura:_____

## DEDICATÓRIA

Aos meus pais, Lorena Dalberto Dal Molin, Sérgio Odir Dal Molin e a minha irmã Denise Maria Dal Molin.

# AGRADECIMENTOS

Ao professor Dr. Antonio Caliri por ter me aceito como aluno de doutorado. Também agradeço a todos os alunos anteriores que investiram parte de sua formação acadêmica no agora chamado modelo estereoquímico (todos os seus nomes e contribuições se encontram no Apêndice A). Sem o trabalho deles e a orientação do professor Caliri esta tese não existiria, meu muito obrigado.

Agradeço aos colegas e amigos do laboratório: Ricardo Oliveira dos Santos Soares, Guilherme Thomaz Pereira Brancini, Rodrigo Antonio Faccioli, Pablo Andrei Silva, Renata Almeida Garcia Reis, Leandro Oliveira Bortot e Túlio Marcus Ribeiro Calixto, pelo ambiente.

Agradeço ao Programa de Pós-Graduação em Ciências Farmacêuticas da Faculdade de Ciências Farmacêuticas de Ribeirão Preto, do campus de Ribeirão Preto – Universidade de São Paulo pela criação e investimento na área de Física Biológica dentro do programa.

Agradeço aos professores que me apresentaram a Mecânica Estatística Não Extensiva Dr. Luís Carlos Malacarne e Dr. Ervin Kaminski Lenzi. Principalmente ao último, por seu incentivo.

Por fim agradeço ao CNPq, que foi a agência que financiou minha iniciação científica, mestrado e doutorado. Esta tese de doutorado foi financiada com a bolsa destinada ao processo 142651/2007-9.

## EPÍGRAFE

"Algumas pessoas vêem as coisas como elas são, e dizem: por quê? Eu sonho com coisas que nunca existiram, e digo: por que não?"
George Bernard Shaw

# RESUMO

DAL MOLIN, J. P. (**O papel do código estereoquímico e das flutuações térmicas locais no processo de *folding* de proteínas**). 2011. Tese (Doutorado). Faculdade de Ciências Farmacêuticas de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2011.

O problema do *folding* de proteínas tem sido investigado intensamente há mais de sessenta anos. Entretanto ainda não é encontrado na literatura um modelo que seja capaz de explicar plenamente qual é o mecanismo responsável pelo processo de *folding*. Neste contexto, a presente tese de doutorado é uma proposta minimalista para investigar o papel de um código estereoquímico, que é centrado no efeito hidrofóbico e nos vínculos estéricos dos aminoácidos (o modelo estereoquímico), no processo em pauta. Esse modelo quando combinado com um método para incluir a flutuação térmica local no sistema cadeia protéica-solvente, possibilita a investigação de um dos aspectos mais extraordinários do problema, a saber, a rapidez do processo de *folding*, considerado aqui por meio da correlação entre a complexidade da estrutura nativa (alvo) e a taxa de *folding*. Esse método é motivado por argumentos físico-químicos e biológicos, e é fundamentado na Mecânica Estatística Não Extensiva, via o uso do peso de Tsallis em simulações Monte Carlo (MC). O tempo característico de *folding* $\tau$ (obtido daquelas simulações e utilizado aqui como um parâmetro analítico do problema), cobre várias ordens de grandeza para cadeias com o mesmo tamanho. Dois conjuntos principais de simulações foram considerados com a finalidade de análise: (*i*) alguns alvos foram especialmente selecionados e submetidos a simulações MC a várias temperaturas $T$ do reservatório térmico (o meio solvente); e (*ii*) um total de duzentos alvos com topologias diversas foram submetidos a simulações à mesma temperatura $T = 1$ (unidades arbitrárias). Com essas simulações foi possível verificar comparativamente o efeito dos dois pesos estatísticos, o de Boltzmann e o de Tsallis, sobre a cinética do processo de *folding,* e assim revelar comportamentos intrigantes, porém consistentes com o fenômeno focado, como a robustez do processo de *folding* e $\tau$, este último emergindo como uma grandeza que depende da complexidade da estrutura nativa. Para estruturas distintas, $\tau$ cobre quatro ordens de grandeza. Os resultados da investigação da correlação entre $\tau$ e parâmetros globais destinados a avaliar a complexidade da topologia da estrutura nativa, como a ordem de contato e a cooperatividade estrutural, corroboram com a noção de que a rapidez do processo de *folding* é determinada fundamentalmente pela complexidade da topologia da estrutura nativa.

Palavras-chave: processo de *folding*, proteína, modelo estereoquímico, flutuação térmica local, mecânica estatística não extensiva, simulação Monte Carlo, tempo característico de *folding*, ordem de contato, cooperatividade estrutural.

## ABSTRACT

DAL MOLIN, J. P. (**The role of stereochemical code and local thermal fluctuation in the protein folding process**). 2011. Thesis (Doctoral). Faculdade de Ciências Farmacêuticas de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2011.

The protein folding problem has been investigated for more than sixty years. However is not yet found in literature a model that is able to fully explain the mechanism behind the folding process. In this context, the present doctoral thesis is a speculative and minimalist proposal to investigate the role of a stereochemical code -grounded in the hydrophobic effect and steric constraints of amino acids (the stereochemical model), in the discussed process. This model, when combined with a method to include local thermal fluctuations in the protein chain-solvent system, enables the investigation of one of the most extraordinary aspects of the problem, namely, the fastness of the folding process, which is studied here by means of the correlation between the complexity of the native structure (target) and the folding process rate. This method is motivated by physical chemistry and biological arguments, and is based on the Nonextensive Statistical Mechanics, by the use of the Tsallis weight in Monte Carlo simulations (MC), where the entropic index $q$ is adjusted at each new conformation of the protein chain. For chains with the same length, the characteristic folding time $\tau$ (estimated from those simulations and used here as an analytical parameter of the problem) span several order of magnitude. Two main sets of simulations were performed for analysis purposes. First, some targets were specially selected and submitted to MC simulations with several temperatures of the thermal reservoir (the solvent), and then a total of two hundred targets with diverse topologies were submitted to simulations with the same reservoir temperature $T = 1$ (arbitrary units). With such set of simulations, we could compare the effect of two statistical weights, namely the Boltzmann and the Tsallis weight, on the kinetics of the folding process. Intriguing but consistent behavior with respect to the folding phenomenon was reveled, such as the robustness of the process, and about the characteristic folding time $\tau$, which emerges as an amount that depends on the complexity of the native structure. For distinct structures, $\tau$ covers four orders of magnitude. Our results about the correlation between $\tau$ and global parameters to assess the complexity of the topology of the native structure, such as contact order and structural cooperativity, support the notion that the fastness of the folding process is essentially determined by the complexity of the topology of native structure.

Keywords: protein folding process, protein, stereochemical model, local thermal fluctuation, nonextensive statistical mechanics, Monte Carlo simulation, characteristic time of *folding*, contact order, structural cooperativity.

# LISTA DE FIGURAS

Figuras do Capítulo 3
Figura 1. ..... 21
Figura 2. ..... 22
Figura 3. ..... 23
Figura 4. ..... 32

Figuras do Capítulo 4
Figura 1. ..... 36
Figura 2. ..... 37
Figura 3. ..... 41
Figura 4. ..... 47
Figura 5. ..... 50
Figura 6. ..... 51
Figura 7. ..... 51

Figuras do APÊNDICE D
Figura D-1. ..... 76
Figura D-2. ..... 77
Figura D-3. ..... 78
Figura D-4. ..... 79
Figura D-5. ..... 80

# LISTA DE TABELAS

Tabela 1. ..... 36
Tabela 2 . ..... 40

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| MC | Monte Carlo (método de simulação computacional) |
| DM | Dinâmica Molecular (método de simulação computacional) |
| OC | Ordem de Contato |
| Ω | Cooperatividade Estrutural |
| CSA | Estrutura maximamente compactada ("*compact self avoiding structure*") |
| PBS | Gráfico Plaxco Baker Simons |
| NMR | Ressonância magnética nuclear ("*magnetic nuclear resonance*") |
| DNA | Ácido desoxirribonucleico ("*deoxyribonucleic acid*") |
| MCS | Passo Monte Carlo (*Monte Carlo step*) |

**SUMÁRIO**


**1 – INTRODUÇÃO** ..... **1**

**1.1 ASPECTOS GERAIS DO PROCESSO DE FOLDING DE PROTEÍNAS GLOBULARES: REVISÃO BIBLIOGRÁFICA** ..... 1

**1.2 AS CARACTERÍSTICAS COMUNS EM PROTEÍNAS GLOBULARES** ..... 3

**1.3 O CONTEXTO PARA O DESENVOLVIMENTO DOS CONCEITOS ENVOLVIDOS NO PROBLEMA** ..... 4

**1.4 SOBRE AS FORÇAS QUE INDUZEM A CADEIA À CONFORMAÇÃO NATIVA** ..... 8

**1.5 A ABORDAGEM TEÓRICA E COMPUTACIONAL DO PROBLEMA** ..... 9

**1.6 SOBRE A IMPORTÂNCIA DO PROBLEMA** ..... 13

**2 – OBJETIVOS DA TESE** ..... **15**

**3 – O MODELO ESTEREOQUÍMICO E A MECÂNICA ESTATÍSTICA NÃO EXTENSIVA** ..... 16

**3.1 O PROBLEMA DO FOLDING DE PROTEÍNAS E O MODELO ESTEREOQUÍMICO** ..... 18

**3.2 O ALFABETO ESTEREOQUÍMICO** ..... 20

**3.3 ASPECTOS TOPOLÓGICOS DA NATIVA E O DESIGN DA SEQÜÊNCIA** ..... 22

**3.4 A ENERGÉTICA DO MODELO ESTEREOQUÍMICO** ..... 25

**3.5 AS FLUTUAÇÕES TÉRMICAS LOCAIS E A ALTERAÇÃO NO CRITÉRIO DE METROPOLIS: A MECÂNICA ESTATÍSTICA NÃO EXTENSIVA** ..... 27

*3.5.1 Flutuações térmicas locais na escala nanométrica* ..... 28

*3.5.2 O fator de Boltzmann generalizado* ..... 29

*3.5.3 A motivação para o uso do peso de Tsallis* ..... 30

**3.6 ANÁLISE DAS FUNÇÕES EXPONENCIAL USUAL E Q-EXPONENCIAL** ..... 32

**3.7 A ORDEM DE CONTATO E O PARÂMETRO Ω** ..... 33

**4 – RESULTADOS E DISCUSSÕES** ..... **35**

**4.1 O TEMPO CARACTERÍSTICO DE *FOLDING* QUANDO *Q* É CONSIDERADO UM PARÂMETRO FIXO** ..... 35

**4.2 O TEMPO CARACTERÍSTICO DE *FOLDING* QUANDO *Q* É CONSIDERADO UMA VARIÁVEL DINÂMICA** ..... 39

**4.3 A CORRELAÇÃO ENTRE A "VELOCIDADE" DO PROCESSO DE *FOLDING* COM A OC E Ω** ..... 45

*4.3.1 Probabilidade de ocorrência de um contato* ..... 48

**4.4 A ROBUSTEZ DO PROCESSO DE FOLDING** ..... 52

**5 – CONCLUSÕES E PERSPECTIVAS** ..... **53**

**6 - REFERÊNCIAS** ..... **57**

**APÊNDICES** ..... **67**

**APÊNDICE A – O DESENVOLVIMENTO DO MODELO ESTEREOQUÍMICO** ..... 67

**APÊNDICE B – MECÂNICA ESTATÍSTICA NÃO EXTENSIVA** ..... 70

**APÊNDICE C – O TEMPO MÉDIO E O TEMPO CARACTERÍSTICO DE FOLDING** ..... 74

**APÊNDICE D – ALGUMAS ESTRUTURAS ALVO, SEUS MAPAS DE CONTATO E CURVAS CINÉTICAS** ..... 76

# 1 – INTRODUÇÃO

## 1.1 Aspectos gerais do processo de *folding*[1] de proteínas globulares: revisão bibliográfica

Proteínas globulares são heteropolímeros lineares formados a partir de um conjunto básico de vinte aminoácidos [1,2]. Quando se encontram em meio fisiológico adequado, essas macromoléculas assumem uma única e peculiar conformação tridimensional, denominada por estado ou estrutura nativa. Uma proteína desempenha sua atividade biológica somente quando se encontra no seu respectivo estado nativo. Tal conformação é obtida através de um processo físico-químico espontâneo, conhecido por *folding*. Esse processo é guiado por meio das interações entre os elementos atômicos que compõem a cadeia protéica (os aminoácidos), e das interações entre a cadeia e o meio solvente [3]. Apesar de ter sido, e continuar sendo, alvo de constante investigação científica, todos os detalhes desse processo ainda não são bem compreendidos [4]. Atualmente o estudo do *folding* é tratado como uma questão multidisciplinar, pois envolve conceitos e técnicas naturalmente abordados em disciplinas clássicas como a matemática, a física, a química, a biologia e a computação [5, 6].

O desafio relacionado com o problema em pauta apresenta motivações muito especiais, tanto em questões ligadas à saúde e o bem-estar humano, quanto à ciência básica. Pois, além de estar relacionado com possíveis descobertas, e, o aperfeiçoamento de medicamentos (por exemplo, no tratamento e prevenção de uma classe de doenças conhecida por amiloidoses [7]), também está associado com a compreensão de modelos e novos conceitos da física dos sistemas complexos [5], nos quais a abordagem holística é necessária. Por outro lado, os métodos tradicionais da física têm se mostrado muito úteis na construção de modelos minimalistas, tratados por simulação computacional por meio dos métodos Monte Carlo (MC) e Dinâmica Molecular (DM), originalmente aplicados em homopolímeros [8]. Modelos minimalistas têm sido propostos com a intenção de lidar apenas com aquelas características do problema que não podem ser reduzidas para aspectos mais elementares; daí a justificativa do nome “modelos minimalistas” ou “modelos reduzidos”.

---

[1] Nesta tese a palavra *folding*, e suas flexões, não foram traduzidas para o português por razões pragmáticas. As traduções usuais, a saber, dobramento e enovelamento, não representam bem o significado que o termo original encerra.

As simulações computacionais de modelos reduzidos têm originado uma significativa conexão entre teoria e técnicas experimentais, no sentido de auxiliar a compreensão dos mecanismos presentes no processo de *folding* [9]. Uma das razões se deve ao fato de que os modelos minimalistas apresentam a peculiaridade de possibilitar a geração de dados em grande quantidade dentro de intervalos de tempo razoáveis, de modo que a análise estatística possa ser amplamente empregada nas estimativas das grandezas físicas pertinentes [10, 11]. Nesta tese, em virtude dessa característica, foi possível explorar o papel da flutuação térmica local sobre a "velocidade" do *folding*, e analisar as várias ordens de grandeza encontradas para o tempo característico de *folding*, dos diversos alvos (estruturas nativas) de diferentes complexidades topológicas. Com isso, também foi possível abordar a questão da robustez do processo de *folding*, como descrito no Capítulo 4 – Resultados e Discussões.

Apesar da constante dedicação e empenho direcionados para a questão do *folding,* mais de meio século de pesquisa não foi capaz de gerar um modelo que pudesse descrever de modo adequado e completo os mecanismos subjacentes ao processo de *folding*, e assim produzir subsídios que possibilitem predizer a estrutura nativa de cada proteína usando apenas o conhecimento de sua seqüência de aminoácidos (o problema do *folding* direto). Ou, na sua versão inversa, especificada uma estrutura alvo, ser capaz de projetar uma seqüência de aminoácidos que, sob o processo espontâneo de *folding*, assuma exatamente a estrutura prescrita *a priori*. Portanto, o problema permanece como uma das principais questões em aberto da bioquímica e da biologia molecular [12, 13]. Além disso, como já mencionado, o referido problema representa um desafio para a compreensão da física dos sistemas complexos, visto que nestes sistemas a emergência de novos fenômenos, como a criticalidade auto-organizada, se encontra presente [14].

Na referência [4] Dill e colaboradores apresentam o problema do *folding* de proteínas como um conjunto de três problemas: (*i*) a definição de um código termodinâmico para o *folding*, (*ii*) a elaboração de um bom algoritmo para previsão computacional de estruturas nativas, e (*iii*) responder à questão da conjectura de Levinthal que consiste em esclarecer quais mecanismos cinéticos fazem com que o *folding* seja relativamente tão rápido. No ano de 2008, alguns dos autores da referência [4] publicaram um artigo de revisão [15], cujo núcleo é de particular importância para esta tese. Na referência [15], discorre-se sobre o gráfico Plaxco Baker Simons (PBS), na qual é apontada a existência de um dos poucos padrões universais da cinética do processo de *folding* – originalmente proposto por Plaxco, Simons e Baker em [16].

Esse padrão é a correlação entre a rapidez do processo de *folding* (representada por log($K_{eff}$)) com as propriedades topológicas globais da estrutura nativa. Um modo de avaliar tais propriedades topológicas, como o conteúdo de estruturas secundárias presentes na estrutura nativa, é através do conceito de ordem de contato (OC), que em suma, é a média da distância ao longo da cadeia entre os resíduos que se contatam [16]. Contudo, é possível avaliar tais propriedades topológicas com maior significado físico, por meio do conjunto dos contatos intra-cadeia que definem a cooperatividade estrutural (Ω), a qual será apresentada oportunamente no final do Capítulo 3 e explorada no Capítulo 4. Qual é a importância dessa correlação entre a rapidez do *folding* com os padrões estruturais da nativa? Dentre muitas outras motivações, a referência [17] fornece um indício particular sobre essa importância, pois a correlação pode ser usada como um critério para avaliar o quão razoável é a proposta de um mecanismo para o processo de *folding*. É útil acrescentar que esse tipo de correlação é relevante quando o processo estudado é classificado como de não equilíbrio em relação à termodinâmica clássica (termodinâmica de equilíbrio) [18]. E esse parece ser o caso no processo de *folding*, como é argumentado logo adiante.

## 1.2 As características comuns em proteínas globulares

Cada proteína é identificada através da sua respectiva seqüência de aminoácidos. Apesar de existir um número enorme desses biopolímeros (por exemplo, é estimado que a variedade de proteínas encontradas no corpo humano seja da ordem de centenas de milhar) elas possuem algumas características em comum. A primeira delas é que as proteínas globulares apresentam um suficientemente rápido processo de *folding,* a fim de que a função biológica para qual a proteína se destina seja cumprida. No contexto do modelo explorado neste projeto, o processo *folding* é encarado como um processo composto por duas etapas. A primeira etapa, caracterizada essencialmente como a etapa de *busca* pela estrutura nativa, é dominada pelo efeito hidrofóbico. Já na segunda etapa, a qual se inicia com a cadeia suficientemente próxima da nativa, além do *ajuste fino* pelo qual o sistema procura a conformação nativa dentro de um número mais restrito de configurações, há uma composição de fatores que se combinam construtivamente, em função da estrutura nativa e obrigatoriamente da seqüência de aminoácidos: os resíduos apolares, mais interiorizados, estão em grande parte se contatando e os resíduos polares mais expostos ao solvente; as ligações de hidrogênio do *backbone* (cadeia principal da estrutura primária de uma proteína qualquer composta por uma seqüência de carbonos alfa) estão suficientemente protegidas da

competição com o meio solvente; e, os resíduos em geral se encaixando estericamente, como num quebra-cabeça tridimensional, maximizam as interações atrativas devido às forças de dispersão.

Em textos de bioquímica é um consenso encontrar a seguinte afirmação: a estrutura tridimensional de uma proteína (o estado nativo) controla a sua funcionalidade [1, 2, 19]. A geometria do estado nativo é composta a partir de um número finito de *folds*, o qual é da ordem de um milhar [20] - *folds,* ou *motifs*, são formas modulares constituídas de partes simples; por exemplo, certos *folds* são compostos por combinações de hélices-α (e/ou folhas-β quase planares) com elas mesmas–[21]. A estrutura nativa é geralmente robusta quanto à constituição da proteína, ou seja, é possível a troca de alguns aminoácidos ao longo da cadeia sem que isso afete de modo radical a topologia do estado nativo [9, 22], embora, dependendo da troca realizada, o funcionamento biológico correto pode ser severamente comprometido [23]. Outro aspecto muito importante presente nas proteínas globulares é a sua flexibilidade, pois permite que realizem diversas tarefas [24, 25]. Para a realização dessas tarefas, as proteínas também podem interagir entre si [1,2].

Aquelas características comuns das proteínas permitem especular que, a quantia limitada de *folds* resulta de propriedades que devem ser compartilhadas por todas elas, e que estas formas modulares são originadas do processo de seleção natural. Contudo, identificar os atributos ordinários que determinam a variedade de *folds* [21] representa um grande desafio a ser enfrentado. É possível imaginar que um desses atributos está relacionado com a acessibilidade cinética daqueles padrões estruturais, e que, a composição geométrica desses padrões resulta na especificidade do papel biológico que cada estrutura nativa desempenha.

### 1.3 O contexto para o desenvolvimento dos conceitos envolvidos no problema

Após a introdução da hipótese termodinâmica [3, 12, 18, 26], estudos sistemáticos para se identificar as forças determinantes na indução da cadeia protéica à conformação nativa, e, de sua estabilidade estrutural, tiveram grande impulso, conforme apresentado em um item da referência [15]. Porém, interpretações conflitantes do papel de cada uma das diversas forças envolvidas no processo têm sido comuns. Fundamentalmente, a hipótese termodinâmica estabelece que a conformação nativa deva corresponder ao mínimo global da energia livre do sistema cadeia-solvente, porém esta afirmação nunca foi conclusivamente

demonstrada, isto é, não há prova de que o estado nativo da proteína seja exatamente aquela conformação termodinamicamente mais estável (*controle termodinâmico*). Por outro lado, embora também não haja evidências rigorosas de que o estado nativo seja a conformação de mais fácil acesso (*controle cinético*) [27], no final da década de 1990 alguns experimentos sugeriram que os estados selecionados cineticamente desempenham papel funcional em algumas proteínas, ao mesmo tempo em que também foi identificada a possibilidade da existência de estados energeticamente mais estáveis do que o estado nativo [28]. Portanto, a supremacia do controle termodinâmico na reação do *folding* pode ser contestada. Um trabalho publicado em meados da década de 1990 [29] sugere a necessidade de separar a idéia de que a conformação nativa é determinada apenas pela seqüência de aminoácidos, do conceito de estabilidade termodinâmica. No parágrafo final do artigo [29] os autores, Baker e Agard indicam que os resultados apresentados no texto podem ser usados como uma evidência sobre a possibilidade de que a seleção natural pode ter assegurado não apenas a estabilidade dos estados nativos, mas também sua acessibilidade. Dessa forma é sugerido que as seqüências de aminoácidos sofreram pressões evolutivas para codificar as rotas para o respectivo estado nativo de cada proteína. A conclusão apresentada no artigo [29] defende a idéia que, se o controle cinético é geral, então o mínimo de energia global pode ser simplesmente irrelevante para a biologia. Cabe aqui um comentário adicional: a hipótese termodinâmica é baseada em argumentos da termodinâmica clássica. Contudo, o sistema cadeia protéica-solvente é um sistema heterogêneo, com componentes nanoscópicos. Isso significa que o processo de *folding,* ou seja, a formação espontânea de uma estrutura tridimensional ordenada pode não estar de acordo com os princípios da termodinâmica clássica. Um argumento plausível para defender esse ponto de vista é que os efeitos de superfície presentes – flutuações térmicas locais – no processo não podem ser ignorados [18]. Comentários e argumentos adicionais se encontram no Capítulo 3.

As pesquisas dedicadas ao problema do *folding* ao longo de algumas décadas têm produzido dados experimentais em abundância, como as descrições pormenorizadas de estruturas espaciais de milhares de proteínas globulares, determinadas basicamente por cristalografia de raios-X e espectroscopia por ressonância magnética nuclear (NMR) [30, 31], e principalmente dados termodinâmicos experimentais de proteínas em solução, como entalpia, entropia e diferença da energia livre dos estados *folded* e *unfolded* [32, 33]. Recentemente, a quantidade de informações experimentais referentes à cinética de *folding* tem aumentado, particularmente no que diz respeito à correlação entre taxa de *folding* e a *ordem*

*de contato* (OC) da estrutura nativa [15–17]. Entretanto desde meados da década de 1990, e, em maior número nos últimos anos uma nova classe de experimentos tem sido realizada, são os chamados experimentos de uma única molécula [34]. O grande mérito desses experimentos é que eles possuem a habilidade de isolar as flutuações presentes nas grandezas físicas de interesse quando moléculas individuais são analisadas, daquele comportamento médio observado no *ensamble* em experimentos tradicionais de bioquímica (como aqueles mencionados no início deste parágrafo). Uma proposta para levar em conta essas flutuações em um modelo minimalista destinado a investigar o processo de *folding* é discutida oportunamente no Capítulo 3.

Embora a termodinâmica clássica seja uma das mais poderosas e aplicadas disciplinas no estudo de sistemas biológicos, suas descrições não levam em conta os detalhes dos mecanismos subjacentes envolvidos nos processos analisados. Essa pode ser uma das razões de interpretações distintas (e também conflitantes) dos dados resultantes das medidas experimentais. Além disso, o uso da termodinâmica de equilíbrio no estudo do processo de *folding* tem sido questionado, como pode ser encontrado na referência [18]. Nesse artigo se encontram trechos que chamam atenção para o fato de que, se um sistema do tipo cadeia protéica-solvente obedecer à distribuição de probabilidades de Boltzmann, e, o princípio de ordem de Boltzmann, então a possibilidade da formação espontânea da estrutura protéica tri-dimensional é praticamente nula. Além disso, é impossível (do ponto de vista estatístico) visitar todos os estados conformacionais equivalentes. Aqui também cabe um comentário adicional, pois a referência [18] indica que a formação do estado nativo no sistema considerado se dá por um processo não ergódico; e que o efeito de superfície – no nosso ponto de vista a flutuação térmica local resultante da interação entre as partes que compõe o sistema – não pode ser ignorado. Considerações, implicações e uma alternativa viável para os comentários anteriores são apresentados e explorados, nos Capítulos 3, 4 e 5 desta tese.

Para muitas proteínas, a transição entre o estado desnaturado e nativo ocorre praticamente sem a presença de intermediários detectáveis; tecnicamente, se trata de uma transição *tudo ou nada*, ou seja, ou a proteína se encontra desnaturada ou está no estado nativo [35]. Para esses casos a termodinâmica é simples, e então é possível caracterizar a quantidade de moléculas em cada estado ($U$ e $N$) através da diferença da energia livre $\Delta G$ molar entre os dois estados: $\Delta G = -RT\ln(K_{eq})$, com $K_{ef} = [N]/[U]$, onde $R$ é a constante

universal dos gases, $T$ é a temperatura, $K_{eq}$ é a constante de equilíbrio e $[U]$ é a concentração de moléculas no estado desnaturado. Para uma proteína típica, $\Delta G$ é conhecida como sendo da ordem de –10 Kcal/mol, então à temperatura ambiente ($RT_0$ aproximadamente 0,6 Kcal/mol), a estimativa é que para cada proteína no estado $U$ deva existir aproximadamente $10^7$ moléculas no estado $N$. Essa relação $[N]/[U]$ é supra eficaz para que as funções biológicas sejam cumpridas, porém a observação que se destaca advém da condição de estabilidade energética marginal da proteína, pois a diferença de $\Delta G$ em torno de –10 Kcal/mol é equivalente à energia de algumas poucas ligações de hidrogênio [36].

Desde os trabalhos pioneiros de Pauling e Corey com modelos de estruturas secundárias [37], as ligações de hidrogênio têm sido propostas como as interações determinantes para a estabilidade das proteínas. Um artigo de revisão publicado recentemente (2008) apresenta como argumento central a conexão entre a termodinâmica do processo de *folding* com as pontes de hidrogênio presentes no *backbone* da cadeia protéica [38]. Contudo, o assunto permanece em pauta como um constante tópico de investigação científica, cujos resultados têm produzido interpretações díspares quanto a sua relevância para a estabilidade das proteínas. De fato, já em 1959 Kauzmann argumentou que a estabilização do estado nativo de uma proteína seria fundamentalmente regida pelo efeito hidrofóbico [39]. Então seguiram estudos com pequenas moléculas em solução [40], os quais sugerem que a contribuição energética das ligações de hidrogênio é pequena, ou mesmo dispensável como ocorre com alguns *foldamers*, que são moléculas artificiais que imitam proteínas [41]. Esses *foldamers* adotam uma única conformação (na forma de hélice) sem a existência de ligações de hidrogênio [42]. Talvez a tentativa de conciliar o papel das pontes de hidrogênio na estabilização do estado nativo seja motivada pela intenção de preservar a hipótese termodinâmica [38], a qual não leva flutuações térmicas locais em conta.

O conjunto das referências [39-42] sugere que as ligações de hidrogênio não contribuem significativamente para a estabilidade energética da proteína, quando não atuam no sentido oposto, enquanto que experimentos mais recentes, envolvendo múltiplas mutações de resíduos simples (um de cada vez), sugerem retornar a idéia de que as ligações de hidrogênio estabilizariam as proteínas por aproximadamente entre -1 e -3 Kcal/mol para cada ligação de hidrogênio intra-cadeia [43-46]. Entretanto existem defensores, suficientes em número e prestígio, para que a questão de como uma proteína é estabilizada

energeticamente permaneça em aberto, e por enquanto mantendo vivas diferentes interpretações dos fatos conhecidos sobre o problema do *folding* de proteínas [38, 47].

### 1.4 Sobre as forças que induzem a cadeia à conformação nativa

A característica mais pertinente das proteínas globulares é sua habilidade em adotar uma conformação estrutural específica, que é única para cada representante dessa classe de biopolímeros, porém quando em solução não se mantém estática numa única configuração, como ocorre na sua forma cristalizada. Essa característica peculiar tem ocupado intensamente diversos ramos científicos, pois algumas dúvidas persistem mesmo em aspectos gerais, como já mencionados, ou seja, se a questão do controle do processo de *folding* é termodinâmica ou cinética, à semelhança do que ocorre com alguns produtos químicos orgânicos. Em certas condições, o controle cinético da reação pode ser dominante para o produto (com menor energia de ativação, por exemplo), ainda que o mesmo possua menor estabilidade energética [48].

A identidade de uma proteína em particular é providenciada pelas especificidades químicas (resíduos polares, apolares e neutros) e estéricas (diversos tamanhos e formas) de cada aminoácido, ao longo da cadeia protéica: cada seqüência de aminoácidos é caracterizada por um código estereoquímico. Esse conjunto de propriedades dos aminoácidos tem sido interpretado como as instruções que guiam a cadeia polipeptídica à estrutura nativa [3]. Embora ainda não se conheça como este código estereoquímico seja traduzido dinamicamente por meio das diversas forças efetivas envolvidas[1] [49], as interações entre as cadeias laterais (resíduos) da proteína têm sido tradicionalmente aceitas como determinantes do processo que induz a cadeia polipeptídica para a sua conformação nativa [3, 15].

Apesar da aparente coerência das idéias apresentadas, têm sido propostas abordagens radicalmente adversas, pois mesmo a importância das especificidades das cadeias laterais tem sido questionada. Mesmo com a equivalência química das unidades poliméricas da cadeia principal (o *backbone*), portanto não possuindo o poder discriminatório das cadeias laterais,

[1]Forças entrópicas (efeito hidrofóbico), ligações de hidrogênio e solvatação aquosa da proteína têm sido consideradas como forças fundamentais para o entendimento do processo que leva a cadeia para a estrutura nativa específica.

foi proposta em 2008 uma teoria de *folding* baseada nos doadores e receptores de ligações de hidrogênio intra-cadeia principal, por mais paradoxal que essa abordagem possa parecer [38].

## 1.5 A abordagem teórica e computacional do problema

Na expressão "processo de *folding*" se encontra implícito dois aspectos do problema: (*i*) a sucessão de etapas através do espaço das configurações que levam à estrutura nativa, e (*ii*) a questão da estabilidade do glóbulo protéico. Como já enfatizado em trechos anteriores, os grandes desafios envolvidos em interpretar o processo de *folding* envolvem igualmente esses dois aspectos fundamentais, e parte significativa das dificuldades tem sua origem na complexidade das interações intra-cadeia e cadeia-solvente, limitando a possibilidade de uma descrição precisa do papel de cada componente energético envolvido no sistema proteína-solvente quando tratadas em separado. Por isso, reside tanto na análise teórica, com novas idéias, paradigmas e propostas de novos experimentos, quanto na abordagem computacional, por meio de simulações numéricas de modelos simplificados, a fonte de inspiração e de novas ferramentas para o estudo do problema. Com efeito, espera-se que através de métodos alternativos e novas técnicas experimentais seja possível alimentar a rotina da análise, isto é, isolar e estudar cada componente da questão independentemente, e depois integrá-los apropriadamente com a perspectiva de adquirir a compreensão do(s) mecanismo(s) por trás do processo [9–11].

A investigação teórica e os experimentos computacionais, portanto, são empregados no desenvolvimento dos testes de novas idéias e análises que deverão inspirar e propor novos e estratégicos experimentos. A conjectura de Levinthal, por exemplo, implica na necessidade da existência de caminhos preferenciais para que o *folding* de proteínas se processe em tempo hábil[2] [50-53], motivou a proposição de diferentes abordagens para estudar este processo. Um conceito emergente que procura "resolver" a conjectura de Levinthal pode ser resumido pela metáfora do funil [9, 15]: em condições favoráveis ao *folding* não haveria caminhos específicos para a conformação nativa, mas sim um afunilamento dos possíveis caminhos conforme o sistema evolui para o estado nativo. Essa imagem representa de forma simplificada a superfície multidimensional de energia livre determinada pelo campo de força interacional entre todas as partes constituintes do sistema (portanto, específica a cada

[2] Esta conjetura ficou conhecida como "paradoxo de Levinthal", pois caminhos e estados intermediários do processo de *folding* ainda não foram efetivamente mapeados em experimentos *in vitro*.

proteína). Na fase desnaturada a cadeia teria um número muito grande de configurações acessíveis ($>10^{N/3}$, onde $N$ é o tamanho da cadeia protéica), correspondendo à boca do funil; mas, eventualmente com partes da cadeia já na conformação próxima da correta, o número de configurações possíveis se reduziria drasticamente, isto é, quanto mais elementos da cadeia se encontram na conformação correta, menor se tornaria o espaço amostral disponível. A superfície de energia livre na forma de funil representa o *folding* de proteínas como um processo que, sob condições apropriadas, a cadeia protéica começa sua evolução configuracional a partir de um estado com grande entropia (estrutura desorganizada) e evolui continuamente para um estado de mais baixa entropia da cadeia, mais organizado e com os contatos corretos encontrados na estrutura nativa. Em princípio a superfície de energia livre poderia ser muito rugosa, mas é assumido que o *processo evolutivo* teria selecionado aquelas seqüências que resultassem em superfícies suficientemente lisas, sem armadilhas energéticas, topológicas ou cinéticas, garantindo assim um rápido processo de *folding* [9, 15].

Outras abordagens para o problema considerado procuram detalhar os mecanismos envolvidos no processo de *folding* em diferentes níveis; seguem três exemplos: (*i*)- o *modelo framework* [54] que basicamente postula uma hierarquia temporal de eventos; (*ii*)- o *modelo do colapso hidrofóbico* [39] que propõe o efeito hidrofóbico como o iniciador da reação de *folding*, colapsando a proteína na forma de um glóbulo precursor; e (*iii*)- o modelo da nucleação-condensação [55], o qual postula a formação inicial de um núcleo crítico de alguns resíduos adjacentes, com os contatos nativos corretos; este núcleo teria a propriedade de dirigir o restante do processo de *folding*. Comentários sobre essas três propostas também podem ser encontrados no Capítulo onze do livro de Nölting [56]. Outra referência recente (2008) sobre aquelas propostas para descrever os mecanismos de *folding*, bem como sobre múltiplas rotas e heterogeneidade estrutural no *folding* de proteínas, pode ser encontrada em [57]. Quanto ao modelo utilizado nesta tese de doutorado, sua apresentação se encontra no Capítulo 3.

Conforme indicado nos parágrafos anteriores, apesar dos esforços dedicados ao problema do *folding* de proteínas globulares, muitas questões ainda permanecem latentes. Este trabalho é direcionado para a interpretação das forças que controlam o(s) mecanismo(s) responsável por induzir a cadeia protéica para sua estrutura nativa, e para isso aspectos termodinâmicos e cinéticos do problema são considerados. Particularmente, temas como o da correlação entre taxa de *folding* e os atributos topológicos da estrutura nativa, e questões como: "qual o limite de validade em que a complexa interação cadeia-solvente pode ser representada por potenciais efetivos entre pares de

resíduos da cadeia?" (como proposto originalmente por Kauzmann via o conceito de "ligação hidrofóbica" [39]), têm sido o centro de nossas atenções. Essas questões estão também relacionadas com a perspectiva dual do problema: mecanismo(s) de busca e estabilidade do glóbulo protéico. Daí, se a especificidade da seqüência de resíduos é suficiente para que a estrutura nativa seja definida, então é compreensível que este código possa ser "imitado" por algum esquema de condições interacionais intra-cadeia, representado, em geral, por algum tipo de potencial de pares. Porém uma das questões em aberto diz respeito à universalidade desse potencial efetivo, no sentido de funcionar para todas as proteínas [58, 59]. Adicionalmente, aqui também é analisada a viabilidade do uso do peso estatístico de Tsallis (Mecânica Estatística Não Extensiva) para simular um modelo minimalista destinado a emular aspectos básicos do processo de *folding* em regime fora do equilíbrio termodinâmico. Essa abordagem é justificada pelas flutuações térmicas locais presentes no sistema conforme é discutido no Capítulo 3.

Como ainda não se dispõe de um modelo, ou conjunto de princípios, que satisfaça a necessidade de razoável compreensão sobre o problema do *folding*, esta tese é fundamentada em dados experimentais que possam discriminar as condições de "estabilidade da proteína" daqueles elementos determinantes dos "mecanismos de busca" (sucessão de etapas através do espaço das configurações que levam à sua estrutura nativa). Por exemplo, certas interações energéticas estão presentes em algumas proteínas, mas não em outras, como ocorre com as pontes cistinas e pontes salinas. Conseqüentemente, sob tal perspectiva, essas interações não devem ser fundamentais para o(s) mecanismo(s) de busca, embora sejam fatores importantes na questão da estabilidade. Entretanto, outras interações são ubíquas, como as forças entrópicas (efeito hidrofóbico) e as especificidades estéricas dos aminoácidos, presentes em todas as proteínas e ambas com potencial discriminatório nas interações intra-cadeia e cadeia-solvente [60]. Também é incluído um terceiro ingrediente, a saber, o efeito da flutuação térmica local num sistema finito.

Dado o panorama apresentado no parágrafo anterior, o problema de *folding* de proteínas globulares nesta tese é considerado sob seu aspecto geral, daí as hipóteses aqui adotadas estão condensadas no seguinte conjunto de premissas conseqüentes:

1. O mecanismo básico, universal para a reação de *folding* $\{U_i\} \leftrightarrow \{N_0\}$: a conformação estrutural funcional ($N_0$) de uma proteína natural é específica e única [3].

2. O potencial estereoquímico: “especificidade estérica” (formas e tamanhos) e “perfil polar-apolar” (hidrofobicidade) dos aminoácidos são ingredientes fundamentais dos mecanismos de busca [61].

3. As estruturas secundárias (hélices-α, folhas-β, *turns*, etc) são os padrões configuracionais mais prováveis [62, 63], pois permitem a fácil acessibilidade cinética da estrutura nativa.

4. Os mecanismos de busca e a consecução das condições de estabilidade global, são etapas desacopladas que, respectivamente, se sucedem no tempo [64].

5. A estabilidade da proteína ocorre de forma significativa somente quando sua correspondente estrutura nativa é alcançada, na qual as ligações de hidrogênio intra-cadeia principal estão eficientemente protegidas da competição com o meio solvente, e o efeito atrativo das interações de van der Waals é maximizado.

6. Flutuações térmicas na escala nanoscópica (o tamanho típico das proteínas globulares) são fatores determinantes da cinética do processo de *folding*.

Seguem alguns comentários sobre as seis premissas anteriores:

A – O processo de *folding* pode ser iniciado a partir de uma configuração qualquer da cadeia polipeptídica, contudo – desde que as condições do meio (condições fisiológicas) permaneçam as mesmas ao longo do processo – o resultado final dessa reação é sempre o mesmo: a obtenção do estado nativo.

B – Adota-se que os ingredientes fundamentais para os mecanismos de busca são: (*i*) o efeito hidrofóbico (escala de hidrofobicidade), e, (*ii*) os vínculos estéricos (forma e tamanho) dos aminoácidos; pois estão presentes em todas as proteínas, e isto é refletido na energia livre de transferência (ou seja, o custo energético para transferir um aminoácido do meio polar para o meio apolar [65]).

C – As hélices-α, folhas-β, e *turns* são os padrões configuracionais mais prováveis porque são de fácil acesso cinético. Isso é justificado pela facilidade com que os contatos nativos de uma cadeia são formados por conta da proximidade dos resíduos ao longo da estrutura primária.

D – Se por um lado o efeito hidrofóbico é o grande responsável pelo colapso da cadeia, por outro lado as restrições estéricas influenciam consideravelmente na busca pela nativa, reduzindo o tempo gasto na primeira etapa do processo (etapa de busca), bem como facilitando a volta para a nativa na segunda etapa (etapa de estabilidade). Com relação à primeira etapa, isso é observado nos resultados das simulações para um intervalo significativo de temperaturas $T$ do reservatório térmico como é mostrado no Capítulo 4.

E – A estabilidade de uma proteína é obtida quando uma grande parte dos contatos nativos é formada, pois é somente na conformação nativa que se atinge a condição em que as ligações de hidrogênio intra-cadeia estão protegidas da competição com o meio solvente, e, que o efeito atrativo das forças de dispersão (van der Waals) é maximizado pela complementaridade estérica das cadeias laterais. Em ambas as etapas do processo de *folding*, a estrutura nativa funciona como um atrator.

F – O papel das flutuações térmicas locais ganha relevância neste problema, pois uma única molécula protéica em contato com o meio solvente representa uma componente finita (nanoscópica) de um sistema heterogêneo; uma proteína imersa em um reservatório térmico.

Um dos grandes desafios relacionados com o problema do *folding* de proteínas é explicar as taxas *folding* para diversas proteínas globulares. Embora muito rápidas se comparadas com o comportamento de outras moléculas pequenas, diferentes proteínas podem apresentar diferenças de várias ordens de grandeza na taxa de *folding*, dentro de uma mesma faixa estreita de temperaturas. Como uma proposta para investigar essas taxas, o papel das flutuações térmicas locais é explorado no contexto da Mecânica Estatística Não Extensiva, aplicada ao modelo minimalista utilizado aqui. Isso se encontra no Capítulo 3 e analisado no Capítulo 4 desta tese.

### 1.6 Sobre a importância do problema

A compreensão do processo de *folding* de proteínas tem sido apontada, e, adotada como o caminho precursor para o estabelecimento da relação direta entre a seqüência de aminoácidos (identidade) da proteína e sua função. Como cada gene (e/ou cada uma de suas eventuais e específicas combinações de genes) determina uma proteína, a função de cada uma delas poderia, em princípio, ser lida diretamente a partir do DNA.

Paralelo ao aspecto de importância científico-acadêmico há também o interesse socioeconômico pelo problema, devido à crescente lista de doenças tendo como causa comum falhas no *folding* de proteínas (*misfolding*). Nessa lista se inclui uma classe de doenças denominada "amiloidoses", que é caracterizada pelo acúmulo de proteínas no espaço extracelular como o agregado protéico, na forma de fibrilas nanométricas. Especificamente, são enfermidades recorrentes dadas suas importâncias atuais: o mal de Alzheimer; o carcinoma da tireóide; as encefalopatias espongiformes; o mal de Parkinson; o diabetes tipo-II; entre outras [5,7] e [23,66].

Apresentados os motivos para estudar o problema do *folding*, a conquista da compreensão dos elementos determinantes deste processo se tornou uma tarefa científica de grande importância. Uma parte da solução do problema seria desenvolver a capacidade de prever a estrutura nativa das proteínas de forma automatizada, por qualquer método que seja. Embora *per si* essa conquista represente um grande desafio, no qual se tem investido muito trabalho e recursos[3], esta capacitação não satisfaz questões importantes que dependem de um entendimento pormenorizado do processo de *folding*. Por exemplo: "por que, e como, as proteínas solúveis perdem sua habilidade para o *folding* correto?"; "como prescrever uma seqüência de aminoácidos para obter uma proteína globular numa desejada conformação estrutural?". A habilidade de responder questões como essas apresenta potencial de impacto revolucionário no tratamento e prevenção de doenças [67].

Nos itens seguintes da tese se encontram: 2 – Objetivos da tese; 3 – O modelo estereoquímico e a Mecânica Estatística Não Extensiva, onde o modelo estereoquímico é apresentado juntamente com as motivações para as variações da técnica de simulação computacional empregada; a apresentação da OC e do parâmetro Ω encerra este Capítulo; 4 – Resultados e Discussões, neste item se encontram: (*i*) que o índice entrópico $q$ é estrutura/seqüência dependente; (*ii*) é sugerido que, talvez o índice entrópico $q$ seja uma variável dinâmica do sistema cadeia protéica-solvente; (*iii*) adicionalmente são exploradas as correlações entre o tempo característico de *folding* e dois parâmetros dedicados a descrever a topologia da nativa (OC e Ω); o Capítulo é encerrado com a discussão sobre a robustez do processo de *folding* no contexto do modelo utilizado. Os últimos itens são: 5 – Conclusões, 6 – Referências e Apêndices A, B, C e D.

---

[3]Um exemplo representativo do poder de motivação do problema do *folding* é a construção de um supercomputador http://domino.research.ibm.com/comm/research_projects.nsf/pages/bluegene.index.html com aproximadamente um milhão processadores (dual) e custos na ordem de centenas de milhões de dólares.

## 2 – OBJETIVOS DA TESE

A meta geral desta tese de doutorado é qualificar o papel das forças entrópicas (efeito hidrofóbico) associadas com as especificidades estéricas no processo de *folding* de proteínas globulares (o código estereoquímico), juntamente com o papel desempenhado pelas flutuações térmicas locais inerentes ao sistema cadeia protéica-solvente.

Objetivos específicos:

1. Evidenciar a prevalência do *controle cinético* sobre o *controle termodinâmico* no processo de busca pela nativa, como decorrente dos atributos topológicos da estrutura nativa estimados por parâmetros globais como, por exemplo, a ordem de contato OC e a cooperatividade estrutural $\Omega$.

2. Mostrar como a taxa de *folding* pode ser correlacionada com diversos atributos topológicos das estruturas alvo (*folds* e suas composições cooperativas), e propor um argumento plausível para as várias ordens de grandeza observadas nos tempos característicos de *folding* $\tau$.

3. Usar o peso de Tsallis nas ponderações configuracionais das simulações Monte Carlo, ao invés do peso de Boltzmann. Esse uso é justificado pelas flutuações térmicas locais no sistema cadeia-solvente (sistema heterogêneo com componente nanométrica e fora do equilíbrio termodinâmico), moduladas pelo índice entrópico $q$.

# 3 – O MODELO ESTEREOQUÍMICO E A MECÂNICA ESTATÍSTICA NÃO EXTENSIVA

Com a intenção de contextualizar os argumentos utilizados na descrição do modelo discutido neste Capítulo, são expostos e comentados alguns conceitos e fatos experimentais relacionados com o desenvolvimento, e, especulações teóricas sobre o estudo do problema de *folding* de proteínas. Conforme já mencionado no Capítulo introdutório, uma proteína específica é identificada por sua estrutura primária, isto é, pela ordem em que os diferentes aminoácidos são dispostos ao longo da cadeia polipeptídica. Devemos o reconhecimento deste fato ao bioquímico Frederick Sanger, que efetivamente estabeleceu, no início da década de 1950, a idéia da estrutura primária em proteínas; veja em Pelesko [14]. Durante a década de 1960, os trabalhos conduzidos por Anfinsen e colaboradores, difundiram a idéia de que toda informação necessária para dirigir o processo de *folding* de uma particular proteína se encontra codificada ao longo de sua estrutura primária, e no modo como ela interage com o meio, isto é, pelas condições fisiológicas. Embora seja reconhecido que algumas proteínas necessitem de coadjuvantes do meio celular (chaperones) para se enovelarem, os chaperones também são proteínas. Além disso, não há chaperones disponíveis para todas as outras proteínas; parte delas precisam se enovelar espontaneamente, ou seja, por conta própria. Daí, o grande impacto da proposição de Anfinsen foi que o processo de *folding* pode ser estudado sob o prisma da física e da química somente.

A idéia freqüentemente identificada com os trabalhos de Anfinsen, mas formulada por Lumry e Eyring [68], que muito corroborou com a proliferação de experimentos físico-químicos relativos ao problema, é a hipótese termodinâmica, que associa a conformação nativa ao estado de menor energia livre global do sistema cadeia-solvente [3,18]. Assim, a estrutura nativa seria a conformação termodinamicamente mais favorável, e por conseqüência seria alcançada mediante uma busca ao acaso no espaço das configurações da cadeia. Porém, como também já citado no primeiro capítulo, em meados da década de 1990 foram identificadas proteínas nas quais seus estados funcionais certamente não correspondem ao mínimo de energia livre global, mas sim às conformações cujo acesso cinético é privilegiado [29]. Essa evidência favorece outro argumento sobre o mecanismo subjacente ao processo de *folding*. Esse argumento, primeiramente discutido na década de 1970 por Levinthal, conhecido como a proposta cinética, sugere que as proteínas deveriam seguir caminhos bem definidos ao longo do espaço configuracional, ligando dinamicamente os estados

desnaturados à estrutura nativa. Contudo, o mapeamento das rotas específicas é aguardado até agora. Durante as décadas de 1980 e 1990 Wolynes, Bryngelson, Onuchic e colaboradores apresentaram uma proposta alternativa, conhecida como a metáfora do funil de energia livre. De acordo com essa idéia, para proteínas reais a superfície de energia livre seria lisa o suficiente para permitir que a cadeia protéica escapasse facilmente dos mínimos locais e, à medida que o processo evoluísse, o espaço configuracional seria reduzido, até que a estrutura nativa fosse encontrada na região mais funda do funil [9]. Nessa abordagem o responsável pelas propriedades do funil seria a evolução biológica. Dentro desse contexto é sugerido que uma única rota específica para o *folding* não seria mais necessária, pois diferentes rotas na superfície afunilada da energia livre conduziriam a cadeia à estrutura nativa.

As propriedades gerais do processo de *folding* de proteínas globulares que se procura explicar envolvem, majoritariamente: (*i*) o(s) mecanismo(s) preciso(s) de busca (as estruturas nativas são únicas e geometricamente precisas); (*ii*) a estabilidade marginal da conformação nativa (para uma proteína típica, sua estabilidade é da ordem de apenas 10kcal/mol); (*iii*) a robustez do processo (por exemplo, o processo de busca se dá, para distintos organismos, em temperaturas muito distintas, varrendo um intervalo que vai de aproximadamente zero para mais de 100 $^{o}$C); e (*iv*) a velocidade diferenciada do processo (embora o processo de *folding* de proteínas globulares seja em geral muito rápido, são encontradas taxas de *folding* que diferem por várias ordens de grandeza, mesmo para proteínas de tamanhos similares). Em geral, as idéias e os mecanismos propostos são bem sucedidos para explicar apenas parte dessas propriedades, mas falham em outras [17]. Com o modelo empregado neste trabalho, embora minimalista, pretende-se reproduzir boa parte das propriedades gerais do processo, como descrito acima, no intuito de gerar *insights* para a compreensão do fenômeno em pauta.

Por conta da abordagem especulativa e teórico-computacional escolhida para estudar o problema do *folding* de proteínas, o procedimento analítico-dedutivo foi o escolhido. O modelo do sistema cadeia protéica-solvente, que será identificado a partir daqui por *modelo estereoquímico*, é tratado por meio de algoritmos Monte Carlo (MC), com o critério de Metropolis modificado. Para completar o círculo metodológico, as premissas incorporadas nesse modelo minimalista são analisadas através dos resultados obtidos daquelas simulações MC. A seguir, é feita uma descrição sistemática sobre como o modelo estereoquímico tem sido construído, em seguida é exposta a motivação física para alterar o critério de Metropolis

no método MC, que é centrada na Mecânica Estatística Não Extensiva. Como último tópico do Capítulo os parâmetros globais OC e Ω são apresentados.

## 3.1 O problema do *folding* de proteínas e o modelo estereoquímico

O processo de *folding* pode ser abordado como um problema no qual a cadeia protéica e o solvente são tratados como um todo. Nesse caso temos um sistema heterogêneo (proteína-solvente), no qual um dos componentes possui um número finito de graus de liberdade (uma única cadeia de tamanho limitado), e cujas propriedades relevantes resultam, pelo menos no seu estágio inicial, da busca pela estrutura nativa, de um processo essencialmente fora do equilíbrio termodinâmico. Portanto, pode ser esperado que propriedades não usuais venham a emergir, principalmente quando as interações intra-cadeia e cadeia-solvente são levadas em conta.

Um exemplo daquela propriedade emergente é a evolução temporal não trivial que ocorre durante a transição da cadeia polipeptídica do estado desnaturado para o estado nativo. Para cada simulação MC independente realizada, o tempo de *folding* é representado pelo tempo de primeira passagem, ou seja, o intervalo de tempo registrado em passos Monte Carlo para que a cadeia atinja a conformação nativa (por isso a associação com o termo: "tempo de primeira passagem"). Por meio de um conjunto $\{t_i^S\}_N$ de tempos de primeira passagem (obtidos de forma independente para cada estrutura alvo $S$, e com $N >> 1$) é estimado o tempo característico de *folding* $\tau$ para aquela estrutura $S$ (nativa). Corroborando com dados experimentais, é verificado para o presente modelo que $\tau$ se correlaciona com características topológicas globais da estrutura e certos padrões estruturais específicos da nativa, como analisado com detalhes no Capítulo seguinte.

O modelo estereoquímico é devotado exclusivamente para a etapa de busca da estrutura nativa, como será detalhado adiante. O modelo é baseado em fatos e em parâmetros experimentais como a energia livre de transferência, resultante do efeito hidrofóbico, e as distintas especificidades estéricas dos aminoácidos. De suas propriedades, um reduzido mundo paralelo ao das proteínas globulares emerge o qual permite investigar aspectos cinéticos e termodinâmicos relacionados com o processo sob estudo. Esse modelo tem sido desenvolvido para abordar o problema do *folding* inverso, ou seja, é provida uma regra geral

para o *design* da seqüência de "aminoácidos" para cada estrutura alvo (estrutura nativa) escolhida *a priori*.

A construção do modelo tem sido baseada na escolha dos elementos mínimos necessários para gerar algo que se pareça suficientemente com o sistema que se pretende estudar, em vários de seus aspectos. Esses elementos são: (*i*) espaço configuracional discreto (rede cúbica) – adotado por razões de desempenho computacional; (*ii*) repertório mínimo de "aminoácidos" distintos (alfabeto estereoquímico) que irão compor a cadeia protéica, caracterizados e diferenciados pelos seus respectivos níveis de hidrofobicidade e especificidades estéricas, como as interações ubíquas no universo das proteínas; (*iii*) cadeias lineares de tamanho fixo (27 resíduos); (*iv*) estruturas nativas representadas por configurações maximamente compactas (CSA 3×3×3).

Uma vez que proteínas – bem como o conjunto de basicamente 20 aminoácidos dos quais são formadas – são consideradas como produtos da evolução biológica ao longo de bilhões de anos, é razoável perguntar: como seria possível criar um cenário que imitasse tal resultado? Com o do conceito de evolução em cena, é importante considerar o que é preservado, e dentre todos os aspectos característicos das proteínas naturais, dois deles se destacam, pois são comuns a todas elas: (*i*) as especificidades estereoquímicas dos aminoácidos - e as conseqüentes modalidades de interações entre si e com o meio solvente; e (*ii*) os padrões topológicos das estruturas nativas, os quais aparentam ser relativamente reduzidos (da ordem de um milhar) .

Sabe-se que os aminoácidos são distinguidos por seus distintos tamanhos e formas, assim como por suas propriedades interacionais específicas em relação ao solvente (a água). Proteínas globulares, em suas respectivas estruturas nativas, formam arranjos atômicos extremamente compactados [69]; portanto, a questão da forma e tamanho dos aminoácidos leva a restrições estéricas determinando quais pares (ou mesmo grupos maiores) deles são mais propensos a se acomodarem espacialmente, e desta forma na estrutura nativa não é qualquer tipo de aminoácido que pode ficar espacialmente próximo (em contato) um do outro. Por outro lado, o caráter polar, apolar ou neutro dos aminoácidos é o outro fator que reforça as especificidades de cada aminoácido, pois numa proteína globular, quando em sua estrutura nativa e em solução, é observada a regra "*hydrophobic inside*" [69], isto significa que elementos apolares são encontrados comprimidos no interior da proteína, e os elementos polares, conseqüentemente, estão expostos na superfície em contato direto com o meio solvente.

Para a utilização da regra "*hydrophobic inside*" em um modelo, é necessário conhecer valores da chamada energia livre de transferência $\Delta G$ (ou hidrofobicidade) [67] para cada aminoácido. Os níveis hidrofóbicos de cada aminoácido são obtidos experimentalmente mediante a transferência de resíduos de um meio polar (água) para um meio apolar (óleo). Os resultados para essas energias livres de transferência produzem as denominadas escalas de hidrofobicidade. É interessante notar que, ao menos para pequenas moléculas apolares, $\Delta G$ (ou seja, o efeito hidrofóbico) varia muito pouco na faixa de temperatura que vai de zero a 100 $^{o}$C (detalhes dados no Capítulo 4, veja as tabelas 1 e 2). Os valores numéricos da escala de hidrofobicidade utilizada constituem um elemento importante para o modelo explorado aqui.

### 3.2 O alfabeto estereoquímico

Aqui é apresentado um candidato que leva em conta os aspectos discutidos nos parágrafos anteriores: trata-se da criação de um alfabeto estereoquímico, isto é, um conjunto mínimo de diferentes aminoácidos que consiste num repertório único, a partir do qual toda e qualquer seqüência ("proteína") será "montada". A aproximação de ordem zero na criação deste alfabeto é o estabecimento de quatro classes de aminoácidos em função do grau de exposição ao meio solvente (água), dada a posição espacial que cada aminoácido ocuparia em qualquer estrutura nativa. Essa classificação leva em conta o nível de hifrofobicidade do aminoácido, casando-o com o grau de exposição do mesmo ao solvente. Em nosso caso, para uma cadeia de tamanho $N$=27 ($N$ é o numero de aminoácidos), a estrutura nativa é representada por um cubo 3×3×3 imerso numa rede cúbica infinita, conforme ilustrado na Figura 1 [61]. A cadeia conformada num cubo 3×3×3 é a forma mais compacta possível para a mesma, e por isso tem sido denominada como configuração maximamente compactada (*compact self avoiding* CSA), com cada aminoácido ocupando um único sítio da rede. Os sítios não ocupados pela cadeia são preenchidos pelo solvente. Em outras palavras, no modelo estereoquímico simulado na rede cúbica regular, necessariamente os resíduos e grupos modulares de moléculas do solvente ocupam o mesmo volume; a cadeia ocupa sítios exclusivos e consecutivos de uma rede cúbica tri-dimensional infinita, enquanto os outros sítios, chamados de *sítios solvente*, são preenchidos pelas *moléculas efetivas do solvente.* Assim, os nomes das quatro classes: **0; 1; 2;** e **3**, indicam exatamente o número de contatos que cada um destes "aminoácidos" faz com os sítios solvente quando na estrutura nativa, conforme indicado na Figura 1.

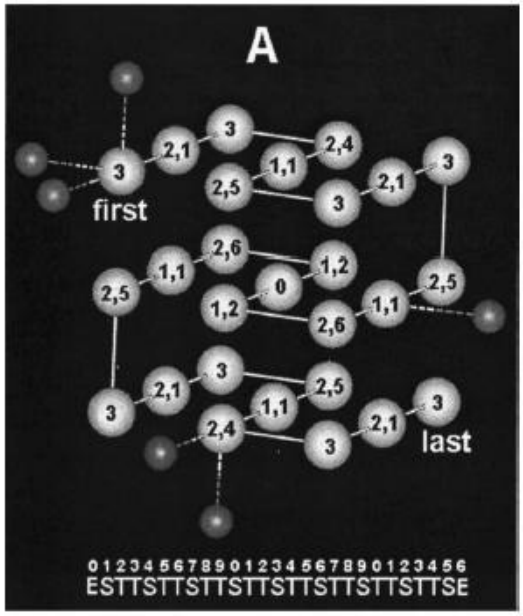


**Figura 1:** Configuração maximamente compacta (CSA) representando uma estrutura nativa. Note que diferentes "aminoácidos" contatam um número diferente de sítios do solvente [61].

Cada classe de "aminoácidos" possui um ou mais integrantes, as classes **0; 1; 2;** e **3** possuem respectivamente: um, dois, seis e um integrantes. Um refinamento, levando em conta detalhes locais da estrutura nativa, determina os valores específicos do nível hidrofóbico de cada "aminoácido" [67, 70], conforme resumido na Figura 2. Adicionalmente, simulações Monte Carlo mostram a importância das restrições estéricas tanto no processo de busca pela nativa, quanto na estabilidade da mesma [64]. Também na Figura 2 são mostrados os pares de "aminoácidos" que são permitidos realizar contatos como primeiros vizinhos na rede. A forma final desse "código estereoquímico", como mostrado na Figura 2, foi obtida por meio de um processo heurístico (tentativa e erro), envolvendo dezenas de simulações MC. Para isso, o tempo registrado em passos MC para encontrar a estrutura nativa pela primeira vez ao longo da simulação foi utilizado como a grandeza física a ser minimizada. Esse requerimento de minimização do tempo de busca pela nativa se justifica não apenas por razões experimentais e teóricas, mas também por razões fisiológicas.

Quanto às distintas configurações maximamente compactas, elas são exatamente 51.704 constituindo um conjunto de configurações que daqui para frente será identificado como "catálogo" de configurações CSA distintas; duas configurações CSA são idênticas se possuem o mesmo conjunto de 28 pares de contato entre primeiros vizinhos. Cada elemento do catálogo representa uma possível estrutura nativa (alvo) a ser estudada; tais configurações CSA's possuem padrões estruturais que as distinguem umas das outras.

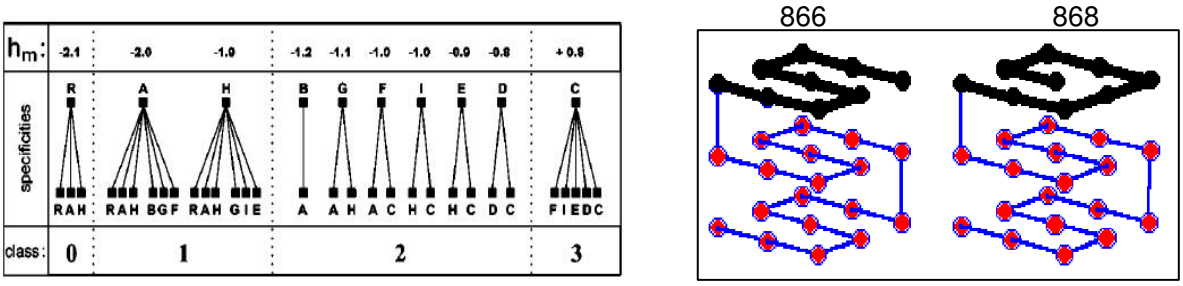


**Figura 2:** À esquerda, especificidades estéricas e níveis hidrofóbicos. Os níveis hidrofóbicos abrangem proporcionalmente o domínio do espectro de hidrofobicidades dos aminoácidos naturais. As letras R; A; H;...; C identificam cada "aminoácido", e os pares de letras conectadas indicam os pares de "aminoácidos" que podem ocupar sítios primeiros vizinhos da rede cúbica. À direita, duas estruturas nativas maximamente compactadas, identificados pelos números ID = 866 e 868, estas configurações são muito parecidas à primeira vista, mas a diferença nos padrões estruturais indicadas em preto implica em tempos característicos de *folding* consideravelmente distintos.

Uma sintaxe (regra) única, baseada no alfabeto estereoquímico, é provida para o *design* da seqüência de "aminoácidos" para qualquer uma das estruturas do catálogo [67]. É importante frisar que, quando se diz que a sintaxe é única, se quer dizer que a mesma regra é utilizada para o *design* de qualquer estrutura alvo escolhida. Assim, dada uma das estruturas alvo do catálogo, a seqüência de "aminoácidos" da "proteína" é automaticamente determinada. Embora esse alfabeto seja constituído por metade do número de aminoácidos de ocorrência natural, o modelo estereoquímico reproduz alguns intrigantes aspectos relacionados com o processo de *folding* como, por exemplo, a capacidade de encontrar a estrutura nativa em tempo hábil (virtualmente, para todas as estruturas do catálogo); a correlação das várias ordens de grandeza observadas para o tempo característico de *folding*, $\tau$, com as características topológicas das estruturas alvo, bem como a robustez do processo de *folding* dentro de um intervalo relativamente estreito de temperaturas.

### 3.3 Aspectos topológicos da nativa e o design da seqüência

Quando uma cadeia com $N$ aminoácidos assume sua compactação máxima em um espaço cúbico discreto ela se conforma em um cubo com dimensões $n \times n \times n$. O número de contatos do tipo primeiros vizinhos, chamados de contatos nativos ($CN$) é dado pela expressão [71]:

$$CN = 2n^3 - 3n^2 + 1 \qquad (1)$$

Aqui $n$=3, pois a cadeia tem 27 resíduos, portanto $CN$=28. Também na tese de doutorado de Inês Regina da Silva [71] se encontra uma classificação para os padrões topológicos que uma

cadeia de 27 resíduos maximamente compactada em um cubo pode apresentar. Essa classificação é apresentada aqui de modo sucinto, para isso o alvo 1128 (Figura 3) serve de exemplo. Para cada uma das 27 partes que compõem a cadeia cabe uma das três classificações: *E* (*End*), *T* (*Turn*) e *S* (*Straight*). Os dois resíduos que se encontram nos extremos da cadeia são classificados por E. Quanto aos 25 aminoácidos restantes eles são da forma *T* e *S* e satisfazem a relação:

$$n_T + n_S = 25 \qquad (2)$$

A grande maioria das estruturas nativas apresenta um número de $n_T$ aproximadamente quatro vezes maior do que o número de $n_S$. O número máximo de unidades do tipo *S* é $n_S^{max}$=11, e isto obriga o número mínimo de *T*'s ser $n_T^{min}$=14. Por outro lado, o número mínimo de unidades do tipo *S* é $n_S^{min}$=2, o que implica em $n_T^{max}$=23. Entretanto, uma parcela muito pequena do total das configurações apresenta esses valores extremos, detalhes em [71].

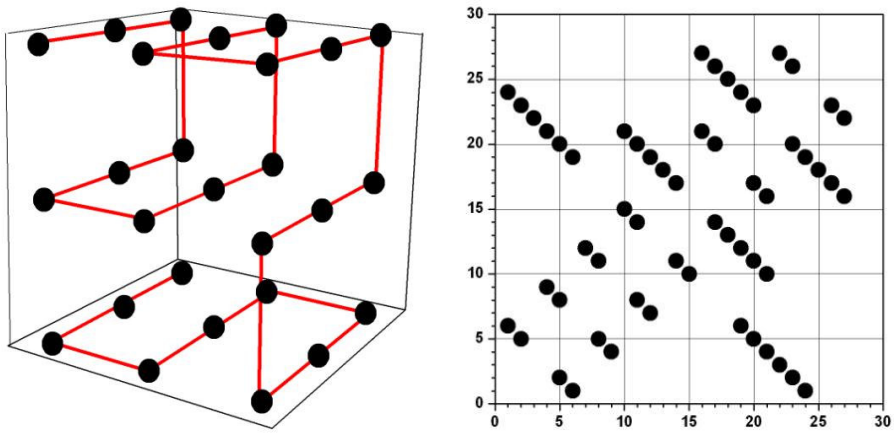


**Figura 3**: À esquerda estrutura 3D do alvo 1128. À direita seu respectivo mapa de contato, onde os contatos nativos são representados em duplicata, observe a diagonal secundária do desenho.

A combinação das unidades *S* e *T* podem formar padrões estruturais que se assemelham às estruturas secundárias encontradas em proteínas reais, por exemplo, as composições [**STTS**], [**TTTT**] e [**TT**], para as quais o asterisco (*) indica que qualquer uma das unidades *E*, *S* ou *T* pode dar continuidade a um destes padrões. Esses tipos de composições geométricas originam com freqüência formas modulares compostas por contatos entre aminoácidos que se encontram próximos ao longo da seqüência primária, e isto sugere o fácil acesso cinético para aqueles três padrões. Contudo, há padrões que representam formas modulares que não são de fácil acesso

cinético, por exemplo, as formas *Snail* e *Staple* (como aquela destacada em negrito no alvo 868 da Figura 1. É natural questionar qual é o efeito da composição dos dois tipos de padrões para a cinética do *folding* de um conjunto de alvos específicos. Realmente, a composição de dois ou mais padrões estruturais quaisquer pode resultar numa estrutura de maior ou menor acessibilidade, o que indica o quão complexa pode ser a determinação do nível de cooperatividade cinética de uma estrutura [70]. Comportamentos anômalos para alguns alvos correspondem a situações em que, por exemplo, estruturas compostas de padrões estruturais de fácil acesso --situação que pode ser facilmente identificada por inspeção no mapa de contato da estrutura em questão, apresentam comportamento cinético (ou seja, o tempo característico de *folding* $\tau$) que não corresponde ao esperado.

É importante mencionar que todos os resultados das simulações MC apresentados em [70] foram obtidos com o critério de Metropolis usual, isto é, a adoção do peso de Boltzmann. Nesta tese, como já mencionado em trechos anteriores, para lidar com as dificuldades apontadas é assumido que flutuações térmicas locais emergem como um atributo fundamental no sistema cadeia protéica-solvente, tratadas aqui com o auxílio do peso de Tsallis.

Quanto ao *design* da seqüência, sobre o alvo escolhido, é levado em conta, em primeiro lugar, o número de contatos que cada resíduo faz com as moléculas do solvente [61]. A partir daí é definido o conjunto $R(i)$, $i = 0, 1, 2, 3$, onde $i$ representa o número de contatos que um monômero encontrado no alvo faz com o solvente. Para cada estrutura maximamente compactada a distribuição das classes é feita da seguinte maneira: (*i*) há um único monômero $R(0)$ no centro do cubo; (*ii*) existem seis monômeros $R(1)$ nos centros das faces; (*iii*) há doze monômeros $R(2)$ no meio das arestas; e oito monômeros $R(3)$ nos vértices do cubo. Quanto à classificação $R(i)$, todas as 51.704 estruturas maximamente compactadas são idênticas, portanto é necessário especificar os integrantes de cada conjunto $R(i)$, para $i$ =1 e 2. Os representantes das classes $R(0)$ e $R(3)$ não possuem subgrupos, sendo simbolizados pelas letras *R* e *C*. Para os membros do conjunto $R(1)$ a decisão é feita por meio da topologia local em que o resíduo se encontra; o primeiro deles, identificado por $r(1,1)$, será adotado toda vez que o número de vizinhos topológicos na forma *S* é maior ou igual ao número de vizinhos topológicos na forma *T* (o monômero do centro do cubo R(0), não é levado em conta); conseqüentemente o subgrupo $r(1,2)$ fica automaticamente determinado. Essas duas subclasses são representadas pelas letras *A* e *H*. De modo semelhante os monômeros do conjunto $R(2)$ serão identificados obedecendo à regra hierárquica:

- *r*(2,1), quando os dois vizinhos topológicos são monômeros do tipo *r*(1,1) e *r*(1,1); sua representação é dada pela letra *B*;

- *r*(2,2), quando os dois vizinhos topológicos são monômeros do tipo *r*(1,1) e *r*(1,2); sua representação é dada pela letra *G*;

- *r*(2,3), quando os dois vizinhos topológicos são monômeros do tipo *r*(1,1) e *r*(3); sua representação é dada pela letra *F*;

- *r*(2,4), quando os dois vizinhos topológicos são monômeros do tipo *r*(1,2) e *r*(1,2), ou *r*(1,2) e *r*(3) na forma *T*, sua representação é dada pela letra *I*;

- *r*(2,5), quando os dois vizinhos topológicos são monômeros do tipo *r*(1,2) e *r*(3) na forma *E*; sua representação é dada pela letra *E*;

- *r*(2,6), quando os dois vizinhos topológicos são monômeros do tipo *r*(3); sua representação é dada pela letra *D*.

A representação dessa sintaxe se encontra condensada na Figura 2 anterior. Como exemplo da regra aplicada em uma estrutura alvo específica (866) veja a Figura 1 anterior.

### 3.4 A energética do modelo estereoquímico

Apresentados os aspectos principais do modelo segue a descrição da sua energética. Formalmente as interações presentes no modelo são de três tipos: (*i*) interações intra-cadeia: o efeito de volume excluído e as restrições específicas que impõem os pares permitidos para os primeiros vizinhos entre as partes que constituem a cadeia; (*ii*) interações solvente-solvente: aqui apenas o efeito de volume excluído é considerado; (*iii*) interações cadeia-solvente: outra vez o efeito de volume excluído é levado em conta, e, adicionalmente o efeito hidrofóbico é incluído através de uma escala numérica de hidrofobicidade. Como já mencionado essa escala é idealizada a partir de dados experimentais e estimativas teóricas para a energia livre de transferência dos aminoácidos [67].

As interações de esfera dura impedem que as partes que compõem a cadeia de uma molécula ocupem o mesmo espaço que já se encontra preenchido por outra parte da mesma molécula, ou que duas moléculas efetivas do solvente ocupem um mesmo sítio da rede [72]. Quanto a interação cadeia–solvente, que em princípio trata da interação explícita entre cada resíduo e o solvente, temos que para modelos em rede nos quais se adota a aproximação na qual a densidade do solvente não flutua, estas interações podem ser representadas de modo

exato pelo potencial hidrofóbico aditivo $g_{m,n} = (h_m + h_n)$ , entre os pares de resíduos [73]. O nível hidrofóbico $h_i$ do resíduo "*i*" representa a variação da energia livre para transferi-lo de um meio polar (solvente) para um meio apolar (interior da proteína). Daí, com a cadeia fixa numa certa configuração *a*, a energia livre do sistema pode ser escrita como $G = G_0 + \Sigma\, g_{m,n}$ (*a*), onde $G_0$ é uma constante. Portanto no contexto do modelo estereoquímico os níveis hidrofóbicos representam o fator químico mais geral e influente atuando no processo de *folding*. Adicionalmente o conjunto de vínculos intra-aminoácidos imita as especificidades estéricas dos resíduos reais pelo estabelecimento de quais deles são permitidos ocupar sítios primeiros vizinhos da rede. Esses vínculos funcionam como especificidades adicionais para os pares de contato do tipo primeiro vizinho, e, sua conseqüência principal é selecionar caminhos para o *folding* e *unfolding* através do espaço configuracional [64]. Em suma, o conjunto de especificidades estéricas acompanhado da variação da energia livre de transferência, que são fixos para cada par de monômeros, constitui os parâmetros do modelo.

Assim, a energia livre do sistema cadeia-solvente $G([\kappa,l])$, quando a cadeia é pensada fixa numa configuração arbitrária $\xi$, definida pelo conjunto $[\kappa,l]$ de $N_\xi$ contatos inter–resíduos do tipo primeiros vizinhos ($1 \le N_\xi \le 28$), pode ser escrita como

$$G([\kappa,l]) = \sum_{\{i,j\}} (g_{i,j} + c_{i,j})\delta_{(i,j),\,[\kappa,l]}\,, \qquad (3)$$

onde o somatório varre todos os pares (*i*, *j*) de resíduos; o fator $\delta_{(l,j),[\kappa,l]} = 1$ indica o caso em que (*i*,*j*) pertence ao conjunto $[\kappa,l]$, caso contrário $\delta_{(l,j),[\kappa,l]} = 0$.

É importante frisar que o modelo estereoquímico não é nativa-centrado, ou seja, informações sobre a estrutura nativa não são empregadas em etapa alguma ao longo das simulações Monte Carlo. A nativa somente é utilizada para o *design* da cadeia, e para a verificação do tempo de primeira passagem pela nativa. O processo como um todo é guiado exclusivamente pelas interações presentes no sistema cadeia-solvente. Com relação ao critério de amostragem entre sucessivas configurações do sistema, é feito o seguinte, antes de checar a variação energética entre uma configuração e outra, a permissão para os pares de primeiros vizinhos entre os resíduos que compõem a cadeia é verificada de acordo com a regra apresentada na Figura 1, e depois a variação energética é conferida.

Modelos fundamentados no potencial estereoquímico (Eq. (3)) têm provado ser eficientes em empacotar a cadeia e encontrar o estado nativo, mas eles falham em providenciar estabilidade ao glóbulo porque tais potenciais aditivos, $h_{i,j}=h_i+h_j$, satisfazem marginalmente o princípio da segregação (a saber, $2h_{i,j}-h_{i,j}-h_{j,j}\leq 0$) mediante o sinal da igualdade, ou seja, $2h_{i,j}-h_{i,i}-h_{j,j}=0$. Todavia, a adição de um conjunto de vínculos estéricos $\{c_{i,j}\}$ no potencial hidrofóbico $\{h_{i,j}\}$, conforme indicado na Eq. (3), leva à conseqüências importantes: por exemplo, eles ajudam a encontrar as rotas para o *folding*, e melhoram a condição de estabilidade do glóbulo no estado nativo como um todo [64].

Em resumo, no modelo apresentado aqui, as interações solvente-solvente e cadeia-cadeia são todas do tipo caroço duro (*hard sphere*), elas produzem o efeito geral conhecido como "volume excluído", e, portanto maior ênfase é dedicada às interações cadeia-solvente. Essa abordagem se justifica por duas razões de caráter experimental. A primeira evidência vem do plote de Ramachandran [74], o qual mostra a pronunciada compactação da estrutura protéica. Por sua vez, a segunda evidência, o foco nas interações cadeia–solvente vem do fato de que todas as interações deste tipo podem igualmente ser feitas entre a cadeia e o solvente, como no caso das ligações de hidrogênio, praticamente cancelando sua capacidade energética de discriminar entre a estrutura nativa e a desnaturada. Realmente, quando a conformação da cadeia se aproxima o suficiente da estrutura nativa é que as ligações de hidrogênio podem se configurar de modo produtivo na estabilidade do glóbulo protéico.

Detalhes sobre o desenvolvimento do modelo estereoquímico se encontram no Apêndice A.

### 3.5 As flutuações térmicas locais e a alteração no critério de Metropolis: a Mecânica Estatística Não Extensiva

Os modelos minimalistas destinados a investigar o processo de *folding* quando simulados com o método Monte Carlo adotam na maioria das vezes o critério de Metropolis usual, centrado apenas na variação energética, para realizar a amostragem estatística [75, 76]. Entretanto, como já mencionado em trechos anteriores, o sistema cadeia-solvente é um sistema heterogêneo e fora do equilíbrio termodinâmico, com um componente finito, pois parte do sistema (a cadeia protéica) pertence à escala nanoscópica [18]. Logo, antes de prosseguir com a argumentação sobre qual peso estatístico (ou quais pesos estatísticos) adotar

no critério de Metropolis, são apresentadas algumas estimativas termodinâmicas envolvendo o sistema finito abordado, bem como as correspondentes motivações experimentais.

### *3.5.1 Flutuações térmicas locais na escala nanométrica*

Os dados tradicionais sobre a cinética e termodinâmica de proteínas são resultados de médias temporais do comportamento coletivo de muitas moléculas, algo em torno de $10^{17}$–$10^{20}$ moléculas/L. Ou seja, esse é um fator determinante para gerar a visão tradicional das proteínas como um glóbulo rígido, compacto e estático. Contudo, quando consideradas individualmente, cada proteína garantidamente está sob efeito de flutuações consideráveis em suas propriedades termodinâmicas. Façamos algumas estimativas, considera-se uma proteína representativa com aproximadamente 250 resíduos em solução – massa molecular $m$ em torno de 4.2 x $10^{-23}$ kg, a qual apresenta capacidade térmica típica ($C_p \cong C_v$) em torno de 0.32 kcal $kg^{-1}K^{-1}$, a temperatura de $T$=300K. Uma estimativa direta da flutuação da energia interna em torno da média, para uma molécula individual, resulta em $\Delta U_{rms}$ = 6.4x$10^{-20}$ cal por molécula ($\Delta U_{rms}=kT^2mC_V$). Marcantemente, ela deve ser da mesma ordem de grandeza da mudança de entalpia típica da desnaturação térmica das proteínas, ou seja, algumas dezenas de Kcal/mol [32], caso todas as moléculas estivessem flutuando em harmonia [77]. Entretanto, as flutuações são individualmente não correlacionadas, logo numa população com um número grande de macromoléculas, elas irão praticamente se cancelar, resultando em uma medição efetivamente sem flutuação nos parâmetros de interesse (veja a argumentação de Popov no quarto parágrafo na página 447 em [18]).

Apesar da quantidade de dados gerados pelos experimentos tradicionais em bioquímica, o processo de *folding* deve ser explicado para uma única molécula protéica em particular. Essa necessidade tem motivado os experimentos de uma única molécula como apontado por Kumar e Li [34], e, recentemente comentados por Shakhnovich [78] sobre o mapeamento das trajetórias de transição entre o estado *unfolded* e *folded* realizados por Eaton e colaboradores para uma proteína globular de dois estados com 56 aminoácidos, a GB1 [79]. A importância desse tipo de medição reside no fato de que o mapeamento do estado de transição demarca onde o *folding* e o *unfolding* realmente ocorrem. Dada a escala nanoscópica da cadeia no sistema cadeia-solvente, é como se a temperatura estivesse flutuando localmente, e isto justifica uma proposta para levar em conta a flutuação térmica local nas simulações Monte Carlo, como é apresentado a seguir.

### *3.5.2 O fator de Boltzmann generalizado*

Por conta da sua típica baixa concentração, para estudar o processo de *folding*, uma proteína em solução pode ser pensada como um sistema heterogêneo constituído por apenas uma única cadeia imersa no meio solvente, o qual funciona como um reservatório de calor à temperatura, digamos $\beta_0^{-1}$. Então, se $\beta^{-1}$ oscila (localmente) em intervalos de tempo menores do que o tempo típico gasto para as trocas configuracionais da cadeia, é possível propor um fator de Boltzmann generalizado $\exp_q(-\beta_0 E)$ como a integral do fator de Boltzmann (função exponencial convencional $\exp(-\beta_0 E)$ ) sobre todas as possíveis flutuações $\beta^{-1}$, ponderado por certa distribuição $f_q(\beta)$ , ou seja:

$$\exp_q(-\beta_0 E) = \int_0^\infty d\beta \exp(-\beta E) f_q(\beta) . \qquad (4)$$

Caso seja assumido que $f_q(\beta)$ obedeça uma distribuição da forma $\chi^2$ - conforme proposto por Beck [80], e explorado por Rajagopal e colaboradores [81] – um caso especial da distribuição Gama da variável $\beta$ que ocorre em muitas circunstâncias comuns [82], então o fator de Boltzmann generalizado se torna

$$\exp_q(-\beta_0 E) = [1-(1-q)\beta_0 E]^{\frac{1}{1-q}} .$$

Ou seja, exatamente a mesma expressão proposta por Constantino Tsallis como peso estatístico no formalismo generalizado da Mecânica Estatística Não Extensiva [83]. Quanto à distribuição $\chi^2$,

$$f_q(\beta) = [\Gamma(n/2)]^{-1}\left(\frac{n}{2\beta_0}\right)^{n/2} \beta^{n/2-1} \exp(-n\beta/2\beta._0) \qquad (5)$$

ela é parametrizada de modo que a temperatura $\beta_0^{-1}$ corresponda à média da flutuação em $\beta$, ou seja: $\beta_0 =< \beta >= \int_0^\infty f_q(\beta)\beta d\beta$. Portanto, o parâmetro $q$, dado por $q = 1 + 2/n$, está associado com a dispersão relativa de $\beta$, $q = 1 + (<\beta^2> - \beta_0{}^2)/\beta_0{}^2$, com $n$ sendo o número de

graus de liberdade do subsistema em contato com o reservatório térmico. A proposta feita por Beck [80] recebe o nome de Super Estatística de Beck-Cohen, alguns detalhes sobre isso são dados no Apêndice B.

Nesta tese, o foco é direcionado para os aspectos cinéticos do curso seguido por uma cadeia que se encontra numa conformação desnaturada qualquer, e que transita para a conformação nativa. Para este estudo, as simulações Monte Carlo são realizadas utilizando o critério de Metropolis, contudo com uma modificação nas probabilidades de transição entre sucessivas configurações, decorrente do mesmo raciocínio utilizado na Eq. (4). Assim, para a transição de uma configuração *a*, digamos com energia $E_a$, para uma configuração *b*, com energia $E_b$, a probabilidade de transição utilizada no processo de Metropolis tradicional, a saber, $\exp(-\beta_0 \Delta E_{ab})$, onde $\Delta E_{ab} = E_b - E_a$, é substituída por

$$\exp_q(\beta_0 \Delta E_{ab}) = < \exp(-\beta_0 \Delta E_{ab}) >, \text{ ou}$$

$$\exp_q(-\beta_0 \Delta E_{ab}) = [1-(1-q)\beta_0 \Delta E_{ab}]^{\frac{1}{1-q}} . \qquad (6)$$

Quanto à amostragem entre duas configurações sucessivas, conforme já mencionado em trechos anteriores, antes de realizar a avaliação energética, como é feito no critério de Metropolis usual, é checada a permissão para os pares de primeiros vizinhos conforme indicado na Figura 2.

### *3.5.3 A motivação para o uso do peso de Tsallis*

O sistema cadeia-solvente, como um todo pode ser pensado no contexto do *ensamble* microcanônico, com a cadeia focada como um subsistema e o meio solvente como o sistema complementar. Assim, todas as configurações do sistema global, envolvendo a cadeia e o solvente, tem o mesmo peso $p = 1/\Gamma_0$. Aqui $\Gamma_0$ é o número total de configurações do sistema como um todo. Entretanto para uma mesma configuração da cadeia há um número muito grande de configurações para o solvente. Daí, se a cadeia é considerada *congelada* em uma configuração particular *a*, a entropia do sistema é reduzida por uma quantia $\Delta S_a$. Portanto, na formulação tradicional, a probabilidade de ocorrência da configuração *a* é dada pelo peso de Bolztmann $p_a = \exp(-\Delta S_a / k_B)$ [84, 85]. Contudo o sistema aqui estudado é heterogêneo. Como já mencionado, se trata de uma cadeia com dimensões nanoscópicas em solução aquosa, logo, uma

vez que o comportamento da cadeia é o foco da nossa atenção, o caos microscópico não pode ser ignorado. Para a aplicação do balanço detalhado, o sistema como um todo pode ser pensado no contexto do *ensamble* microcanônico, e, em equilíbrio térmico macroscópico. Assim sendo uma probabilidade de transição $W(a \rightarrow b)$, da configuração $a$ para a configuração $b$, pode ser estabelecida, ou seja, $W(a \rightarrow b) = \exp(-\beta_0 \Delta S_{ba})$, onde $\Delta S_{ba} = \Delta S_b - \Delta S_a$. Para lidar com as flutuações térmicas locais é adotada a proposta não extensiva de trocar o peso estatístico usual por sua correspondente generalização a $q$-exponencial [86], e a probabilidade de transição fica dada por $W(a \rightarrow b) = \exp_q (- \beta_0 \Delta S_{ba})$. Feita a suposição adicional de que no estágio de busca da nativa não há praticamente variações significativas na energia interna do sistema, a probabilidade de transição é rescrita como $W(a \rightarrow b) = \exp_q (- \beta_0 \Delta G_{ba})$. Onde $G_b$ ($G_a$) é a energia livre do sistema como um todo quando a cadeia se encontra (congelada) na estrutura $b$ ($a$) [84, 85].

A motivação para adotar o peso de Tsallis em nossas simulações MC, como já indicado, reside no fato de que uma molécula protéica está sujeita ao efeito de forças locais não balanceadas, quando em contato com um reservatório térmico, a uma dada temperatura macroscópica $\beta_0^{-1}$. Uma situação muito parecida ocorre com uma partícula de dimensões coloidais, onde essas flutuações se originam da natureza finita (nanométrica) do objeto, modulando as complexas interações entre suas partes. Vários outros exemplos são dados no artigo de Beck e Cohen [80].

No contexto mais geral, as motivações ficam por conta de que: (*i*) o processo de *folding* de proteínas pode ser considerado como um processo estocástico não linear fora do equilíbrio termodinâmico, como defendido por Popov [18] (veja o quarto parágrafo da página 445, bem como o segundo parágrafo da página 446); (*ii*) e que para lidar com a termodinâmica dos sistemas vivos a Mecânica Estatística Não Extensiva pode ser uma escolha conveniente [87]. Feita a opção pelo peso de Tsallis, o efeito da Mecânica Estatística Não Extensiva sobre o sistema estudado é analisado de duas maneiras. Numa etapa preliminar, o índice entrópico $q$ é escolhido ser uma grandeza fixa ao longo de um conjunto de simulações MC. Por sua vez, na segunda etapa, $q$ é deixado variar, ao longo de cada simulação, em função do grau de compactação da cadeia, estimado aqui pelo raio de giração $R_G$ da mesma, enquanto que a temperatura do reservatório permanece fixa. Isso é feito ajustando o valor de $q$, configuração após configuração da cadeia, por meio da relação $q = 1+$

$\alpha / R_G^2$; α é uma constante que determina o valor máximo possível para $q$ (este ponto será retomado mais adiante, no tópico 4.2).

### 3.6 Análise das funções exponencial usual e *q*-exponencial

Os valores significativos de $q$ para o presente problema são relativamente pequenos. Portanto é interessante registrar as comparações entre a função $q$-exponencial e a exponencial usual. No limite de $q \to 1$, a função $q$-exponencial, $\exp_q(-x) = [1-(1-q)x]^{\frac{1}{1-q}}$, apresenta exatamente o mesmo comportamento da função exponencial convencional, ou seja, $\exp_{q\to 1}(-x)/\exp(-x) = 1$. Entretanto, para $q$ = 1,1 a diferença $\Delta_q(x) = \exp_q(-x) - \exp(-x)$ atinge seu maior valor em $x \cong 2.1$, veja na Figura 4, a razão $\Delta_q(x)/\exp(-x) \cong 16\%$.

Assim, em princípio, é possível argumentar que o efeito líquido de usar a $q$-exponencial ao invés da função exponencial convencional, no problema aqui considerado, deve ser o equivalente a um pequeno aumento local da temperatura. Com o intuito de testar essa possibilidade, nesta tese o tempo característico de *folding* τ é adotado como um parâmetro analítico para analisar os efeitos da utilização da Mecânica Estatística Não Extensiva sobre o modelo estereoquímico.

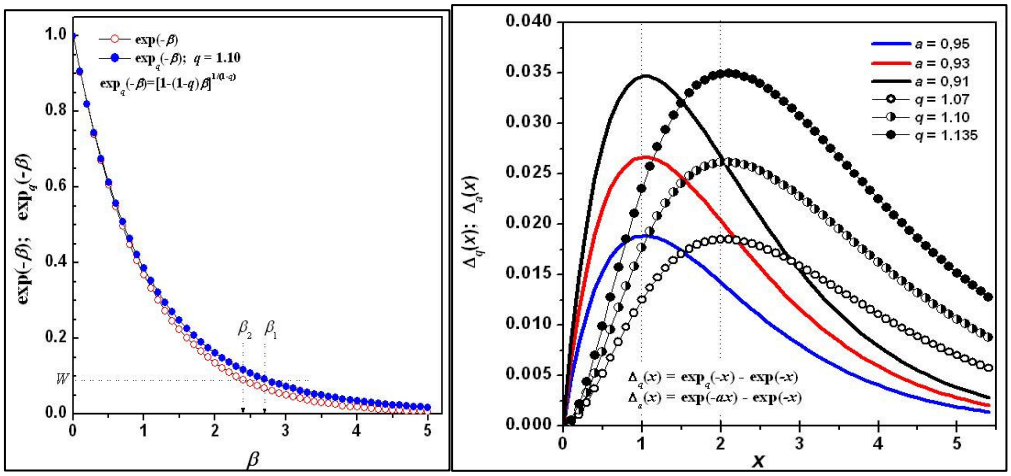


**Figura 4**: À esquerda comparação do comportamento da exponencial generalizada, a $q$-exponencial $\exp_q(-\beta)$ quando $q$=1.1, com a exponencial usual exp(-β). À direita comparação das duas diferenças: $\Delta q(x) = \exp_q(-x)-\exp(-x)$, e $\Delta a(x)=\exp(-ax)-\exp(-x)$. Seus perfis são similares, todavia o máximo para $\Delta a(x)$ ocorre em torno de $x_0 = 1$, e, para $\Delta q(x)$ ele é em torno de $x_0 = 2$.

Outros aspectos da Mecânica Estatística Não Extensiva são retomados no Capítulo seguinte, e também no Apêndice – B. No próximo item os parâmetros OC e $\Omega$ propostos para avaliar a topologia da estrutura nativa são discutidos.

### 3.7 A Ordem de Contato e o parâmetro Ω

Como mencionado no final do Capítulo introdutório um dos principais desafios relacionados com o problema do *folding* de proteínas é explicar as taxas *folding* para diversas proteínas globulares. Proteínas semelhantes em tamanho (quantia de aminoácidos na estrutura primária), mas com padrões topológicos distintos presentes na estrutura nativa, podem apresentar diferentes ordens de grandeza para a taxa de *folding*, numa mesma faixa estreita de temperaturas. A proposta feita por Plaxco e colaboradores [16, 17] indica que, de alguma maneira a taxa de *folding* deve estar relacionada com parâmetros geométricos globais da nativa, e assim o parâmetro chamado Ordem de Contato (OC) foi proposto.

A OC relativa, proposta por Kevin Plaxco e colaboradores [16, 17], tem sido apontada como um elemento importante no estudo da taxa de *folding* [15]. Essencialmente ela determina a distância média ao longo da cadeia entre todos os pares $(i,j)$ de resíduos que se apresentam em contato (próximos) na estrutura nativa, ou seja:

$$OC = \frac{1}{N_c} \sum_{(i,j)} \frac{d_{i,j}}{L} \qquad (7)$$

onde $d_{ij}/L$ é a separação ao longo da cadeia entre os resíduos $(i,j)$ em contato, normalizado pelo comprimento total da cadeia. A soma é realizada sobre todos os pares de contato $(i,j)$, e $N_c$ é o número total de contatos presentes na estrutura nativa. Em geral os aminoácidos são considerados em contato quando a distância entre os seus carbonos-α (ou centro de massa dos resíduos) for menor do que 6 Å. Porém, variações desse valor têm sido usadas, por exemplo, 5,5 e 7,5 Å [88]. Quanto aos valores para a OC do conjunto das 51.704 estruturas maximamente compactas eles se situam no intervalo $0.238095 < OC < 0.494709$. Apesar de o parâmetro OC ser amplamente explorado na literatura [17], ele não apresenta nenhum apelo físico como, por exemplo, avaliar a probabilidade de formação dos padrões estruturais que compõem a estrutura nativa, isto é, ser capaz de predizer a acessibilidade cinética da nativa mediante uma inspeção direta sobre o mapa de contato da estrutura escolhida.

Um possível parâmetro que possa avaliar a probabilidade de formação de contatos quando a cadeia se encontra imersa em um banho térmico, e, apenas interações do tipo caroço-duro controlam seus movimentos, deve ponderar favoravelmente que os contatos mais prováveis são aqueles com menor distância $d_{ij}$ de separação entre os resíduos em contato. Para associar algum significado físico à $d_{ij}$ associamos a ela certa probabilidade $p_{ij}$ de ocorrência do contato entre o para $(i,j)$ de resíduos. Na aproximação da cadeia ideal (cadeia gaussiana) [89] esta probabilidade pode ser escrita como:

$$p_{ij} \propto d_{ij}^{-3/2}, \qquad (8)$$

como será demonstrado no Capítulo seguinte. Nessa aproximação cada contato intra-cadeia é considerado como independente dos demais, daí aqueles contatos prévios que tornam os seguintes mais fáceis, não são levados em conta. Conseqüentemente, cada estrutura nativa é ponderada em seu menor nível de cooperatividade. Assim a cooperatividade estrutural $\Omega$ pode ser representada pela média aritmética do correspondente conjunto de probabilidades $\{p_{ij}\}$. Logo é razoável especular que, se os fatores cooperativos adicionais são incluídos, a acessibilidade cinética fica adequadamente caracterizada por $\Omega$. Detalhes adicionais sobre a motivação física relacionada com o parâmetro proposto, bem como as correlações entre as taxas de *folding* para 200 estruturas alvo com os parâmetros OC e $\Omega$ se encontram no próximo Capítulo.

# 4 – RESULTADOS e DISCUSSÕES

A apresentação dos resultados das simulações MC é feita em quatro partes. Em todas elas o tempo característico de *folding* $\tau = \tau(T, q)$ é adotado como a grandeza física de relevância no estudo da primeira etapa do processo de *folding*, a busca pela estrutura nativa. No primeiro item o índice entrópico $q$, juntamente com a temperatura $T$ do reservatório, foi escolhido ser uma grandeza fixa ao longo dos conjuntos de simulações, e então o comportamento de $\tau(T, q)$ foi analisado para três estruturas alvo distintas. Na segunda parte $q$ foi deixado variar, ao longo de cada simulação, em função do grau de compactação da cadeia, conforme já mencionado no Capítulo anterior, e o valor médio dos $q$'s acessados é mostrado para os mesmos alvos do tópico anterior, novamente o comportamento de $\tau(T, q)$ é analisado. O segundo item é encerrado com exibição da freqüência de visitação dos valores de $q$ ao longo das simulações para seis alvos. Na terceira parte são mostradas as correlações entre a rapidez do *folding* $\log(1/\tau)$ com os parâmetros OC e $\Omega$ para um conjunto de 200 estruturas alvo distintas simuladas com o índice $q$ considerado como uma variável dinâmica. Finalmente como último item a questão da robustez na primeira etapa do processo de *folding* é brevemente discutida. Em todas as quatro partes o papel da topologia da nativa na taxa do *folding* é frisado.

## 4.1 O tempo característico de *folding* quando *q* é considerado um parâmetro fixo

A Tabela 1 mostra $\tau$ para três estruturas nativas como função da temperatura $T$ e do índice entrópico $q$, no intervalo $0.9 \leq T \leq 1.5$ e $1 \leq q \leq 1.4$. Naquela Tabela 1 observa-se que $\tau$ depende da topologia da estrutura, de $T$, e $q$. Em geral, são realizadas cento e cinqüenta simulações independentes para cada par $(T, q)$. Dentre as estruturas presentes na Tabela 1 (ID 866, 1128, 36335) a estrutura ID 36335 é a que apresenta a maior complexidade topológica, um fato refletido pelo maior valor do seu tempo característico em relação aos outros dois alvos. Todavia, para qualquer estrutura do catálogo, dada uma temperatura $T_i$ existe um valor específico $q_= q(T_i)$ que minimiza $\tau$, e para cada estrutura particular há um par $(T_i, q(T_i))$, para qualquer $T_i$, que dá aproximadamente $\tau \cong \tau_{min}$ (Tabela 1, células escuras). A incerteza $\delta\tau$ em $\tau$ foi estimada pelo desvio padrão da média das médias, considerando 20 amostras distintas $\{t_i\}_N$ tomadas de um conjunto de $10^4$ simulações independentes ($N$ =150). A incerteza $\delta\tau$ depende do par $(T, q)$ e é fortemente influenciada pelo tamanho do conjunto de simulações independentes quando o sistema se aproxima do regime vítreo, situação na qual a cadeia pode ficar presa em conformações metaestáveis por longos intervalos de tempo. As menores incertezas (em torno de 5%) ocorrem

para aqueles valores específicos de $q = q(T_i)$ os quais minimizam o tempo característico de *folding* $\tau = \tau_{min}$. As respectivas estruturas tridimensionais e os mapas de contato para os três alvos 866, 1128 e 36335 se encontram no Apêndice D.

| | ID 1128 | | *q* | | | ID 866 | | *q* | | | ID 36335 | | *q* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***T*** | 1.0 | 1.1 | **1.2** | 1.3 | 1.4 | 1.00 | **1.1** | 1.2 | 1.3 | 1.4 | 1.0 | 1.1 | **1.2** | 1.3 | 1.4 |
| 0.9 | 560 | 54 | 20 | 22 | 48 | 330 | 30 | 31 | 30 | 45 | 8700 | 870 | 240 | 310 | 470 |
| **1.0** | 130 | 28 | **19** | 26 | 67 | 120 | **24** | 31 | 36 | 81 | 3800 | 560 | **190** | 380 | 1100 |
| 1.1 | 40 | 22 | 25 | 40 | 130 | 47 | 26 | 30 | 50 | 130 | 1450 | 230 | 210 | 520 | 1700 |
| 1.2 | 39 | 18 | 29 | 65 | 320 | 24 | 26 | 36 | 90 | 310 | 680 | 200 | 260 | 780 | 3200 |
| 1.3 | 21 | 21 | 52 | 190 | 470 | 21 | 26 | 57 | 190 | 620 | 460 | 180 | 370 | 1560 | ... |
| 1.4 | 23 | 30 | 80 | 280 | 1000 | 24 | 36 | 92 | 320 | 1400 | 170 | 210 | ... | ... | ... |
| 1.5 | 28 | 47 | 170 | 810 | 2300 | 26 | 59 | 170 | 710 | 3100 | 200 | 290 | ... | ... | ... |

**Tabela 1**: Tempos característicos de *folding* em função da temperatura *T*, do parâmetro *q*, e da complexidade da estrutura nativa para três alvos distintos: 1128, 866, 36335. Observe as células escuras, elas correspondem ao $\tau = \tau_{min}$ para determinados pares de valores (*T*,*q*).

O cenário apresentado na Tabela 1 sugere que a cinética dos mecanismos de busca em ambos os casos são igualmente reproduzidas, não importando se as configurações são ponderadas com o peso de Tsallis a uma dada temperatura *T*, ou pelo peso de Boltzmann convencional com a temperatura um tanto elevada em relação à temperatura *T* (macroscópica) do reservatório. Realmente, essa possibilidade se torna clara caso o comportamento $\tau$ é plotado, para cada *q* específico como uma função da escala de temperatura transladada $T_q = T - T_q^*$, onde $T_q^*$ é a temperatura na qual $\tau = \tau_{min}$ para aquele valor de *q*, como mostrado na Fig. 1: essencialmente todas as curvas se comportam de um modo muito próximo para todos os casos em que $\tau \approx \tau_{min}$.

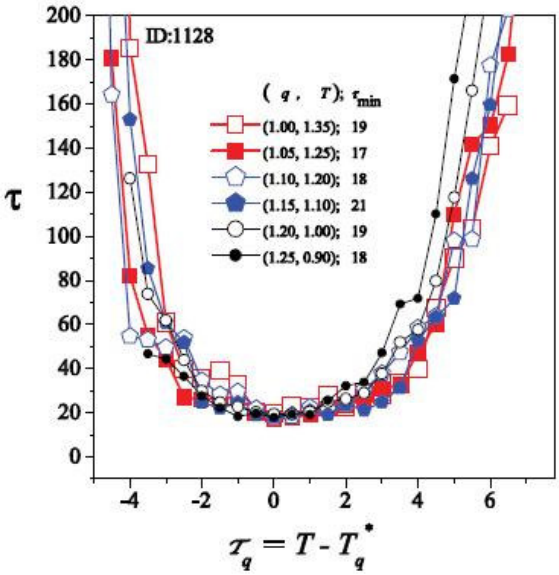


**Figura 1**: $\tau$ para várias combinações (*q*, *T*) que convergem para $\tau$ ótimo ($\tau = \tau_{min}$); os eixos da temperatura foram transladados para $T_q = T\text{-}T_q^*$, onde $T_q^*$ é a temperatura na qual $\tau$ aproxima-se do valor ótimo para cada caso.

Um exame mais detalhado mostra que as distribuições dos tempos de *folding* são essencialmente as mesmas em ambos os casos, como mostrado na Figura 2 para a estrutura representativa ID 866 em diferentes circunstâncias ($q$ e $T$), e $N = 10^4$ simulações independentes para cada caso. As distribuições são mais bem ajustadas com uma ou mais curvas log-normal, dependendo da temperatura. A distribuição para ($q$, $T$) = (1.1, 1) (peso de Boltzmann generalizado; curva contínua mais fina) é praticamente a mesmo que para ($q$, $T$) = (1, 1.25) (peso de Boltzmann convencional; pequenos círculos vazios), os quais implicam, na convergência de $\tau$ para os dois casos, a saber, $\tau_{min} = 24$ (Tabela 1). Essa concordância confirma que o efeito do peso de Boltzmann generalizado sobre a cinética do mecanismo de busca pela nativa é equivalente, da perspectiva do peso de Boltzmann usual, a um aumento na temperatura (ou seja, a inclusão das flutuações térmicas locais). Contudo, apesar da aparente equivalência entre o efeito do peso de Boltzmann a temperaturas elevadas, e o peso de Tsallis com o índice entrópico $q$ devidamente ajustado para encontrar $\tau_{min}$, há uma diferença entre as duas abordagens. Essa diferença pode ser encontrada quando o tempo médio $\langle t \rangle$ é calculado. Com a abordagem não extensiva $\tau_{min}$ e $\langle t \rangle$ são muito próximos, o que não ocorre quando o peso de Boltzmann é utilizado para encontrar o $\tau_{min}$.

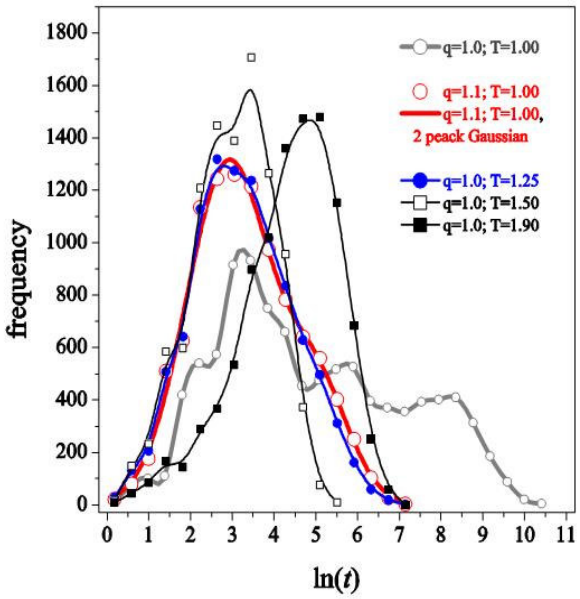


**Figura 2**: Distribuições dos tempos de *folding* $\Phi$ para o alvo 866. A distribuição log-normal para $q$=1.1 e $T$=1.0 (círculos vazios maiores) é mais bem ajustada usando uma Gaussiana com dois picos (linha contínua em vermelho – amplitude ($A_1$; $A_2$) = (1310 ± 40; 430 ± 70); largura ($w_1$; $w_2$) = (0.49 ± 0.07; 0.8 ± 0.3); e média ($\mu_1$; $\mu_2$) = (2.91 ± 0.09; 5.0 ± 0.2)). Para $q$=1.0 o comportamento do tempo de *folding* é mostrado para várias temperaturas. Em particular, as distribuições dos tempos de *folding* para ($q$, $T$) = (1.0; 1.25) e (1.1; 1.0) são praticamente as mesmas, isto confirma que o peso de Tsallis ($q$>1) a uma temperatura $T$, corresponde ao fator de Boltzmann usual ($q$=1.0) com certo aumento em $T$. Na região dos menores tempos de *folding* ($ln(t)$ < 2.5) e para temperaturas no intervalo $1.0 \leq T \leq 1.5$, todas as curvas se aglutinam – inclusive o caso ($q$, $T$)=(1.1, 1.0). Em $T$=1.0 e $q$=1.0 (linha cinza) o sistema se aproxima do regime vítreo e manifestações das dificuldades ergódicas são evidentes. Na temperatura $T$=1.50 a distribuição apresenta o menor domínio, ou seja, 0 < $ln(t)$ < 5.5 (quadrados vazios), e conforme $T$ aumenta a partir deste ponto, o pico da distribuição se move para os $t$'s maiores, conforme ilustrado para $T$=1.90 (quadrados preenchidos).

Para $(q, T) = (1, 1)$, o sistema se aproxima do regime vítreo com manifestações das dificuldades ergódicas, como indicado pela curva Gaussiana com três picos. Em $T \cong 1.5$ a distribuição apresenta seu menor domínio, a saber, $0 < \ln(t) < 5.5$, e progressivamente para $T>1.5$, o tamanho do domínio aumenta e o pico da curva se desloca para os maiores valores de $t$. Quanto à região dos menores tempos de *folding*, a saber, para $\ln(t) <2.5$, todas as curvas no intervalo $1.2 \leq T \leq 1.5$ apresentam o mesmo comportamento incluindo o caso onde $(q, T) = (1.1, 1)$. Isso significa que, mesmo com a temperatura 25% maior, entre as configurações iniciais há algumas que combinadas com certa evolução configuracional, podem levar a cadeia rapidamente à estrutura nativa.

Como já mencionado em um dos parágrafos anteriores é importante prestar atenção nas diferenças entre as duas abordagens aqui discutidas, principalmente por conta das peculiaridades do sistema considerado. Realmente, a temperatura do sistema não pode ser mudada arbitrariamente; não é uma simples questão de renormalização de parâmetros. O mecanismo de busca pela nativa deve ser universal, trabalhando igualmente no intervalo de temperatura entre zero e $100^{o}$C. Contudo, uma vez que a estrutura nativa é encontrada, outra etapa do processo de *folding* entra em cena. Essa etapa é a estabilização da nativa como um todo, a qual é uma função estritamente dependente da temperatura do reservatório térmico. Para cada proteína há um intervalo da temperatura de trabalho, fora do qual ela perde sua funcionalidade. Portanto, a temperatura $T$ do sistema (banho térmico) deve ser mantida como referência, medida macroscopicamente, e todas as peculiaridades térmicas de um corpo em escala nanoscópica, em resposta às flutuações térmicas locais (por exemplo, o tempo característico ótimo $\tau = \tau_{min}$) seriam controladas pelo índice entrópico $q$.

Para qualquer proteína alvo do catálogo, mantida à temperatura constante (temperatura macroscópica do reservatório), o ajuste de um valor específico de $q$ para obter o $\tau_{min}$ é sempre possível. Contudo, a questão de como o índice entrópico $q$ pode ser associado a cada proteína em particular (sua seqüência de aminoácidos ou sua estrutura nativa) deve ser investigada. Para o caso presente, a saber, uma cadeia que evolui no espaço configuracional para um glóbulo compacto, a idéia direta vem da observação que a flutuação de $\beta$ é dependente da escala espacial conforme sugerido por Beck, Lewis e Swinney [80, 90].

Mediante a relação entre $q$ e o número de graus de liberdade $n$ da distribuição da temperatura, ou seja, $q=1+2/n$ é possível reconhecer uma sutil associação entre $q$ e a complexidade da estrutura nativa. Pode-se perceber, em primeiro lugar, a natureza dinâmica dos graus de liberdade $n$. Conforme a cadeia muda seu grau de compactação ao longo do tempo, $n$ muda de acordo: então, para uma cadeia aberta $q_{\min}=2/n_{\max} \rightarrow 1$, e para um glóbulo plenamente compactado $q_{\max}=2/n_{\min}$. Considerando que para uma cadeia plenamente compactada $n_{\min}=3$ é possível obter um valor máximo para $q$, ou seja, $q_{\max}=1+2/3 \cong 1.67$. Realmente, durante o processo de *folding*, armadilhas energéticas e topológicas devem ser superadas até que a estrutura alvo seja alcançada. Tais armadilhas fazem com que a cadeia colapse em conformações erradas, e em diferentes graus de compactação, as quais devem ser desfeitas de modo que o processo de *folding* possa ser retomado. Portanto, a cadeia em diferentes níveis de compactação é suscetível aos efeitos térmicos locais de diferentes maneiras, que ao longo do processo de simulação são governados por valores distintos de $q$. Assim, dependendo do número e peculiaridades das armadilhas em potencial – determinadas pela combinação da complexidade da estrutura nativa e do *design* da cadeia (ou seja, a seqüência dos resíduos), certo valor médio (ou valor ótimo) do índice entrópico $q$ pode ser identificado por um processo de triagem. Isso é o que foi descrito em trechos anteriores como mostrado na Tabela 1 para três estruturas alvo diferentes. Contudo, uma inspeção direta desse processo pode ser realizada usando um processo dinâmico que muda (apropriadamente) o valor de $q$ ao longo da simulação; este é o foco do próximo item.

## 4.2 O tempo característico de *folding* quando *q* é considerado uma variável dinâmica

Aqui os três alvos analisados na Tabela 1 são novamente o foco de análise, com uma diferença fundamental: desta vez o índice entrópico $q$ foi deixado variar ao longo de cada simulação MC em função do raio de giração, mediante a expressão $q = 1 + 2(q_{max}-1)$. Na configuração compacta $R_G^2 = 2$ e assim, $q = q_{max}$, por outro lado com a cadeia aberta $q$ deve apresentar o comportamento $q \rightarrow 1$. O valor de $q_{max} = 1{,}33$ parece ser universal para o presente modelo, uma vez que para este valor e com a temperatura do reservatório $T = 1$, $\tau \rightarrow \tau_{\min}$ para qualquer estrutura simulada.

| | ID 1128 | | | ID 866 | | | ID 36335 | | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | $\tau$ | <*t*> | <q> | $\tau$ | <*t*> | <q> | $\tau$ | <*t*> | <q> |
| 0,8 | 37 | 44 | 1,16 | 47 | 84 | 1,16 | 330 | 354 | 1,19 |
| 0,9 | 23 | 30 | 1,15 | 43 | 45 | 1,15 | 283 | 294 | 1,18 |
| **1** | 18 | 25 | 1,15 | 29 | 31 | 1,14 | 243 | 253 | 1,17 |
| 1,1 | 21 | 25 | 1,14 | 27 | 31 | 1,13 | 242 | 246 | 1,16 |
| 1,2 | 27 | 31 | 1,13 | 29 | 32 | 1,13 | 292 | 296 | 1,15 |
| 1,3 | 41 | 45 | 1,13 | 40 | 44 | 1,12 | 387 | 387 | 1,14 |
| 1,4 | 68 | 73 | 1,12 | 64 | 68 | 1,11 | 551 | 553 | 1,13 |

**Tabela 2**: Tempos característicos de *folding* $\tau$, e tempos médios de *folding* <*t*> em função da temperatura *T*, e do índice entrópico *q*, e $q_{max}$ = 1,33. Aqui, como mostrado na Tabela 1, a complexidade da estrutura nativa para os três alvos 1128, 866, 36335 influi em $\tau$ e <*t*>. Observe as células azuis, elas correspondem ao $\tau = \tau_{min}$ dentro do domínio de temperaturas *T* do reservatório utilizadas, e que nestes casos $\tau$ e <*t*> são muito próximos.

O cenário apresentado na Tabela 2 mostra que: (*i*) para os menores tempos característicos $\tau = \tau_{min}$, os valores de <*t*> e $\tau_{min}$ são muito próximos; (*ii*) que o $\tau_{min}$ para os três alvos é obtido numa parte estreita dentro do domínio da temperatura do reservatório *T*; (*iii*) o valor médio de *q* acessado ao longo das simulações é próximo daquele valor de *q* o qual fornece $\tau_{min}$ quando o índice entrópico é utilizado como parâmetro (veja na Tabela 1, e imagine o caso em que uma interpolação entre os valores vizinhos de *q* é feita).

Na série de figuras seguintes, são mostradas distribuições dos valores de *q* para seis alvos diferentes. Para todos os casos adotamos *T* = 1, $q_{max}$ = 1,33, e um conjunto total de 150 simulações para cada estrutura foram necessárias para a determinação do tempo característido de *folding* $\tau$. A ênfase é dada para a distribuição de *q*, mas na parte superior de cada figura é também mostrada a curva cinética, a qual indica a evolução do número de proteínas desnaturadas, cujos dados computacionais são ajustados por uma, ou em alguns casos, por duas exponenciais.

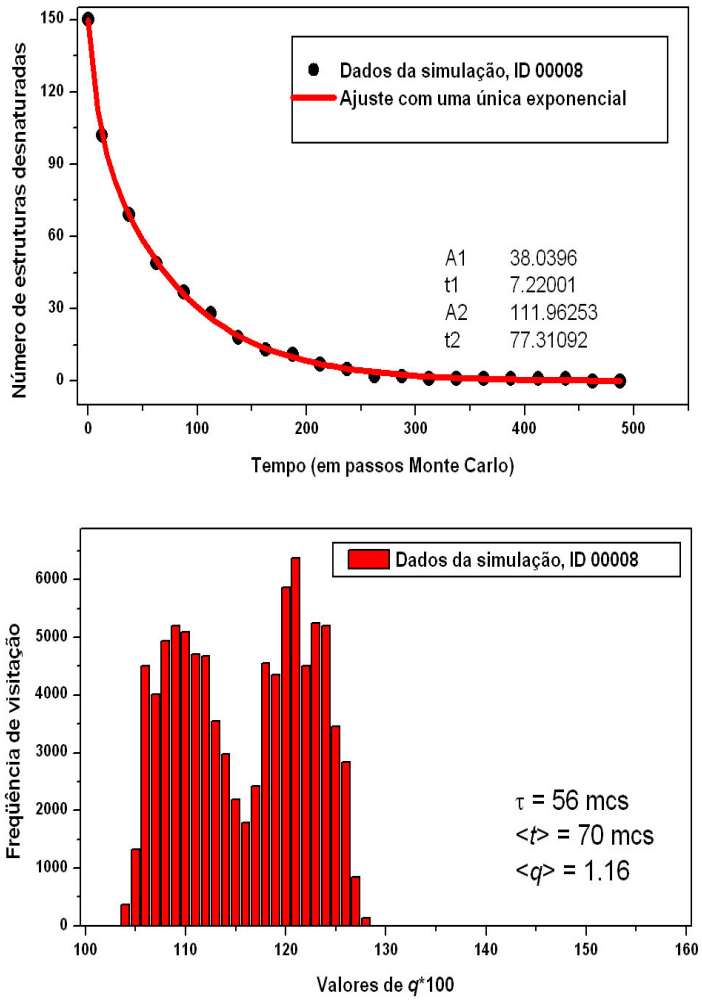


**Figura 3-A**: Freqüência de visitação dos valores de *q* para o alvo ID 8. Observe a presença de uma distribuição bimodal para a distribuição dos valores visitados pela variável *q*.

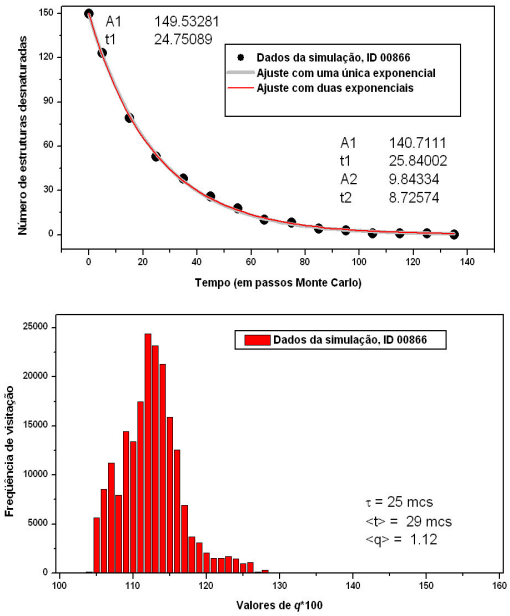


**Figura 3-B**: Freqüência de visitação dos valores de *q* para o alvo ID 866. Neste caso, temos uma distribuição unimodal para a freqüência dos valores visitados pela variável *q*. Portanto, os valores para τ, <*t*> e <*q*> apresentados neste gráfico são consistentes com aqueles presentes nas Tabelas 1 e 2 anteriores para o alvo 866.

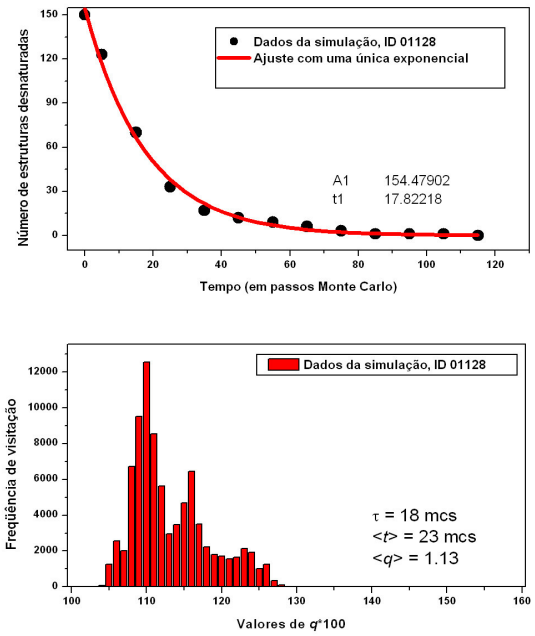


**Figura 3-C**: Freqüência de visitação dos valores de *q* para o alvo ID 1128. Aqui, também podemos ver tendências para uma distribuição bimodal, embora uma das componentes da distribuição predomine visivelmente. Os valores para τ, <*t*> e <*q*> apresentados neste gráfico também são consistentes com aqueles presentes nas Tabelas 1 e 2 anteriores para o alvo 1128.

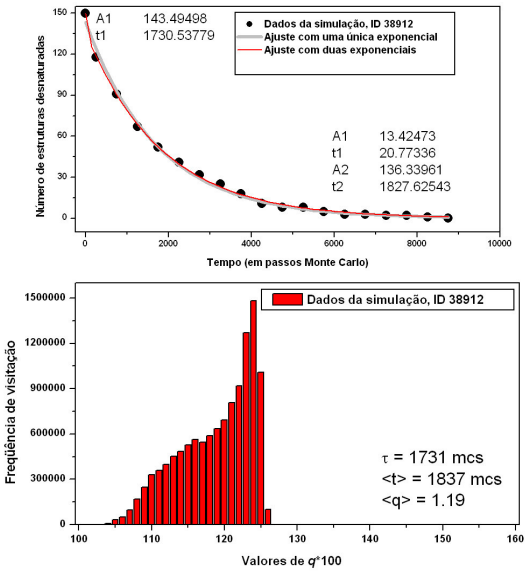


**Figura 3-D**: Freqüência de visitação dos valores de *q* para o alvo ID 38912. Note a assimetria acentuada desta distribuição, com a predominância dos valores maiores de *q*. Essa tendência significa que na evolução configuracional, muitas estruturas compactas, diferentes da nativa são visitadas, indicando sinonímia (degenerecência): uma mesma sequência (palavra) levando a mais de uma estrutura compacta (significado).

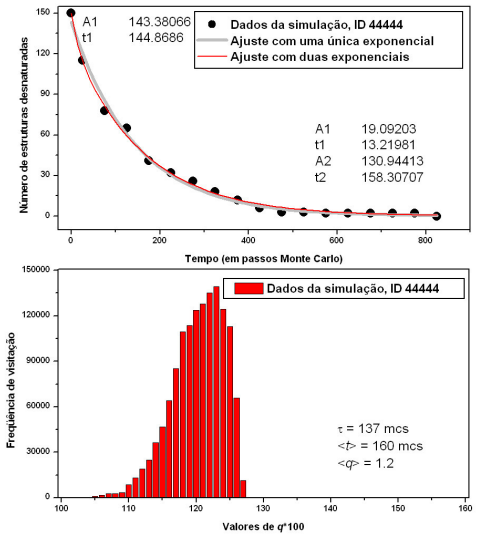


**Figura 3-E**: Freqüência de visitação dos valores de *q* para o alvo ID 44444. Aqui se tem um caso muito similar ao anterior, Figura 3-D.

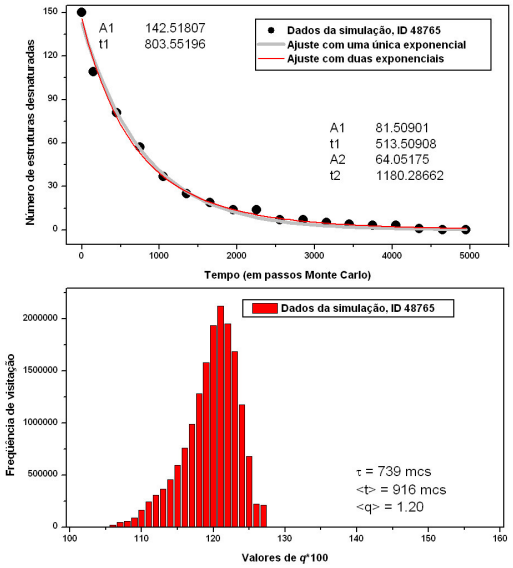


**Figura3-F**: Freqüência de visitação dos valores de $q$ para o alvo ID 48765. Nesse caso também a predominância dos valores de $q$ acessados indica que a cadeia freqüentemente acessa conformações compactas diferentes da estrutura nativa, o que faz com que a cinética do processo de *folding*, neste caso seja mais lenta.

Da série de gráficos apresentados nas Figuras acima (Figura 3 A–F), vemos que a distribuição de $q$ é bem variada. Para algumas estruturas os valores de $q$ acessados durante a simulação concentram-se ao redor de valores bem menores do que $q_{max}$; para outras estruturas, os valores de $q$ tendem a $q_{max}$; ou ainda para outras estruturas a distribuição dos $q$'s é bimodal. Cada uma destas tendências singulares revela uma classe de comportamento cinético. No primeiro caso, quando os valores de $q$ são predominantemente pequenos, isto significa que as rotas para a estrutura nativa são, na sua grande maioria, muito bem determinadas e rápidas, o que é verificado nas estruturas compostas por padrões estruturais mais cooperativos. No caso oposto, quando a frequência dos $q$'s é maior para valores próximos de $q_{max}$, o que se constata é que a cadeia se compacta muitas vezes, ao longo da simulação, porém em estruturas diferentes da nativa, e assim deve ser desfeita sistematicamente pela ação mais acentuada da flutuação térmica local. Como resultado, o tempo característico de *folding,* $\tau$, para esses casos são significativamente maiores. Estruturas com topologias mais complexas, pobres em padrões com contatos próximos ao longo da cadeia, tendem a mostrar este comportamento. No caso intermediário, com distribuição bimodal, temos a coexistência de dois grupos de rotas distintas à nativa; por algumas rotas o acesso é mais rápido, enquanto que por outras o processo de *folding* é marcadamente muito

mais lento, pois provavelmente muitas conformações maximamente compactas erradas serão acessadas e conseqüentemente desfeitas.

De acordo com a discussão apresentada no parágrafo anterior, o peso de Tsallis para pequenas proteínas cuja cinética é de dois estados [35] pode desempenhar o papel dos "chaperones": quanto mais compacta for a estrutura, mais acentuadas serão as flutuações térmicas locais, portanto desfazendo mais facilmente o glóbulo. Por isso, armadilhas energéticas e topológicas, muito mais freqüentes nas conformações compactas, serão mais facilmente sobrepujadas e assim o espaço conformacional é mais bem visitado. Evidentemente, o mesmo efeito será exercido sobre a estrutura nativa, e a tendência será igualmente de desfazê-la pelo efeito das flutuações térmicas mais acentuadas nas estruturas compactas. Contudo, nas proximidades da conformação nativa, e somente nela, é que todas as interações se compõem de forma cooperativa e produtiva, permitindo a estabilidade do glóbulo.

A distribuição dos $q$'s também nos revela que o valor ótimo de $q$, isto é, aquele valor $q = q^*$ que minimiza o tempo característico de *folding* $\tau$, é no final praticamente igual ao valor médio $\langle q \rangle$ da distribuição dos $q$'s, ou seja, $q^* \approx \langle q \rangle$. Realmente, os tempos característicos de *folding* $\tau$ obtidos com o parâmetro fixo $q = q^*$ e com $q$ como uma variável convergem muito satisfatoriamente.

### 4.3 A correlação entre a "velocidade" do processo de *folding* com a OC e Ω

À semelhança do que se encontra apresentado nas duas Tabelas anteriores (1 e 2) seria possível escolher como referência um daqueles três alvos, por exemplo, a estrutura 866, e então investigar como $\tau$ se altera quando apenas mudanças estruturais sutis são permitidas. A análise desse tipo de investigação não foi dada nesta tese, visto que ela envolveria a classificação e descrição detalhada dos aspectos topológicos de cada um dos padrões estruturais envolvidos, e assim umas poucas estruturas alvo poderiam ser tratadas. Alternativamente, escolhemos aqui simular um conjunto considerável e representativo de estruturas alvo (para esta tese foram simulados 200 alvos distintos), sujeitos a certa temperatura do reservatório, e então estimamos $\tau$ para cada uma delas. E em seguida, investigamos a correlação entre a rapidez do processo de *folding* (a busca pela nativa) com parâmetros globais destinados a avaliar a topologia dos alvos escolhidos, isto é OC e Ω.

A proposta feita aqui tem como fundamentação o trabalho de Plaxco e Baker [16]. Adicionalmente acreditamos que, o trabalho que mais se aproxima com o que se encontra mostrado a seguir, seja o artigo de Kaya e Chan [93]. Lá, em [93], o papel da topologia da nativa nas taxas de *folding* é o alvo principal das análises feitas. No fim do artigo, no último parágrafo, os autores apontam algumas lacunas na compreensão da energética das proteínas reais. Contudo, o nosso ponto de vista indica (como é mostrado nos gráficos a seguir), que as deficiências apontadas no final daquele artigo ficam por conta de: (*i*) a simplificação demasiada assumida por modelos do tipo Go (os quais usam informações da nativa ao longo das simulações); (*ii*) o peso estatístico usado nas simulações MC; (*iii*) o modo com que a nativa é avaliada. Na tese de doutorado de Inês Regina da Silva [71] são encontrados avanços na descrição da topologia da estrutura nativa mediante a classificação dos padrões estruturais, também são propostas modificações na OC, e são dadas correlações similares aquelas exploradas na seqüência. Os principais resultados da tese de Inês se encontram condensados no artigo [70]. Com o intuito de avançar com a investigação dada em [70, 71] propomos o parâmetro Ω.

Ainda que o processo do *folding* de proteínas seja estocástico ele não se processa de forma aleatória, pois apresenta taxas de *folding* com dezenas de ordens de grandeza mais rápidas do que se poderia esperar de uma busca aleatória pela estrutura nativa [50]. Contudo, o tempo característico de *folding* pode abranger várias ordens de grandeza, variando de frações do microsegundo a horas. O tamanho da proteína é um dos fatores determinantes do tempo característico de *folding*, mas detalhes topológicos também exercem grande influência na acessibilidade cinética à estrutura nativa. Realmente, proteínas pequenas com algumas dezenas a centenas de aminoácidos podem apresentar taxas de *folding* muito distintas, como é o caso discutido por Gillespie e Plaxco em [17].

Como já mencionado no Capítulo anterior, para proteínas pequenas de dois estados (aquelas sem estados intermediários detectáveis entre o estado nativo e não-nativo [35]), tem sido mostrado que o logaritmo da taxa de *folding* correlaciona-se negativamente com a OC [16, 35]. O curioso é OC ser considerado um parâmetro global, pois ele essencialmente é a média das distâncias ao longo da cadeia entres os resíduos que fazem contato na estrutura; Eq. (7) do Capítulo 3.

A Eq. (7) do Capítulo 3 (definição da OC) carece de um significado físico preciso. Contudo, intuitivamente não é difícil aceitar que a OC de cada estrutura reflete de alguma forma a facilidade com que a mesma possa ser acessada, pois contatos entre resíduos a uma distância λ ao longo da cadeia ocorrem com uma freqüência que aumenta com o inverso de λ [61]. Assim, estruturas com menores valores de OC, seriam mais acessíveis no processo de busca estocástica.

Com o intuito de aprimorar a idéia de representar a acessibilidade da estrutura nativa por um único parâmetro global, é possível substituir cada termo da Eq. (7) anterior por sua probabilidade de ocorrência. Feita essa identificação, em sua formulação mais simples, o resultado seria uma probabilidade média, determinada pelas probabilidades de ocorrência de cada elemento do conjunto de contatos intra-cadeia $\{i,j\}$. A dedução rigorosa de um parâmetro como esse é uma tarefa hercúlea. Contudo, quando a aproximação da cadeia ideal para homopolímeros é utilizada (cadeia gaussiana), uma primeira aproximação é de obtenção imediata, como é dado a seguir.

Considere uma cadeia polimérica simples com $N$ unidades idênticas confinadas numa rede cúbica. Cada configuração consiste numa seqüência de $N$ passos consecutivos, representada graficamente por um traço conectando $N$ pontos consecutivos da rede. A Figura 4 abaixo ilustra uma configuração arbitrária numa rede 2-D.

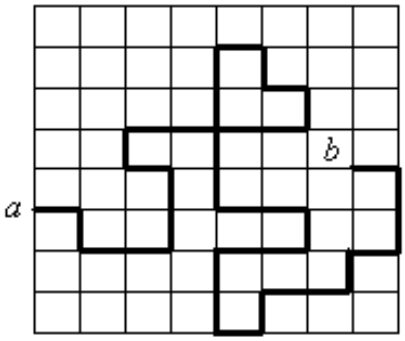


**Figura 4**: Configuração ***C*** arbitrária no espaço bi-dimensional. Qualquer configuração pode ser representada graficamente pelo caminho descrito por $N$ passos mediante os pontos consecutivos da rede. O comprimento $L$ total da cadeia é então $L = N\ell$, onde $\ell$ é o parâmetro de rede. A distância entre os extremos $a$ e $b$ da cadeia é chamada "distância ponta-a-ponta" (dpp). Perceba que no presente caso, à semelhança dos gases ideais, a superposição entre as unidades da cadeia é permitida [72].

A entropia $S(\mathbf{r}) = k\ln\Gamma(r)$ é calculada sobre o número $\Gamma(r)$ de todas as configurações disponíveis para a cadeia. Quanto às distâncias entre as extremidades $a$ e $b$, elas são calculadas como: $\mathbf{r}^2 = (i_a - i_b)^2 + (j_a - j_b)^2$ (todas as configurações são equiprováveis). Portanto o número de todas as possíveis configurações fica dado por $\Gamma_N = \sum_r \Gamma(r) = z^N$, onde z +1 é a coordenação da rede. No presente caso, a rede cúbica, $z = 5$.

Para uma dada configuração arbitrária ***C,*** o vetor ponta-a-ponta **r,** que liga um extremo ao outro da cadeia é dado pela soma dos *N* passos entre os primeiros vizinhos na rede, ou seja, $\mathbf{r} = \sum_{i=1}^{N} \ell_i$. Com o vetor **r** é possível a determinação analítica de algumas grandezas físicas relevantes em polímeros. Por exemplo, quando se leva em conta todas as possíveis configurações Γ da cadeia, é possível mostrar que a média do quadrado da distancia ponta-a-ponta, $R_0^2$, é $R_0^2 =< \mathbf{r}^2 >= \sum_{nn}\sum_{m} \ell_n \cdot \ell_m = \sum_{n} \ell^2 = N\ell^2$, isto é, $R_0 = \ell N^{1/2}$ (os termos cruzados, $n \neq m$, se anulam). Outra grandeza de interesse, dada uma distância fixa *r*, é a probabilidade de ocorrência *p*(*r*) de uma configuração arbitrária, a qual tem a forma gaussiana, isto é,

$$p(r) = \frac{A}{N^{3/2}} \exp\left( -\frac{3r^2}{2N\ell^2} \right), \tag{2}$$

onde *A* é uma constante. A rigor esse resultado é válido somente nas condições de cadeia ideal e quando são assumidas as seguintes aproximações: (*i*) o número de monômeros que compõem a cadeia é muito grande (*N*>>1); e (*ii*) quando a distância *r* é muito menor do que o comprimento da cadeia, quando completamente esticada (*Nl*), satisfaz, *r/Nl << 1*. Detalhes sobre os cálculos de $p(r)$ e outras grandezas decorrentes, podem ser encontrados no primeiro Capítulo de um texto em português [91], ou em inglês e nas referências contidas em [92]. Apresentada a expressão para *p*(*r*) segue como utilizá-la para definir a cooperatividade estrutural.

### *4.3.1 Probabilidade de ocorrência de um contato*

Para uma cadeia com 27 resíduos que se encontra em uma rede cúbica, a sua configuração maximamente compacta (que representa uma estrutura nativa) é composta por 28 contatos intra-cadeia (primeiros vizinhos). Cada estrutura nativa apresenta um conjunto de contatos topologicamente único, que pode ser mais ou menos complexo, e isto resulta em distintos padrões estruturais. Essas diferenças estruturais implicam na peculiar acessibilidade cinética ao longo do processo de *folding*. O contato entre os resíduos *i* e *j* ocorre quando eles são primeiros vizinhos na rede cúbica (a uma distância espacial fixa $\ell$). Um contato dessa natureza é caracterizado pela distância $d_{ij}$ entre os resíduos *i* e *j* ao longo da cadeia. Na Eq. (2) anterior é assumido que $r = \ell$, ou seja, a distância espacial entre resíduos contatantes (igual ao

parâmetro de rede) e, considera-se que $N = d_{ij}$, como a distância ao longo da cadeia entre os resíduos contatantes. Assim podemos associar uma probabilidade $p_{ij}$ a cada contato, em função da distância $d_{ij}$ ao longo da cadeia, ou seja,

$$p_{ij} = \frac{A}{d_{ij}^{3/2}} \exp\left(\frac{-3}{2d_{ij}}\right). \tag{3}$$

Logo, com a substituição direta de $d_{ij}$ pela estimativa de sua probabilidade de ocorrência $p(d_{ij}) = p_{ij}$, na Eq. (7) (Capítulo 3) para a OC, é introduzido um novo conceito para se estimar, em primeira ordem, a cooperatividade de uma dada estrutura, com a intenção de avaliar a sua acessibilidade cinética, por um único parâmetro global, o qual será representado por $\Omega_1$. Dessa forma temos que,

$$\Omega_1 = \frac{1}{N_c} \sum_{\{i,j\}} p_{ij} \,. \tag{4}$$

Embora a Eq. (4) acima represente uma aproximação grosseira, diferentemente do parâmetro global OC, ela tem um significado físico claro: $\Omega$ mede a acessibilidade cinética pela probabilidade média de ocorrência de cada contato inter-resíduo.

Quanto ao fator exponencial $c_{ij} = \exp\left(\frac{-3}{2d_{ij}}\right)$, ele é definido no intervalo $0{,}6 < c_{ij} < 1$. Na prática, esse fator pode ser ignorado sem maiores prejuízos, mantendo-o fixo; $c_{ij} = 1$. Feita essa escolha, a cooperatividade mínima $\Omega_0$ para uma dada estrutura fica definida como

$$\Omega_0 = \frac{1}{N_c} \sum_{\{i,j\}} \frac{A}{d_{ij}^{3/2}} \,. \tag{5}$$

Nós simulamos 200 estruturas alvo distintas, todas elas com a temperatura de referência do banho térmico em $T = 1$. Cada alvo foi simulado cento e cinqüenta vezes, isto gerou um conjunto de 150 tempos de primeira passagem, com o qual calculamos $\tau$. Os valores de $\tau$ para as diversas estruturas consideradas, cobrem quatro ordens de grandeza, e com eles investigamos as correlações entre $\log(1/\tau)$ e OC, $\Omega_0$ e $\Omega_1$, veja as figuras 5, 6 e 7 dadas em

seguida. Quanto à $\Omega_1$ sua estimativa foi realizada levando o fator exponencial da Eq. (3) em conta. Entretanto como é mostrado na Figura 5 a correlação entre $\log(1/\tau)$ e $\Omega_1$ não difere significativamente daquela entre $\log(1/\tau)$ e $\Omega_0$. Realmente $\Omega_1$ varia linearmente com $\Omega_0$, com a Figura 7 podemos inferir que $\Omega_1 = \alpha\,\Omega_0$, onde $\alpha < 1$.

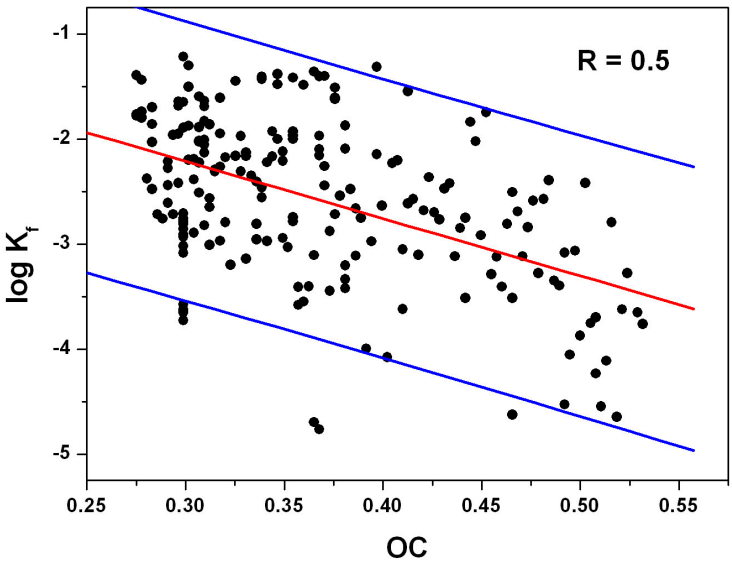


**Figura 5** - Correlação R (Pearson) entre as taxas de *folding* e OC para o conjunto de 200 alvos simulados, e temperatura $T$ =1 (unidades de *energia/K*); R =50% . Observe que $K_{eff}$ varre quatro ordens de grandeza.

Perceba na Figura 5 as ordens de grandeza varridas pelas taxas, temos tempos característicos de *folding* que vão da dezena a dezena de milhar em passos MC. Isso significa que a combinação do modelo estereoquímico com a Mecânica Estatística Não Extensiva é capaz de encontrar as estruturas alvo para todas as estruturas consideradas (sem nenhum critério de exclusão) em intervalos de tempo razoáveis, porém amplos. Entretanto, o mesmo não ocorreu quando o peso de Boltzmann foi utilizado para simular alvos mais complexos, ainda que uma janela de tempo com 1.000.000 de passos MC, ou mais, tenha sido utilizada em alguns casos. Quanto à correlação mostrada na Figura 5, o valor de 50% indica que a OC não é satisfatória para descrever a topologia da nativa e assim contribuir para o elucidamento do(s) mecanismo(s) de *folding*. Nas duas figuras seguintes as correlações entre o $\log(K_f)$ e variações de $\Omega$ são exploradas.

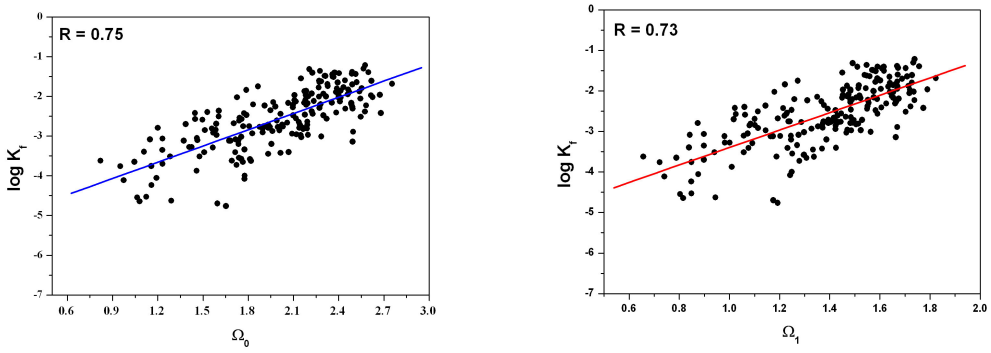


**Figura 6**: Correlações entre as taxas de *folding* para o conjunto de alvos simulados com $\Omega_0$ (à esquerda) e $\Omega_1$ (à direita). Observe à esquerda que com $\Omega_0$ a correlação aumentou em 50% em relação àquela com a OC, por outro lado com $\Omega_1$ a correlação aumentou em 46% novamente em relação àquela realizada com a OC.

A Figura 6 acima indica que a cooperatividade estrutural $\Omega$ em suas versões mais básicas ($\Omega_0$ e $\Omega_1$) é capaz de melhorar em aproximadamente 50% a correlação com as taxas de *folding*. Portanto, podemos dizer que $\Omega$ é um parâmetro mais conveniente do que a OC para avaliar a topologia da nativa. Adicionalmente também podemos dizer que a acessibilidade cinética é um fator muito importante relacionado com o mecanismo envolvido na primeira etapa do processo de *folding*, a busca pela nativa. Entretanto, a análise detalhada do papel da composição de padrões mais cooperativos (acessíveis), com outros não tão cooperativos na cinética de cada uma das 200 estruturas simuladas não é dada nesta tese, conforme já foi mencionado antes. Para encerrar este item cabe uma análise sobre a diferença nas correlações mostradas na Figura 7 para $\Omega_0$ e $\Omega_1$.

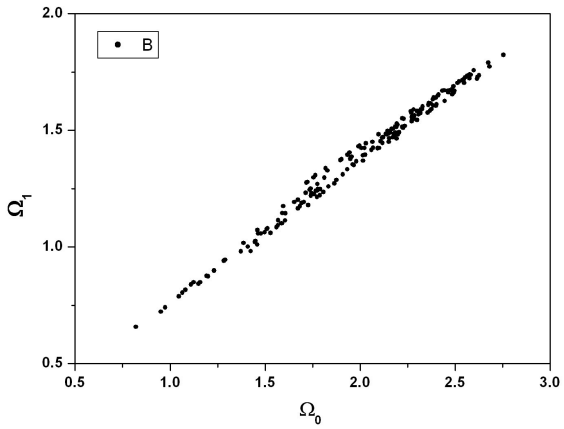


**Figura 7**: A comparação entre $\Omega_0$ e $\Omega_1$ para vários alvos, por conta do fator exponencial da Eq. (3) calculado em $\Omega_1$, percebe-se que seu efeito é praticamente negligenciável.

O resultado mostrado na Figura 7 nos dá uma boa margem de segurança para usar $\Omega_0$ em nossas análises, bem como na caracterização da cooperatividade estrutural para os alvos simulados. Apresentadas as correlações anunciadas este Capítulo é encerrado com um breve comentário sobre a robustez do processo de *folding*.

### 4.4 A robustez do processo de *folding*

A questão da robustez na busca pela estrutura nativa no processo de *folding* pode ser colocada em termos diretos da seguinte maneira. Caso a regra para a construção da cadeia (a sintaxe) é alvo independente, e, se para um alvo particular ela é capaz de ditar a seqüência que se enovela no alvo especificado, dentro de uma janela de tempo hábil, dada uma temperatura de referência para o banho térmico ($T_r$=1); então é esperado que para temperaturas um tanto maiores ou menores do que $T_r$, a estrutura nativa também pode ser encontrada, porém com valores maiores para o tempo característico de *folding*. Adicionalmente a universalidade da sintaxe deve garantir o mesmo comportamento para os $\tau$'s dos demais alvos do catálogo quando a temperatura do banho térmico se afasta de $T_r$. Em particular, esse comportamento pode ser visto em um dos gráficos do artigo [94], no qual a estrutura alvo de interesse foi a 866, ou para a estrutura alvo 1128 na Figura 1 anterior.

# 5 – CONCLUSÕES e PERSPECTIVAS

Ao longo desta tese de doutorado o problema do *folding* de proteínas foi abordado sob a perspectiva do problema do *folding* inverso através de um modelo minimalista, o modelo estereoquímico [61]. Dentro do contexto desse modelo o processo investigado é pensado como sendo composto por duas etapas: (*i*) a busca pela estrutura nativa, e (*ii*) a estabilização da mesma [64]. O principal foco da análise aqui apresentada se deu sobre a primeira etapa do processo, isto é, a busca pela nativa, e conseqüentemente a investigação do papel da topologia do alvo sobre o tempo característico de *folding*. Para isso utilizamos 200 alvos distintos em nossas simulações Monte Carlo, sendo que o comportamento cinético de alguns deles foi analisado em detalhes (estruturas 866, 1128, 36335). Em nossas simulações MC o peso estatístico de Tsallis foi utilizado. Para justificar essa escolha, foram usados tanto argumentos físico-químicos, quanto biológicos que implicam na necessidade das estruturas nativas serem alcançadas agilmente, em janelas de tempos razoáveis e numa faixa relativamente estreita de temperaturas. Alguns argumentos plausíveis para essa escolha podem ser encontrados no conjunto de referências [14, 15, 16, 34, 50, 18, 29, 77, 80, 81, 82, 83, 87, 94] e mais recentemente em [95, 96]. Com isso, um dos nossos objetivos principais foi alcançado, como ilustra a Figura 1 do Capítulo 4. Nesta figura mostramos os menores tempos característicos de *folding* do alvo 1128 para diversos pares de valores ($T$, $q$); naquela Figura, o eixo das temperaturas foi transladado, fazendo com que as curvas fossem sobrepostas, facilitando assim a inspeção visual. Esse gráfico (Figura 2) ilustra o fato de que o efeito do peso de Tsallis é equivalente a certa elevação da temperatura macroscópica do sistema; porém, tendo em mente que a questão da estabilidade do glóbulo protéico, esta simples renormalização de temperatura não torna as duas abordagens equivalentes. Assim, o índice entrópico $q$ do peso estatístico de Tsallis representa convenientemente as flutuações térmicas locais do sistema cadeia-solvente, permitindo que a temperatura $T$ permaneça associada com a noção usual de uma grandeza macroscópica de referência, isto é, a temperatura do banho térmico.

A meta de prover evidências sobre a competição entre *controle cinético* e *controle termodinâmico* no processo de busca pela nativa, como decorrente dos atributos topológicos da própria estrutura alvo, está resumida nas Tabelas 1 e 2 do Capítulo 4. A inspeção direta nos respectivos mapas de contato dos alvos utilizados (Apêndice D) ajuda a entender que os atributos topológicos controlam os tempos característicos de *folding* apresentados nas referidas Tabelas. E quanto a mostrar como a taxa de *folding* pode ser correlacionada com

diversos atributos topológicos das estruturas alvo (*folds* e suas composições cooperativas), mediante a proposta de um argumento plausível para as várias ordens de grandeza observadas nos tempos característicos de *folding* $\tau$, isso foi feito estudando as correlações entre $\log(1/\tau)$ e OC, e entre $\log(1/\tau)$ e $\Omega$, conforme ilustrado nas Figuras 5 e 6 do Capítulo 4. Esse estudo nos permitiu concluir que a Ordem de Contato não seria o parâmetro mais indicado para avaliar a topologia da nativa como um todo, embora para casos específicos de proteínas reais tal correlação seja bem mais acentuada do que a encontrada aqui. Porém no nosso caso, em que as estruturas não foram selecionadas por nenhum critério excludente, pois foram incluídos os mais diversos padrões estruturais, as correlações com $\Omega$ apresentam um aumento de aproximadamente 50%, o que é considerável.

Outro resultado importante e ilustrativo, também apresentado no Capítulo 4, diz respeito à distribuição dos valores de $q$, ilustrado por seis alvos com topologias distintas (Figuras 3A –3F, Capítulo 4). A determinação do valor máximo permitido para a variável $q$, $q_{max}$ = 1.33, foi realizada por meio de uma análise de varredura com diferentes e representativas estruturas. A referida análise constatou evidências de que este valor para $q_{max}$ se trata de um valor universal para o presente modelo. Isto é, para todos os casos analisados até agora, para uma mesma temperatura $T = 1$ (unidades arbitrárias), este é o valor que resulta no tempo característico de *folding* mínimo, ou seja, $\tau \rightarrow \tau_{min}$. Para qualquer outro valor significativamente maior ou menor do que $q_{max}$ =1,33, $\tau$ se afasta de $\tau_{min}$.

Se a variável $q$ é identificada com o parâmetro que controla a flutuação térmica local, então podemos afirmar que as estruturas de fácil acesso cinético não necessitam de valores consideráveis dessa flutuação, e, portanto os valores maiores de $q$ praticamente não serão visitados. Portanto para essas estruturas a primeira etapa do processo de *folding* é muito rápida, como é o caso do tempo característico de *folding* dos alvos 866 e 1128, que é da ordem de dezenas de passos Monte Carlo. Se assim for, é possível especular que a flutuação térmica local funcione, até certo ponto, com substituto da metáfora do funil [9].

Em relação ao cenário originado da combinação do modelo estereoquímico com a Mecânica Estatística Não Extensiva, aliado com a proposta do parâmetro $\Omega$, é possível imaginar os seguintes avanços: aumentar o tamanho da cadeia (número de aminoácidos); violar a regra para a construção da cadeia (o que equivale a fazer mutações na cadeia); usar

com critério outros pesos estatísticos mais especulativos para simular o modelo além do de Boltzmann ou o de Tsallis (por exemplo, a distribuição de Levy); considerar flutuações na densidade do solvente; migrar do espaço discreto para o contínuo; melhorar a precisão do parâmetro $\Omega$ e aplica-lo em proteínas reais; experimentar variações no alfabeto estereoqímico; e, introduzir interações que representem as ligações de hidrogênio e as forças de dispersão. Para que tudo isso possa ser realizado será necessário recursos computacionais distribuídos, ou seja, rodar os programas em paralelo, e, se possível desenvolver uma interface amigável para os programas utilizados. Parte disso está sendo realizado por Pablo Andrei Silva, aluno de mestrado trabalhando no Grupo de Física Biológica (Departamento de Física e Química – FCFRP – USP). Uma vez realizados esses desenvolvimentos e generalizações, uma das possibilidades imediatas será simular todo o catálogo das 51704 estruturas maximamente compactas, com diferentes pesos estatísticos, e diferentes temperaturas $T$ do reservatório (solvente), e assim investigar com detalhe o papel da combinação de composições cooperativas e não cooperativas para a cinética de um dado alvo, e esclarecer a pretensa universalidade do valor de $q_{\max}$. Com isso, talvez seja possível classificar os alvos em subgrupos de acordo com a cooperatividade estrutural e dos padrões estruturais presentes em cada um deles.

Seria natural alguém perguntar: O que se pretende com isso tudo? Bem, o que se espera de um modelo minimalista é que ele desempenhe o papel de uma caricatura, a mais fiel possível, de um sistema ou processo sob análise, e com isso obter *insights* sobre o problema real. Apesar de suas limitações, conforme os resultados apresentados nesta tese, o modelo estereoquímico se mostrou capaz de resolver o problema do *folding* inverso –dada um estrutura, prever a sua correspondente seqüência e encontra-la numa janela de tempo pré-estabelecida, por meio de um processo de simulação (MC), a partir de uma conformação aberta (sem a presença de qualquer contato do tipo primeiro vizinho, condição inicial). Adicionalmente, o modelo mostrou ser capaz de reproduzir quatro ordens de grandeza para a taxa de *folding*. Todos esses resultados podem de certa forma, ser reproduzidos com a estatística de Boltzmann, porém somente quando a temperatura do reservatório térmico é ajustada de forma adequada para cada estrutura alvo. Contudo essa abordagem para o presente problema é artificial, uma vez que o conjunto das proteínas de um ser vivo específico, ou de uma célula em particular, funciona apenas dentro de um regime estreito de temperaturas. Daí, com o peso de Tsallis a questão não é encontrar uma temperatura de *folding* para cada seqüência de aminoácidos, e sim perguntar: dada uma seqüência providenciada pela sintaxe

do modelo estereoquímico, quanto de flutuação térmica local é necessária (distribuição dos *q*'s) para que um conjunto de seqüências idênticas encontre a estrutura nativa no menor tempo característico possível? Esta pergunta é respondida automaticamente, deixando o índice entrópico *q* variar adequadamente.

Finalmente, com o conceito da cooperatividade estrutural, contido em Ω, a relevância do papel da topologia da estrutura nativa e conseqüentemente a facilidade com que os padrões que compõem a nativa são acessados, esclarece em termos físicos o *insight* dado pela ordem de contato; OC.

Talvez em breve, com auxílio do modelo estereoquímico, apesar de todas as aproximações contidas em sua abordagem, e dentro de certos limites, será possível vislumbrar como a seleção natural garantiu não apenas a estabilidade de uma particular estrutura nativa, mas também sua acessibilidade [29]. Pois, aqui foi mostrado que a combinação do código estereoquímico proposto, com o modo de incluir as flutuações térmicas locais no sistema cadeia protéica-solvente em simulações Monte Carlo, é capaz de reproduzir aspectos relevantes do processo de *folding* de proteínas.

# 6 - REFERÊNCIAS

[1] Branden C, Tooze J. Introduction to Protein Structure, Garland Science, 2nd edition 1999.

[2] Schulz GE, Schirmer RH. Principles of Protein Structure, Springer-Verlag 4th edition, 1996.

[3] Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(96):223–30.

[4] Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? Current Opinion in Structural Biology. 2007;17:342–46.

[5] Faísca PFN. O mistério da forma das proteínas. Gazeta da Física. 2006;29(3):34–9 .

[6] Echenique P. Introduction to protein folding for physicists. Contemporary Physics. 2007;48(2):81–108.

[7] Ecroyd H, Carver JA. Unraveling the mysteries of protein folding and misfolding. IUBMB Life. 2008;60(12):769–74.

[8] Binder K (Editor). Monte Carlo and Molecular Dynamics Simulations in Polymer Science. Oxford University Press. 1995.

[9] Onuchic JN, Wolynes PG. Theory of protein folding. Current Opinion in Structural Biology. 2004;14(1):70–75.

[10] Head-Gordon T, Brown S. Minimalist models for protein folding and design. Current Opinion in Structural Biology. 2003;13:160–7.

[11] Araújo AFP de, Protein folding cooperativity: basic insights from minimalist models. Protein and Peptide Letters. 2005;12(3):223–8.

[12] Galzitskaya OV, Garbuzynskiy SO. Entropy capacity determines protein folding. Proteins: Structure, Function, and Bioinformatics. 2006;63:144–54.

[13] Snow CD, Sorin EJ, Rhee YM, Pande VS. How well can simulation predict kinetics and thermodynamics of protein folding? Annu. Rev. Biophys. Biomol. Struct. 2005;34:43–69.

[14] Prigogine I, From being to becoming – time and complexity in the physical sciences. W. H. Freeman and Company. 1980; Pelesko J. Self Assembly: The Science of Things That Put Themselves Together. Chapman & Hall/CRC. 2007; Bak P, Paczuski M. Complexity, contingency, and criticality. Proc Natl Acad Sci USA. 1995;92(15):6689–96.

[15] Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys. 2008;(37): 289-316.

[16] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol. 1998;277:985–94.

[17] Gillespie B, Plaxco KW. Using protein folding rates to test protein folding theories. Annu Rev Biochem. 2004;73:837–59.

[18] Popov EM. Protein folding as a nonlinear nonequilibrium thermodynamic process. IUBMB Life. 1999;47(3):443–53.

[19] Koolman J, Roehm KH. Color Atlas of Biochemistry. Thieme Books. 2nd edition. 2005.

[20] Chothia C. One thousand families for the molecular biologist. Nature. 1992;357(18):543–44.

[21] Lesk AM, Rose GD. Folding units in globular proteins. Proc Natl Acad Sci. 1981;78(7): 4304–8.

[22] Rose GD. Protein folding and the Paracelsus challenge. Nature Structural Biology. 1997;4(7):512–4.

[23] Selkoe DJ. Folding proteins in fatal ways. Nature. 2003;426:900-4.

[24] Huang F, Nau WM. A Conformational Flexibility Scale for Amino Acids in Peptides. Angew Chem Int Ed. 2003;42:2269–72.

[25] Kamerzell TJ, Middaugh CR. The complex inter-relationships between protein flexibility and stability. Journal of Pharmaceutical Sciences. 2008;97(9):3494–517.

[26] Haber E, Anfinsen CB. Studies on the reduction and re-formation of protein disulfide bonds. J Biol Chem. 1961;236:1361-3.

[27] Lazaridis T, Karplus M. Thermodynamics of protein folding: a microscopic view. Biophys Chem. 2003;100:367–95.

[28] Baker D. Metastable states and folding free energy barriers. Nature Structural Biology. 1998;5:1021–4.

[29] Baker D, Agard DA. Kinetics versus Thermodynamics in Protein Folding. Biochemistry. 1994;33(24):7505–9.

[30] Em: http://www.rcsb.org/pdb/home/home.do

[31] Fersht AR. From the first protein structures to our current knowledge of protein folding: delights and skepticisms. Nature Molecular Cell Biology. 2008;9:650–4.

[32] Privalov PL. Capítulo três em Protein Folding, editado por Creighton TE. W.H. Freeman and Company. New York. 1992.

[33] Nölting B, Salimi N, Guth U. Protein folding forces. Journal of Theoretical Biology. 2008;251:331–47.

[34] Kumar S, Li MS. Biomolecules under mechanical force. Physics Reports. 2010;486:1–74.

[35] Jackson SE. How do small single-domain proteins fold? Folding & Design. 1998;3(4):81–91.

[36] Lattman EE, Rose G. Protein folding - what's the question? Proc Natl Acad Sci.1993;90:439–441.

[37] Pauling L, Corey RB. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc Natl Acad Sci. 1951;37(5):235-40.

[38] Bolen DW, Rose GD. Structure and energetics of the hydrogen-bonded backbone in protein folding. Annu Rev Biochem. 2008;77:339–62.

[39] Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem. 1959;14:1–63.

[40] Klotz IM, Franzen JS. Hydrogen bonds between model peptide groups in solution. J. Am Chem Soc. 1962;84:3461–6.

[41] Hill DJ, Mio MJ, Prince RB, Hughes T, Moore JS. A field guide to foldamers. Chem Rev. 2001;101:3893–4011.

[42] Nelson JC, Saven JG, Moore JS, Wolynes PG. Solvophobically driven folding of nonbiological oligomers. Science. 1997;277:1793-6.

[43] Richardson JM, Lopez MM, Makhatadze GI. Enthalpy of helix-coil transition: missing link in rationalizing the thermodynamics of helix-forming propensities of the amino acid residues. Proc Natl Acad Sci. 2005;102:1413–8.

[44] Shirley BA, Stanssens P, Hahn U, Pace CN. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. Biochemistry 1992;31:725–32.

[45] Myers JK, Pace CN. Hydrogen bonding stabilizes globular proteins. Biophys J. 1996;71: 2033–9.

[46] Dill KA, Polymer principles and protein folding. Protein Science 1999;8:1166–80.

[47] Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. Proc Natl Acad Sci. 2006;103:16623–33.

[48] Fox MA, Whitesell JK. Organic Chemistry, Jones & Bartlett Publishers; 3rd 2004.

[49] Auton M, Bolen DW. Predicting the energetics of osmolyte-induced protein folding/unfolding. Proc Natl Acad Sci. 2005;102:15065–68.

[50] Levinthal C. Are there pathways for protein folding? Journal de Chimie Physique et de Physico-Chimie Biologique, 1968;65:44.

[51] Harrison SC, Durbin R. Is there a single pathway for the folding of a polypeptide chain? Proc Natl Acad Sci., 1985;82:4028–30.

[52] Lazaridis T, Karplus M. "New View" of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations. Science. 1997;278: 1928–31.

[53] Chekmarev SF, Krivov SV, Karplus M. Folding of Ubiquitin: A Simple Model Describes the Strange Kinetics. J Phys Chem B. 2006;110:8865–9.

[54] Kim PS, Baldwin RL. Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. Annu Rev Biochem. 1982;51:459–89.

[55] Fersht A. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc Natl Acad Sci. 1995;92:10869–73.

[56] Nölting B. Protein Folding Kinetics - Biophysical Methods. Springer. 2nd edition. Capítulo 11. 2006.

[57] Udgaonkar JB. Multiple routes and structural heterogeneity in protein folding. Annu Rev Biophys. 2008;37:489–510.

[58] Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys. 1998;109:11101–8.

[59] Salvi G., De Los Rios P. Effective interactions cannot replace solvent effects in a lattice model of proteins. Phys Rev Let. 2003;91:258102-1.

[60] Rose GD, Wolfenden R. Hydrogen bonding, hydrophobicity, packing, and protein folding. Annu. Rev. Biophys. Biomol. Struct. 1993;22:381-415.

[61] da Silva RA, da Silva MAA, Caliri A. Deterministic folding: The role of entropic forces and steric specificities. J Chem Phys. 2001;114:4235–42.

[62] Rohit VP, Rose GD. A simple model for poly-proline II structure in unfolded

states of alanine-based peptides. Protein Science. 2002;11:2437–55.

[63] Head-Gordon T, Stillinger FH, Wright MH, Gay DM. Poly(L-alanine) as a universal reference material for understanding protein energies and structures. Proc Natl Acad Sci. 1992;89:11513–7.

[64] Tarragó MEP, Rocha LFO, da Silva RA, Caliri A. Steric constraints as folding coadjuvant. Phys Rev E. 2003;67:031901.

[65] Tanford C. The hydrophobic effect and the organization of living matter. Science. 1978;200(2):1012–8; Chandler D. Two faces of water. Nature. 2002;417:491.

[66] Zoghbi HY, Orr HT. Glutamine Repeats and Neurodegeneration. Annu. Rev. Neurosci., 2000;23:217–47.

[67] Silva IR, Caliri A. The protein folding problem. Current Methods in Medical Chemistry and Biological Physics. Taft CA, Silva HTPC (Editors). V.1, Research Signpost. India. 2007.

[68] Lumry R, Eyring H. Conformation Changes of Proteins. J Phys Chem. 1954;58(2):110–20.

[69] Dill KA. Dominant forces in protein folding. Biochemistry. 1990;29:7133–55.

[70] Silva IR, dos Reis LM, Caliri A. Topology-dependent protein folding rates analysed by a stereochemical model. J Chem Phys. 2005;123:154906.

[71] Tese de doutorado de Inês Regina da Silva. Enovelamento protéico: fatores topológicos (2005). Hospedada em: http://www.teses.usp.br/teses/disponiveis/59/59135/tde-24082005-112844/

[72] Hill TL. An Introduction to Statistical Thermodynamics. Dover Publications. New York. p. 225, 1986.

[73] Rocha LFO, Silva IR, Caliri A. Distinct conformational properties determined by implicit and explicit representation of protein–solvent interactions. An analytical and computer simulation study. Physica A. 2009;338:4097–104.

[74] Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 1963;7:95–9.

[75] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH. Equation of state calculation by fast computing machines. J Chem Phys. 1953;21(6):1087–92.

[76] Schwartz R. Biological Modeling and Simulation: A Survey of Practical Models, Algorithms, and Numerical Methods. The MIT Press. Cambridge, Massachusetts, United States 2008.

[77] Cooper A. Thermodynamic fluctuations in protein molecules. Proc. Natl. Acad. Sci. USA. 1976;73(8):2740–1.

[78] Shakhnovich E. Protein folding roller coaster, one molecule at a time. Proc Natl Acad Sci. 2009;106(29):11823–4.

[79] Hoi Sung Chung, John M. Louis, William A. Eaton, Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. Proc. Natl. Acad. Sci., v. 2009;106(29):11837–11844.

[80] Beck C. Non-additivity of Tsallis entropies and fluctuations of temperature. Europhys Lett. 2002;57(3):329–33; Beck C, Cohen EGD. Superstatistics. Physica A. 2003;322:267–75.

[81] Rajagopal AK, Pande CS, Abe S. Nanothermodynamics – A generic approach to material properties at nanoscale, Nano-Scale Materials: From Science to Technology 2006: 241–8. Disponível em: http://arxiv.org/ftp/cond-mat/papers/0403/0403738.pdf .

[82] Touchette H. Temperature fluctuations and mixtures of equilibrium states in canonical ensemble, Capítulo do livro: Nonextensive Entropy – Interdisciplinary Applications. Gell-Mann M, Tsallis C (Editores). Oxford University Press. Oxford. 2004.

[83] Tsallis C. Possible Generalization of Boltzmann-Gibbs Statistics. J Stat Phys. 1988;52:479–87.

[84] Rocha LFO, da Silva MAA, Caliri A. Entropic force and folding of macromolecules. Phys Letters A. 1996;220:178–82.

[85] Rocha LFO, Tarragó Pinto ME, Caliri A. The water factor in the protein-folding problem. Brazilian Journal of Physics. 2004;34(1):90–101.

[86] Tsallis C. Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World. Springer. Berlin. 2009.

[87] Tsallis C. Thermostatistically approaching living systems: Boltzmann–Gibbs or nonextensive statistical mechanics? Physics of Life Reviews. 2006;3:1–22.

[88] Micheletti C, Lattanzi G, Maritan A. Elastic properties of proteins: Insight on the folding process and evolutionary selection of native structures. J Mol Biol. 2002;321:909–21.

[89] de Gennes PG. Scaling concepts in polymer physics. Cornell University Press. Cornell. 1979.

[90] Beck C, Lewis GS, Swinney HS. Measuring non-extensitivity parameters in a turbulent Couette-Taylor flow. Physical Review E. 2001;63:035303.

[91] Dissertação de mestrado de João Paulo Dal Molin. Correções de cadeia finita para a teoria elástica clássica (2006). Hospedada em: http://www.dominiopublico.gov.br/pesquisa/PesquisaObraForm.do?select_action=&co_autor=9031

[92] Doi M, Edwards SF. The Theory of Polymer Dynamics. International Series of Monographs on Physics 73. Oxford Science Publications. Clarendon Press – Oxford. 1986.

[93] Kaya H, Chan HS. Contact order dependent protein folding rates: kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences. Proteins: Struct Funct Genet. 2003;52:524–33.

[94] Dal Molin JP, da Silva MAA, da Silva IR, Caliri A. Nonextensive statistical mechanics applied to protein folding problem: kinetics aspects. Brazilian Journal of Physics. 2009;39(2a):435–8.

[95] Thirumalai D, O'Brien EP, Morrison G, Hyeon C. Theoretical Perspectives on Protein Folding. Annu Rev Biophys. 2010;39:159–83.

[96] Beck C. Generalized statistical mechanics for superstatistical systems. Phil Trans R Soc A, 2011;369:453–65.

[97] http://tsallis.cat.cbpf.br/TEMUCO.pdf

[98] Tsallis C. Nonextensive Statistical Mechanics and Thermodynamics: Historical Background and Present Status. Em Nonextensive Statistical Mechanics and Its Applications, Abe S, Okamoto Y (Editores). Springer. Berlin. 2001.

[99] Tsallis C. Nonextensive statistical mechanics: construction and physical interpretation. Capítulo do livro Nonextensive Entropy – Interdisciplinary Applications. Gell-Mann M, Tsallis C (Editores). Oxford University Press. Oxford. 2004.

[100] Tsallis C. What should a statistical mechanics satisfy to reflect nature? Physica D. 2004;193:3–34.

[101] Tsallis C. Nonadditive entropy: The concept and its use. European Physical Journal A. 2009;40(3):257–66.

[102] Tsallis C. Nonadditive entropy and nonextensive statistical mechanics – an overview after 20 years. Brazilian Journal of Physics. 2009;39(2A):337–56.

[103] Lenzi EK, Anteneodo C, Borland L. Escape time in anomalous diffusive media. Phys. Rev. E. 2001;63:051109.

[104] Klafter J, Sokolov IM. Anomalous diffusion spreads its wings. Physics World. 2005;8:29–32.

[105] Plastino AR, Lima JAS. Equipartition and virial theorems within general thermostatistical formalisms. Phys Letters A. 1999;(260):46–56.

[106] Martínez S, Pennini F, Plastino A. Equipartition and virial theorems in a nonextensive optimal Lagrange multipliers scenario. Phys. Letters A. 2000;278:47–52.

[107] Tsekouras GA, Tsallis C. Generalized entropy arising from a distribution of $q$ indices. Physical Review E. 2005;71:046144.

[108] Curado EMF, Tsallis C. Generalized statistical mechanics: connection with thermodynamics. J. Phys. A, 1991;24:69–72. Correções. 1991;24:3187 e 1992;25;1019.

[109] Tsallis C, Mendes RS, Plastino AR. The role of constraints within generalized nonextensive statistics. Physica A, 1998;261:534–54.

[110] Sokolnikoff IS. Advanced Calculus. McGraw-Hill. USA. 1939.

# APÊNDICES

## APÊNDICE A – O desenvolvimento do modelo estereoquímico

Segue uma lista em ordem cronológica com o nome das pessoas que contribuíram para a construção e avanços do modelo estereoquímico:

i) – Lucila M. dos Reis. Dissertação de Mestrado: "*Folding* de Proteínas: Simulação Monte Carlo de um modelo em rede com potencial randômico". Mestrado em Física Aplicada à Medicina e Biologia – Faculdade de Filosofia Ciências e Letras de Ribeirão Preto – USP (1995). O principal resultado desse trabalho é a elaboração do catálogo com as estruturas maximamente compactadas em um cubo 3×3×3.

ii) – Luiz Faustino de Oliveira Rocha. Dissertação de Mestrado: "Estabilidade das Conformações de Macromoléculas". Mestrado em Física Aplicada à Medicina e Biologia – Faculdade de Filosofia Ciências e Letras de Ribeirão Preto – USP (1996). Nessa dissertação foi mostrado que o efeito hidrofóbico é capaz de compactar a cadeia eficientemente e conferir maleabilidade ao glóbulo protéico [84].

iii) – Roosevelt Alves da Silva. Dissertação de Mestrado: "Força Entrópica e *Folding* de Macromoléculas". Mestrado em Física Aplicada à Medicina e Biologia – Faculdade de Filosofia Ciências e Letras de Ribeirão Preto – USP (1997). Nessa dissertação o principal resultado foi mostrar que, um alfabeto de três letras é insuficiente para empacotar e garantir a estabilidade do glóbulo protéico de modo satisfatório. As principais implicações dessa dissertação serviram como ponto de partida para uma tese de doutorado realizada pelo mesmo aluno.

iv) – Roosevelt Alves da Silva. Tese de Doutorado: "*Folding* de proteínas: o papel das interações hidrofóbicas e estéricas". Doutorado em Física Aplicada à Medicina e Biologia – USP (2002). Nessa tese se encontra a gênese do alfabeto estereoquímico com 10 letras fundamentado no efeito hidrofóbico e nas especificidades estéricas dos aminoácidos. Uma sintaxe para o *design* da seqüência de aminoácidos, dado um alvo (*folding* inverso) também está presente nessa tese. Com isso um aspecto muito importante relacionado com o problema do *folding* foi satisfeito, que é o compromisso entre a estrutura alvo e a seqüência, e que, adicionalmente o alfabeto independe do alvo escolhido. Aspectos cinéticos, como

o tempo característico de enovelamento $\tau$, também foram investigados nessa tese [61].

*v)* – Maria Eulália Pinto Tarragó. Tese de doutorado: "*Folding* de macromoléculas: ordem de contato e estrutura geométrica da configuração nativa". Doutorado em Física Aplicada à Medicina e Biologia – USP (2003). Nessa tese o papel da topologia da estrutura nativa no contexto do modelo estereoquímico começou a ser investigado. Simulações computacionais relacionadas com a termodinâmica do processo, a atividade configuracional e a freqüência de visitação da nativa foram realizadas. Essas simulações permitiram concluir que o papel dos vínculos estereoquímicos é o de selecionar caminhos preferenciais para o processo de *folding*. Além disso, nessa tese foi possível propor que o processo de *folding* pode ocorrer em duas etapas desacopladas que se sucedem no tempo: (*i*) a etapa de busca, e, (*ii*) a consecução das condições de estabilidade global [64].

vi) – Inês Regina Silva. Tese de Doutorado: "Enovelamento protéico: fatores topológicos". Doutorado em Física Aplicada à Medicina e a Biologia – Faculdade de Filosofia Ciências e Letras de Ribeirão Preto – USP (2005). Aqui o papel da ordem de contato (OC) em descrever a topologia da estrutura nativa como um todo foi questionado. Foi proposto um meio de avaliar o conceito de cooperatividade: quando a estrutura nativa é rica em combinações de padrões estruturais de fácil acesso cinético, o processo de enovelamento é extremamente rápido, pois os contatos locais são naturalmente estimulados pelas flutuações térmicas do sistema cadeia-solvente [70, 71].

vii) Luiz Faustino de Oliveira Rocha, Tese de Doutorado: "Mecanismo de *folding* de macromoléculas: estudo sobre o papel das especificidades topológicas e químicas das interações intramoleculares em cadeias poliméricas lineares". Doutorado em Física Ibilce – Universidade Estadual Paulista Júlio de Mesquita Filho (2005). O resultado central deste trabalho é a demonstração de que, quando não há flutuação na densidade do solvente as interações do sistema cadeia-solvente podem ser bem representadas pelas interações de pares intra-cadeia [73].

viii) No sétimo Capítulo do livro [67] pode ser encontrado um resumo de tudo o que foi realizado nas dissertações e teses indicadas nos itens anteriores.

ix) Quanto a presente tese suas principais propostas de contribuições para o modelo estereoquímico são: (*i*) a alteração do critério de Metropolis, a fim de que a cadeia seja sensível às flutuações térmicas locais sem mudar a temperatura macroscópica

do banho térmico, o que é viável com o uso do peso estatístico de Tsallis, (*ii*) propor um modo de avaliar a cooperatividade das estruturas nativas através do parâmetro Ω, que em primeira aproximação nos informa a probabilidade de formação de um par de contatos nativo, e, assim investigar o papel da topologia da nativa *versus* log(1/τ), e, (*iii*) testar se as premissas propostas nos trabalhos anteriores são satisfatórias para descrever o processo de *folding*, que quanto a primeira etapa do processo de *folding,* isto foi realizado mediante milhares de simulações Monte Carlo.

## APÊNDICE B – Mecânica Estatística Não Extensiva

Originalmente a agora chamada Mecânica Estatística Não Extensiva foi proposta por Constantino Tsallis em 1988 [83]. Em quase duas décadas e meia essa proposta chamou, e, têm chamado a atenção de pesquisadores pertencentes as mais diversas áreas científicas. Uma bibliografia regularmente atualizada sobre os fundamentos e aplicações dessa proposta pode ser encontrada em [97]. A Mecânica Estatística Não Extensiva é destinada a ser uma generalização da Mecânica Estatística de Boltzmann-Gibbs, e ela parece ser conveniente para lidar com sistemas dinâmicos não lineares, que exibem caos fraco [87].

Artigos de revisão e Capítulos de livro sobre a Mecânica Estatística Não Extensiva podem ser encontrados em [86, 87, 98–102]. Recentemente foi publicado um livro dedicado ao assunto escrito pelo próprio autor da proposta [86]. No livro são encontrados tópicos como o pano de fundo histórico, e as motivações físicas para generalizar a forma entrópica de Boltzmann-Gibbs (BG); por conseguinte a Mecânica Estatística de BG também é objeto de análise. Em seguida, no livro, é apresentada uma investigação para as soluções de três equações diferenciais ordinárias progressivamente mais complicadas, onde a tentativa de encontrar uma solução generalizada para os casos analisados leva à obtenção da $q$-exponencial, que é obtida quando a linearidade é violada, sendo a solução da equação $dy/dx=y^q$ (y(0)=1; $q \in \Re$ ) . Os Capítulos seguintes são destinados à fundamentação da teoria, bem como aos casos em que ela tem sido aplicada. Quanto aos exemplos de aplicações da Mecânica Estatística Não Extensiva, um exemplo marcante é a generalização da lei de Arrhenius feita por Anteneodo e colaboradores [103], outro é a associação dela com processos difusivos anômalos como indicado por Klafter e Sokolov [104]. Ainda com relação às implicações da forma entrópica proposta por Tsallis, temos a generalização do teorema do virial e de eqüipartição da energia feita por Plastino, Plastino e colaboradores [105, 106].

Com relação a $q$-exponencial, ela apresenta como sua função inversa o $q$-logaritmo $\ln_q(x) \equiv (x^{1-q}-1)/(1-q)$ , o qual é usado para escrever a forma entrópica

$$S_q = k\frac{1-\sum_{i=1}^{W} p_i^q}{q-1}, \qquad q \in \Re . \tag{B-I}$$

Caso seja assumida a existência de um sistema composto por dois subsistemas independentes *A* e *B*. As probabilidades do sistema composto ficam dadas por $p_{ij}^{(A+B)} = p_i^A p_j^B, (i = 1,2,...,W_A\,;\, j = 1,2,...,W_B\,;W_{A+B} = W_A W_B)$. Daí, se $k$=1 é verificado que

$$S_q(A+B) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B) \qquad \text{(B-II)}$$

O terceiro termo na equação (B-II) sugere que a forma entrópica de Tsallis é não aditiva, e que pode haver algum tipo de conexão entre os subsistemas *A* e *B*. Há sistemas para os quais $q$>1, e, portanto a entropia do sistema composto é menor do que a soma das entropias dos subsistemas isolados, este caso corresponde à sub-extensividade. Por outro lado há sistemas nos quais $q$<1, situação esta chamada de super-extensividade, pois a entropia do sistema composto é maior do que a dos subsistemas quando isolados. Na presente tese, dada a argumentação proposta por Beck [80], e recentemente discutida em [96] o nosso foco foi direcionado para valores de $q$>1. Quanto ao nome da proposta de Beck-Cohen (a Super Estatística), ele fica por conta de que para alguns sistemas indicados em [80, 96], e nas referências contidas neles, ocorre a superposição de duas (ou mais) estatísticas. Pense em situações nas quais é possível que o sistema assuma diferentes valores para o índice entrópico $q$, como é o caso dos resultados das simulações mostradas nesta tese. Ou ainda em uma forma abstrata como proposto por Tsekouras e Tsallis em [107].

Quando o sistema em estudo é dito isolado a situação física é descrita no contexto do *ensamble* microcanônico, no qual o único vínculo existente é aquele sobre as probabilidades, sendo dado por

$$\sum_{i=1}^{W} p_i = 1. \qquad \text{(B-III)}$$

Desde que a probabilidade seja não nula, cada uma delas é igual a $1/W$. Por outro lado, se o *ensemble* canônico é adotado, é necessário incluir um vínculo apropriado para a energia. Há três propostas para o segundo vínculo, a original feita por Tsallis [83], a segunda feita por Curado e Tsallis [108], e a terceira feita por Tsallis, Mendes e Plastino [109]. Com a finalidade de obtenção direta da $q$-distribuição, aqui foi adotada a segunda proposta [108], a qual utiliza a técnica dos multiplicadores de Lagrange, como é feito na Mecânica Estatística usual (BG), isto é mostrado a seguir.

Se a forma entrópica $S_q$ é dada pela expressão (B-I), deseja-se obter uma expressão para a distribuição $P_q$ quando os seguintes vínculos são assumidos

$$\sum_{i=1}^{W} p_i = 1, \qquad U_q = \sum_{i=1}^{W} E_i p_i^q. \qquad \text{(B-IV)}$$

De acordo com os vínculos (B-IV), são acrescentados dois zeros na forma entrópica $S_q$, portanto

$$R_q(p_i) = \frac{1-\sum_{i=1}^{W} p_i^q}{q-1} + \lambda_1\left(1-\sum_{i=1}^{W} p_i\right) + \lambda_2\left(U_q - \sum_{i=1}^{W} E_i p_i^q\right). \quad \text{(B-V)}$$

Se a variação de $R_q$ é feita em relação a uma configuração $p_j$ qualquer

$$\frac{\delta R_q(p_i)}{\delta p_j} = \frac{q p_j^{q-1}}{q-1} + \lambda_1 + q\lambda_2 E_j p_j^{q-1} = 0\,, \qquad \text{(B-VI)}$$

$$q p_j^{q-1} + \lambda_1(q-1) + q\lambda_2(q-1)E_j p_j^{q-1} = 0\,,$$

$$p_i = C_q[1-(1-q)\lambda_2 E_i]^{1/(1-q)}\,. \qquad \text{(B-VII)}$$

Do vínculo de normalização das probabilidades, e por inspeção $\lambda_2 \to \beta_0 = \dfrac{1}{k_B T}$, segue

$$C_q = \frac{1}{\sum_{i=1}^{W}[1-(1-q)\beta_0 E_i]^{1/(1-q)}}\,. \qquad \text{(B-VIII)}$$

Daí,

$$P_i = \frac{1}{Z_q}[1-(1-q)\beta_0 E_i]^{1/(1-q)}, \qquad \text{(B-IX)}$$

onde $Z_q$ é a função de partição generalizada, e $P_i$ é a distribuição $q$-exponencial no contexto do ensamble canônico. Por completeza as expressões para a energia interna média e a energia livre generalizada ($F_q = U_q - TS_q$) também são dadas

$$U_q = -\frac{\partial}{\partial\beta}\left(\frac{Z_q^{(1-q)}-1}{1-q}\right), \qquad F_q = -\frac{1}{\beta}\left(\frac{Z_q^{(1-q)}-1}{1-q}\right). \qquad \text{(B-X)}$$

Um comentário final deve ser feito. Quando $\lambda_2$ é identificado com $\beta_0$, isto significa, como foi utilizado em nossas simulações MC, que a temperatura do reservatório térmico é sempre mantida constante. Por outro lado, quando o valor para o índice entrópico $q$ é considerado fixo ou deixado variar , isso permite realizar um conjunto de simulações, como aquelas descritas nos Capítulos 3 e 4, onde a temperatura flutua localmente. Para testar essa interpretação, o índice entrópico $q$ pode ser deixado variar ao longo de cada uma das simulações, isto é feito em função do raio de giração da cadeia, configuração após

configuração. Daí, a freqüência de visitação dos valores de $q$ é registrada, e calcula-se um valor $\langle q \rangle$, que depois é comparado com o valor fixo do índice entrópico $q$ que fornece o menor tempo característico para o mesmo alvo, dada uma temperatura fixa $T$ do reservatório. Com isso é mostrado que os valores de $q$, tanto como parâmetro quanto variável dinâmica, são muito próximos quando o menor tempo característico de *folding* ($\tau_{min}$) é adotado como parâmetro analítico para analisar o efeito do peso de Tsallis na classe de sistemas heterogêneos abordada.

## APÊNDICE C – O tempo médio e o tempo característico de folding

Em Resultados e Discussões (Capítulo 4), a taxa de *folding*

$$\mathrm{T}(t) = \left| \frac{dN(t)}{dt} \right| \qquad \text{(C-I)}$$

foi discutida. Na expressão (C-I), tem-se que $N(t)=N_0e^{-\lambda t}$, e, que $\lambda=1/\tau$, onde $N_0$ é o número total de proteínas no estado desnaturado no instante $t$=0. No Capítulo 4, quando o índice entrópico $q$ é considerado uma variável dinâmica, foi mostrado que os valores do tempo médio de *folding* $\langle t \rangle$, e, $\tau$ podem ser bem próximos. Como o tempo característico de *folding* é obtido do ajuste de uma ou mais funções exponenciais sobre os dados do histograma, para um determinado conjunto de tempos de primeira passagem, aqui são mostradas as deduções de dois casos em que $\tau$ e $\langle t \rangle$ assumem valores próximos. O primeiro deles quando $\tau$ é obtido com o ajuste de uma única exponencial, e depois com duas exponenciais.

Para o ajuste com uma única exponencial, da expressão (C-I) segue que

$$\mathrm{T}(t) = \lambda N_0 e^{-\lambda t}. \qquad \text{(C-II)}$$

Se $\langle t \rangle$ é adotado como

$$\langle t \rangle = \frac{1}{N_0} \int_0^\infty T(t) t dt. \qquad \text{(C-III)}$$

Então segue que

$$\langle t \rangle = \lambda \int_0^\infty t e^{-\lambda t} dt = \lambda \left( \frac{-\partial}{\partial \lambda} \int_0^\infty e^{-\lambda t} dt \right)$$

$$= \lambda \left( -\frac{\partial}{\partial \lambda} \int_0^\infty e^{-\lambda t} dt \right)$$

$$= \lambda \left[ -\frac{\partial}{\partial \lambda} \left( \frac{e^{-\lambda t}}{-\lambda} \right)_0^\infty \right]$$

$$= \lambda \frac{\partial}{\partial \lambda} \left( \frac{1}{-\lambda} \right) = \lambda \frac{\partial}{\partial \lambda} (-\lambda^{-1})$$

$$= \lambda (\lambda^{-2}) = \frac{1}{\lambda} = \tau$$

$$\langle t \rangle = \tau.$$

Aqui a diferenciação sob o sinal da integral foi utilizada [110]. Quando o tempo característico é obtido com o ajuste de duas funções exponenciais o tempo médio fica

$$\langle t \rangle = \frac{1}{N_0}\left( \int_0^\infty \lambda_1 N_1 e^{-\lambda_1 t} dt + \int_0^\infty \lambda_2 N_2 e^{-\lambda_2 t} dt \right)$$

$$= \frac{1}{N_0}\left\{ \lambda_1 N_1 \left[ -\frac{\partial}{\partial \lambda_1}\left( \int_0^\infty e^{-\lambda_1 t} dt \right) \right] + \lambda_2 N_2 \left[ -\frac{\partial}{\partial \lambda_2}\left( \int_0^\infty e^{-\lambda_2 t} dt \right) \right] \right\}$$

$$\langle t \rangle = \frac{1}{N_0}\left( N_1 \tau_1 + N_2 \tau_2 \right).$$

Os gráficos com as verificações dos dois casos para estruturas alvos distintas são dados no Apêndice D seguinte.

## APÊNDICE D – Algumas estruturas alvo, seus mapas de contato e curvas cinéticas

Aqui são apresentadas as representações de cinco alvos distintos em suas formas 3D, juntamente com seus mapas de contato. Para cada alvo a curva cinética obtida de um conjunto de simulações com $T$=1 e $q$=1.1 fixos, e, $T$=1 com $q$ variando ao longo das simulações também é dada. A primeira das estruturas é o alvo 00866.

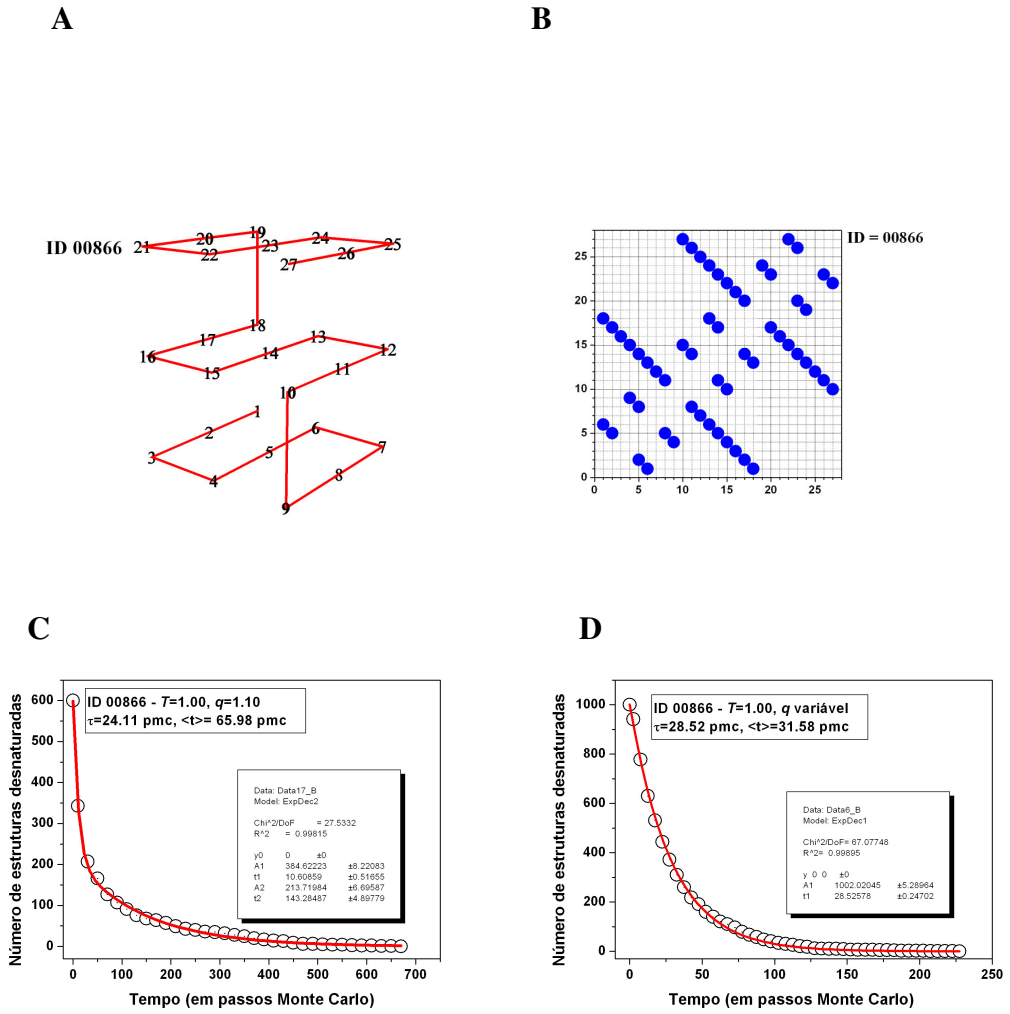


**Figura D-1**: No canto superior esquerdo (A) se encontra a representação tridimensional da estrutura 866, a sua direita em B, está o seu mapa de contato. No plano inferior, em C é dada a curva cinética para um conjunto de seiscentas simulações MC com $T$=1 e $q$=1.1, perceba que o ajuste sobre os dados do histograma foi obtido usando duas exponenciais. Entretanto, no canto inferior direito (D) quando $q$ é deixado variar ao longo da simulação em função do raio de giração da molécula o ajuste sobre os dados do histograma é feito com uma única exponencial, então $\tau$ e <t> assumem valores bem próximos.

A estrutura 866 tem sido adotada com um dos principais protótipos tanto na construção do modelo estereoquímico, quanto na checagem das simulações MC para as duas etapas do processo de *folding*: (*i*) a busca pela estrutura nativa; e (*ii*) a estabilidade da nativa [61,64]. A principal motivação para adotar aquele alvo fica por conta da proximidade ao longo da cadeia entre os resíduos que realizam os contatos nativos. O valor ótimo do índice entrópico $q$ que minimiza $\tau$ para o alvo 866 ($q$ = 1.10) é usado como referência para as outras quatro estruturas apresentadas aqui. Segue o alvo 1128.

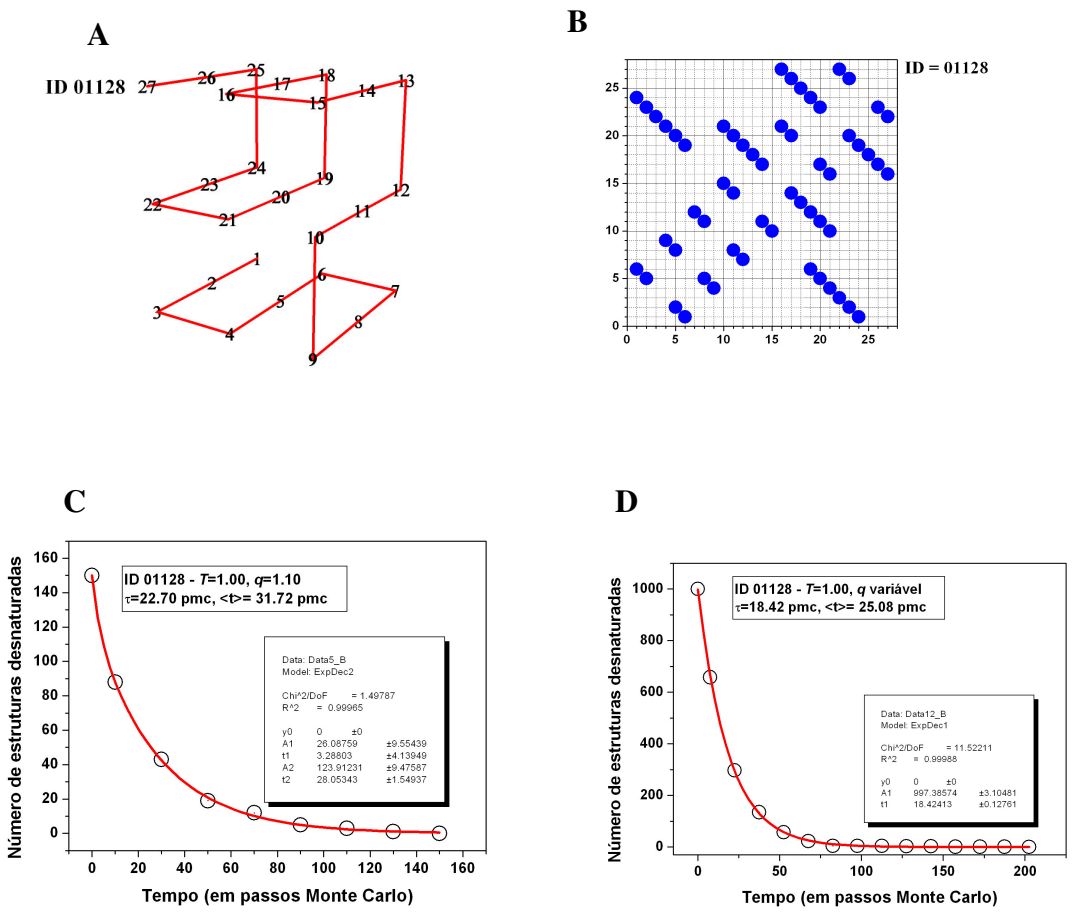


**Figura D-2**: No canto superior esquerdo (A) se encontra a representação tridimensional da estrutura 1128, a sua direita em B, está o seu mapa de contato. No plano inferior, em C é dada a curva cinética para um conjunto de cento e cinqüenta simulações MC com $T$=1 e $q$=1.1, perceba que o ajuste sobre os dados do histograma foi obtido usando duas exponenciais. Entretanto, no canto inferior direito (D) quando $q$ é deixado variar ao longo da simulação, da mesma forma como apresentado para o alvo 866 o ajuste sobre os dados do histograma é feito com uma única exponencial, então $\tau$ e <t> assumem valores bem próximos.

Segue o alvo 36335.

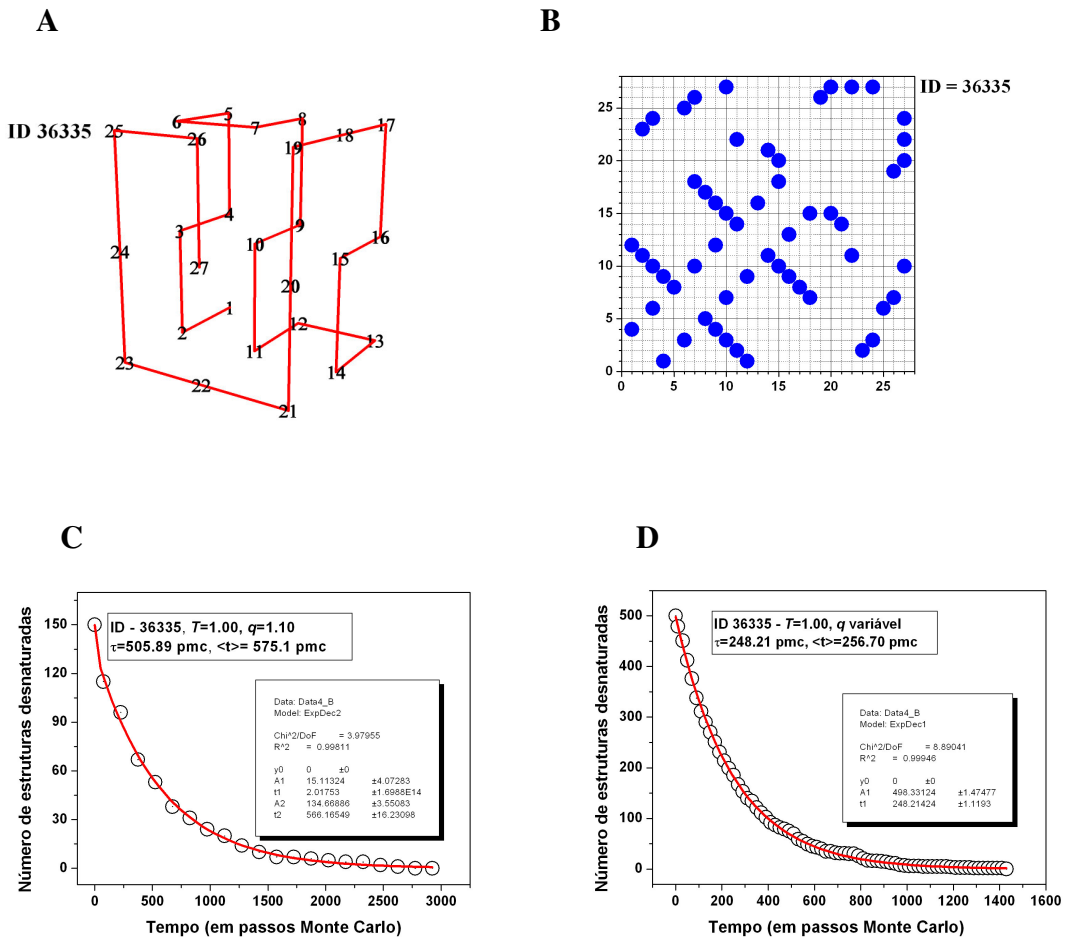


**Figura D-3**: No canto superior esquerdo (A) se encontra a representação tridimensional da estrutura 36335, a sua direita em B, está o seu mapa de contato. No plano inferior, em C é dada a curva cinética para um conjunto de cento e cinqüenta simulações MC com $T$=1 e $q$=1.1, perceba que o ajuste sobre os dados do histograma foi obtido usando duas exponenciais. Entretanto, no canto inferior direito (D) quando $q$ é deixado variar ao longo da simulação, e, outra vez como apresentado nos dois alvos anteriores o ajuste sobre os dados do histograma é feito com uma única exponencial, então $\tau$ e <t> assumem valores bem próximos.

Seguem agora duas estruturas mais complicadas, os alvos 51201 e 51476.

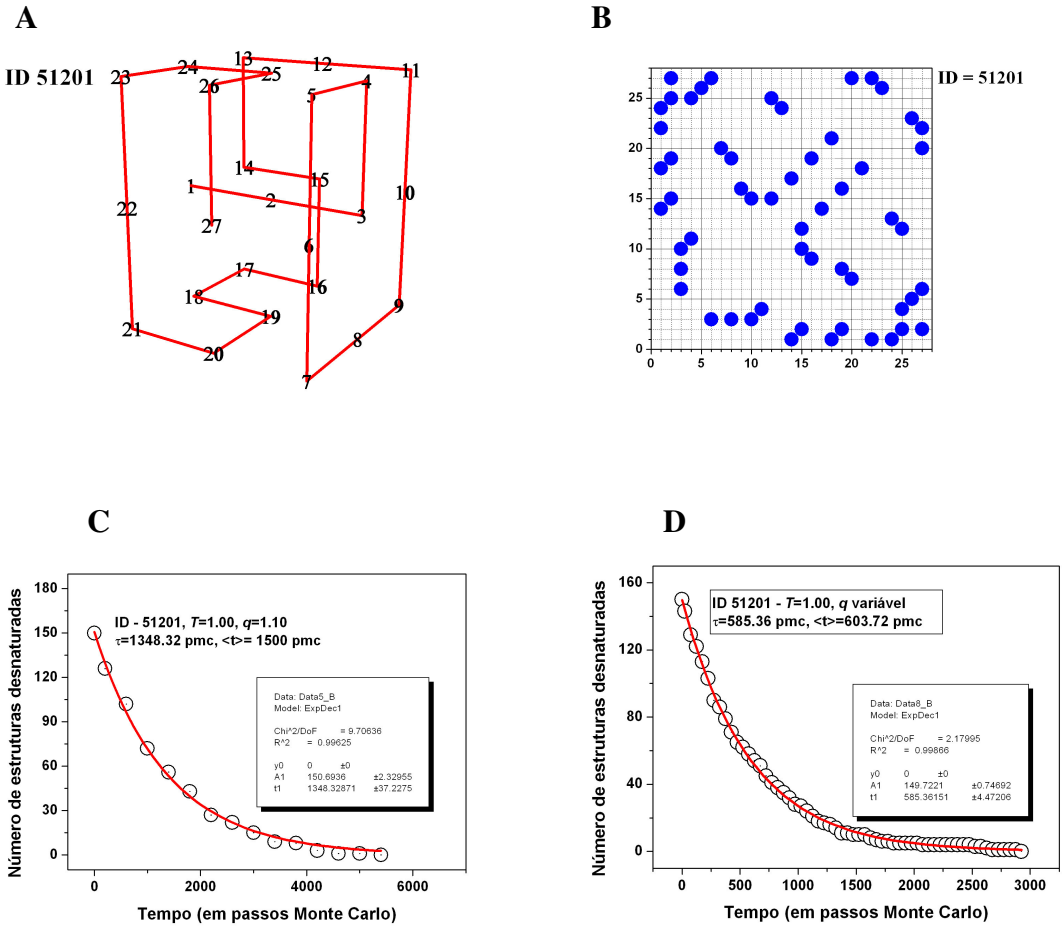


**Figura D-4**: No canto superior esquerdo (A) se encontra a representação tridimensional da estrutura 51201, a sua direita em B, está o seu mapa de contato. No plano inferior, em C é dada a curva cinética para um conjunto de cento e cinqüenta simulações MC com $T$=1 e $q$=1.1, perceba que o ajuste sobre os dados do histograma foi obtido usando uma única exponencial. Contudo, no canto inferior direito (D) quando $q$ é deixado variar ao longo da simulação o ajuste sobre os dados do histograma é feito com uma única exponencial, e novamente $\tau$ e <t> assumem valores bem próximos.

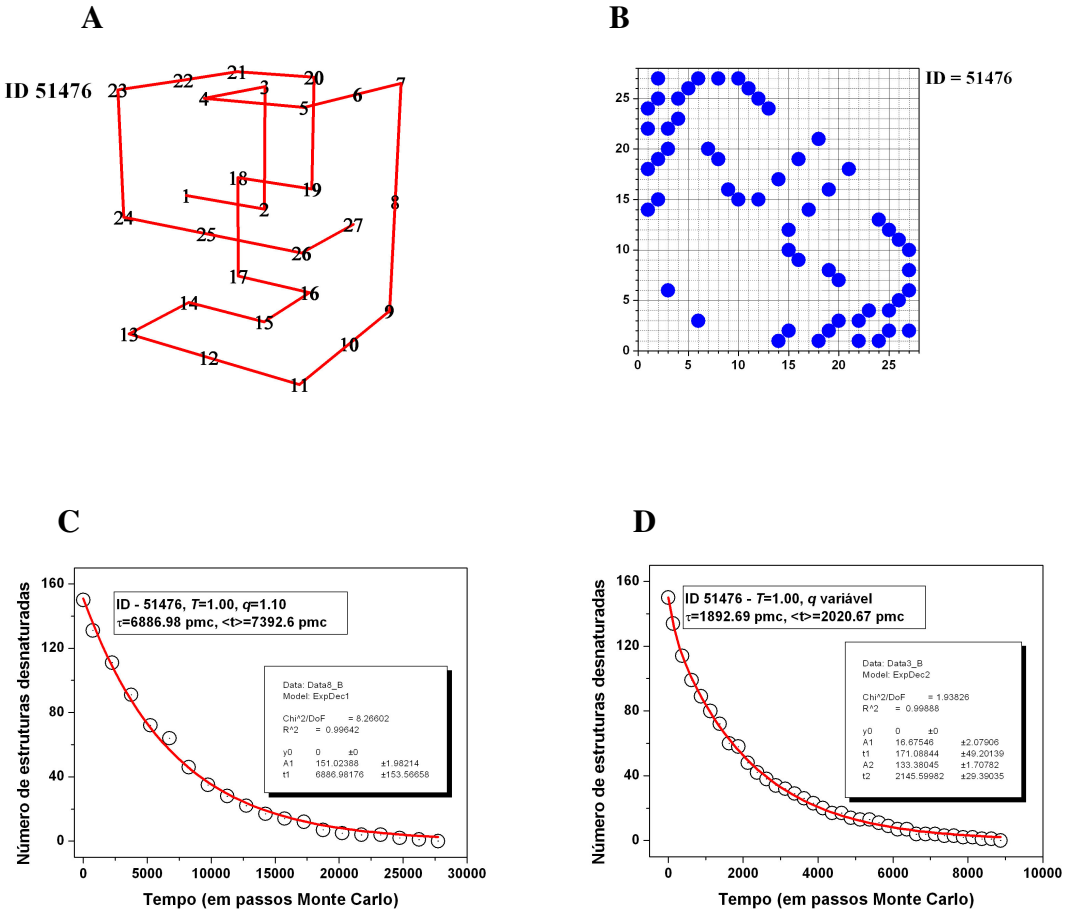


**Figura D-5**: No canto superior esquerdo (A) se encontra a representação tridimensional da estrutura 51476, a sua direita em B, está o seu mapa de contato. No plano inferior, em C é dada a curva cinética para um conjunto de cento e cinqüenta simulações MC com $T$=1 e $q$=1.1, perceba que o ajuste sobre os dados do histograma foi obtido usando uma única exponencial. Contudo, no canto inferior direito (D) quando $q$ é deixado variar ao longo da simulação, desta vez o ajuste sobre os dados do histograma é feito com duas exponenciais, e apesar da complexidade do alvo, novamente $\tau$ e <t> assumem valores bem próximos.