

Universidade de São Paulo

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

Departamento de Física e Matemática

Pós-graduação em Física aplicada à Medicina e Biologia

Inês Regina Silva

Enovelamento protéico: fatores topológicos.

v.1

Ribeirão Preto

2005

Universidade de São Paulo  
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto  
Departamento de Física e Matemática  
Pós-graduação em Física aplicada à Medicina e Biologia

Inês Regina Silva

Enovelamento protéico: fatores topológicos.

Tese apresentada à Faculdade de  
Filosofia, Ciências e Letras de Ribeirão  
Preto, da Universidade de São Paulo  
para obtenção do título de Doutor em  
Ciências.

Área de Concentração: Física aplicada à  
Medicina e Biologia

Orientador: Prof. Dr. Antonio Caliri

v.1

Ribeirão Preto

2005

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.

Catálogo-na-Publicação

Serviço de Documentação Bibliográfica

Faculdade de Filosofia, Ciências e Letras da Universidade de São Paulo

SILVA, Inês Regina

Enovelamento protéico: fatores topológicos. Ribeirão Preto, 2005.

75 f.: il.; 30 cm.

Tese de Doutorado, apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/USP – Área de concentração: Física aplicada à Medicina e Biologia.

Orientador: Antonio Caliri.

1. Enovelamento de proteína. 2. Modelo em rede. 3. Restrições estéricas. 4. Caracterização topológica.

## **DEDICATÓRIA**

Ao meu pai, Hortêncio Silva, à minha mãe, Maria Conceição de Martin Silva, pela minha vida. Com amor, admiração e gratidão por toda paciência que tiveram comigo, por sua compreensão, carinho, presença e incansável apoio ao longo do período de elaboração deste trabalho.

Às minhas irmãs, Rosa Maria Silva Donato e Marli Cristina Silva Vidotti pela compreensão nos momentos estressantes e pelo apoio dado nos momentos difíceis da minha vida.

Aos meus cunhados, Antonio Aparecido Vidotti e Gilberto Donato Junior, pela presença nos momentos críticos.

Ao meu sobrinho Daniel Henrique Silva Donato por me proporcionar momentos de agradáveis brincadeiras e risos e à minha sobrinha Amaitê que está por chegar.

A todos os meus amigos que contribuíram direta ou indiretamente para que eu desenvolvesse este trabalho de doutorado.

## AGRADECIMENTOS

Ao Prof. Dr. Antonio Caliri que, nos anos de convivência, muito me ensinou, contribuindo para meu crescimento científico e intelectual. Pela orientação e completo envolvimento com este projeto de pesquisa.

Ao Prof. Dr. José Drugowich de Felício e Prof. Dr. Nelson Augusto Alves, pela atenção inicial durante o período de definição da linha de pesquisa e de minha orientação.

À Universidade de São Paulo pela oportunidade de realização do meu curso de doutorado.

Ao Departamento de Física e Química da Faculdade de Ciências Farmacêuticas de Ribeirão Preto, por disponibilizar os computadores utilizados neste trabalho e uma sala de estudos.

A todas as pessoas que participaram direta ou indiretamente do desenvolvimento deste trabalho de doutorado.

“Se um homem tem um talento e não tem capacidade de usá-lo, ele fracassou.

Se ele tem um talento e usa somente a metade deste, ele fracassou parcialmente.

Se ele tem um talento e de certa forma aprende a usá-lo em sua totalidade, ele triunfou gloriosamente e obteve uma satisfação e um triunfo que poucos homens conhecerão”.

Thomas Wolfe

## RESUMO

SILVA, I. R. **Enovelamento protéico: fatores topológicos**. 2005. 75 p. Tese (Doutorado). Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, Ribeirão Preto, 2005.

O entendimento dos princípios básicos do enovelamento protéico pode conduzir a muitas aplicações importantes. Embora não se conheçam todos os aspectos significativos envolvidos neste problema, experimentos e aproximações teóricas têm produzido avanços relevantes na sua compreensão. Um fato experimental importante tem sido a descoberta de que o logaritmo da taxa de enovelamento  $\log k_f$  se correlaciona linearmente com parâmetros estruturais globais, como a ordem de contato relativa  $\chi$ . Com o propósito de contribuir para o entendimento do processo de enovelamento, o objetivo primordial deste trabalho consiste em explicar o porquê de certas proteínas não seguirem o comportamento linear entre  $\log k_f$  e  $\chi$ , verificado para outras proteínas da mesma classe (usualmente proteínas pequenas e com termodinâmica descrita pela aproximação de dois estados). Para isso foi necessário identificar os parâmetros topológicos da estrutura nativa que constituíssem importantes determinantes da cinética do enovelamento de proteínas globulares. Também se estudou como as especificidades estéricas dos aminoácidos afetam o processo do enovelamento de proteínas, assim como influenciam na correlação entre a ordem de contato relativo e a taxa de enovelamento.

Empregou-se neste estudo um modelo simplificado em rede cúbica, que foi tratado por meio de simulações Monte Carlo. Um conjunto de 52 estruturas maximamente compactas, correspondendo a cadeias de tamanho  $L = 27$  monômeros, foi usado para representar estados nativos; estas estruturas foram escolhidas de forma a representar uma variedade significativa de padrões estruturais, independentemente de  $\chi$ . Através de uma análise detalhada da influência de parâmetros topológicos das configurações nativas na cinética do enovelamento, conclui-se que a taxa de enovelamento é fortemente dependente daquilo que denominamos aqui como “conteúdo de estruturas tipo-secundárias” da estrutura nativa. Adicionalmente, observou-se que aquela (taxa), independentemente do valor da ordem de contato relativo, é fortemente influenciada pelos padrões configuracionais e suas combinações presentes na nativa.

Por meio dessa premissa, foi então possível explicar de forma consistente os casos que não obedecem a pretensa relação linear entre  $\log k_f$  e  $\chi$ , levando a concluir que o logaritmo da taxa de enovelamento e a ordem de contato relativo são linearmente dependentes somente para aquelas configurações em que há uma certa quantidade equilibrada (que depende de  $\chi$ ) de padrões estruturais, mesclando contatos efetivos de curto alcance (alto conteúdo de estruturas tipo-secundárias), com outros de longo alcance (baixo conteúdo de estruturas tipo-secundárias). Estruturas nativas que quebram este equilíbrio têm sua cinética de enovelamento afetada com respeito à reta de regressão linear ajustada para o conjunto de todas as configurações consideradas. Dessa forma, verificou-se que o mecanismo físico básico que relaciona o conteúdo de estruturas tipo-secundárias e a taxa de enovelamento, envolve o conceito de cooperatividade: se a estrutura nativa é rica em combinações de padrões estruturais ricos em contatos efetivos de curto alcance, o processo de enovelamento é mais rápido porque contatos locais são naturalmente estimulados por flutuações térmicas.



**ABSTRACT**

SILVA, I. R. **Protein folding: topological determinants**. 2005. 75 p. Thesis (Doctoral). Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, Ribeirão Preto, 2005.

The understanding of basic principles of the protein folding problem can lead to many important applications. Although not all the involved significant aspects of this problem are known, experiments and theoretical approaches have produced important advances in its understanding. An important experimental fact has been the discovery that the logarithm of the folding rate  $\log k_f$  correlates linearly with global structural parameters, like the relative contact order  $\chi$ . In order to contribute for the understanding of folding process, the primordial goal of this work consists in to explain why certain proteins do not follow the linear behavior between  $\log k_f$  and  $\chi$ , as verified to other proteins from the same class (usually small two states proteins). For this, it was necessary to identify those topological parameters of the native structure that are important to the folding kinetic of globular protein. It was also studied how steric specificities of the aminoacids affect the protein folding process, as well how they influence the correlation between the relative contact order and the folding rate.

It was employed in this study a simplified cubic lattice model, treated by Monte Carlo simulation. A set of 52 maximum compact structures, corresponding to chains of size  $L = 27$  monomers, was used to represent the native states; these structures were chosen in such a way to represent a significant diversity of structural patterns, independently of  $\chi$ . Through a detailed analysis of the influence of topological parameters of the native configurations on the folding kinetic, it was concluded that the folding rate is strongly dependent of what we call here as “content of type-secondary” of the native. Additionally, it was observed that  $\log k_f$  is, independently of  $\chi$ , strongly influenced by the configurational patterns and its combinations in the native.

Through this premise it was possible to consistently explain the cases that do not obey the pretense linear relation between  $\log k_f$  and  $\chi$ , leading to conclude that the logarithm of the folding rate and the relative contact order are linearly related only for those configurations in that there is a certain balanced amount of structural patterns (which depend on  $\chi$ ) mixing short-range effective contacts (high contents of secondary-

type structures) and long-range contacts (low contents of secondary-type structures). Structures that break this balance have its folding kinetic affected with respect to the linear fitting adjusted for the set of all the considered configurations. Of this form, it was verified that basic physical mechanism that relates the content of type-secondary structures and the folding rate involves the cooperativity concept: if the native structure presents combinations of structural standards rich in effective contacts of short-range, the folding process is faster because local contacts are naturally stimulated by thermal fluctuations.

## LISTA DE FIGURAS

- Figura 1.1. Os 20 aminoácidos naturais da proteína: (1) Glicina, (2) Alanina, (3) Valina, (4) Leucina, (5) Isoleucina, (6) Serina, (7) Treonina, (8) Cisteína, (9) Metionina, (10) Prolina, (11) Aspártico, (12) Asparagina, (13) Glutâmico, (14) Glutamina, (15) Arginina, (16) Lisina, (17) Histidina, (18) Fenilalanina, (19) Tirosina, (20) Triptofano . . . . . .03
- Figura 2.1. Ligação peptídica (retângulos tracejados). Associação entre o carbono da carboxila e o nitrogênio do grupo amino, com eliminação de água. A seta à direita indica a orientação convencional da cadeia . . . . . .08
- Figura 2.2. Ligação peptídica. Ângulos de rotação phi  $\Phi$  entre o nitrogênio do plano peptídico da esquerda e o carbono alfa, e o psi  $\Psi$  entre o carbono alfa e o carbono do plano peptídico seguinte . . . . . .08
- Figura 2.3. Os quatro níveis estruturais das proteínas . . . . . .09
- Figura 2.4. Proteína real 1BY0 (*wild-type plastocyanin from silene*) que possui sua cadeia de aminoácidos na conformação de alfa hélice. Pares de aminoácidos contactantes ( $i$  e  $i+4$ ) estão separados por poucas unidades ao longo da cadeia . .10
- Figura 2.5. Proteínas naturais. Na seqüência da esquerda para a direita: 1LMB ( *$\lambda$  repressor-operator complex*) formada somente por alfa hélices; 1CSP (*universal nucleic acid-binding domain*) formada somente por folhas-beta; 1HRC (*horse heart cytochrome c*) da forma mista. A união entre alfa hélices e folhas beta é feita por meio de *loops* ou *turns* . . . . . .11
- Figura 2.6. Proteína 1L2Y considerada atualmente a mais rápida para enovelar-se. Possui apenas 20 aminoácidos, é da forma mista (alfa hélice e folha beta) e possui tempo de enovelamento da ordem de quatro milionésimos de segundos . . .12
- Figura 2.7. Movimentos utilizados para a simulação do enovelamento da proteína: as linhas pontilhadas indicam as possíveis posições para mudanças configuracionais . . . . . .20
- Figura 3.1. Modelos minimalistas. (a) modelo estrutural, (b) colapso hidrofóbico, (c) mecanismo de nucleação-condensação. . . . . .22
- Figura 3.2. Atributos topológicos básicos dos monômeros, a saber,  $S$ ,  $T$  e  $E$ , numa configuração CSA . . . . . .24

Figura 3.3. Ilustração dos grupos e subgrupos hidrofóbicos aplicados numa CSA específica. . . . .	.26
Figura 3.4. Cubo 3×3×3 com seus sítios (27) identificados por inteiros consecutivos . . . . .	.29
Figura 4.1. Os círculos abertos mostram os sete casos dos finais da cadeia. Os padrões estruturais [*STTS*], [*TTTT*], [*TT*], [*STS*], <i>Snail</i> e <i>Staple</i> estão enfatizados . . . . .	.41
Figura 4.2. Estruturas tridimensionais das proteínas reais 1DMX e 1ZNC, mostrando a ocorrência de pseudo-nós . . . . .	.43
Figura 4.3. Mapa de contato da estrutura número 10.448 mostrando na parte superior esquerda 11 monômeros compondo a diagonal secundária (pontos incluídos pelas linhas pontilhadas) e na parte inferior direita, 10 monômeros compondo a diagonal primária (pontos sequenciais incluídos pelas linhas pontilhadas). . . . .	.47
Figura 5.1. Mapa de contatos inter-resíduos (primeiros vizinhos na rede) das estruturas mais rápidas: baixa ordem de contato ( $L\chi$ ) e alto conteúdo de estruturas tipo-secundárias ( $H\sigma$ ), facilmente identificado através da ocorrência persistente de linhas ao longo (e próximo) da diagonal principal e linhas paralelas à diagonal secundária, lembrando $\alpha$ hélices e folhas $\beta$ . Ver também Tabela 5.1 . . . . .	.55
Figura 5.2. Estrutura 17 mostrando uma seqüência ininterrupta de quinze Ts, que pode atuar como facilitadora no processo de enovelamento. . . . .	.56
Figura 5.3. Mapa de contatos inter-resíduos (primeiros vizinhos na rede) das estruturas com valor de ordem de contato intermediária ( $I\chi$ ) e alto conteúdo de estruturas tipo-secundárias ( $H\sigma$ ). Ver também Tabela 5.1 . . . . .	.57
Figura 5.4. Mapa de contatos inter-resíduos (primeiros vizinhos na rede) das estruturas com valor de ordem de contato baixa ( $L\chi$ ) e baixo conteúdo de estruturas tipo-secundárias ( $L\sigma$ ). Ver também Tabela 5.1 . . . . .	.58
Figura 5.5. Mapa de contatos inter-resíduos (primeiros vizinhos na rede) das estruturas com valor de ordem de contato intermediária e alta ( $I$ e $H\chi$ ) e baixo conteúdo de estruturas tipo-secundárias ( $L\sigma$ ). Ver também Tabela 5.1. Note a baixa densidade de pontos ao longo da diagonal principal. . . . .	.59

Figura 5.6. Mapa de contatos inter-resíduos (primeiros vizinhos na rede) das estruturas pertencentes ao grupo com padrões *Snail* e/ou *Staple* (O) e estruturas que nunca enovelaram (★). Ver também Tabela 5.1 . . . . .59

Figura 5.7. Estruturas números 1 e 7 da Tabela 5.1. Possuem ordem de contato próximas,  $\chi = 0,2381$  e  $\chi = 0,24603$ , mas taxa de enovelamento alterada em quase três ordens de magnitude, devido à transformação de dois padrões [\*STTS\*] consecutivos, estrutura 1, em um padrão *Snail*, estrutura 7. . . . .60

Figura 5.8. Mapa de contatos inter-resíduos (primeiros vizinhos na rede) das estruturas pertencentes ao grupo que apresenta uma mistura balanceada de contatos topológicos (curto e longo alcance) de acordo com a ordem de contato  $\chi$ . Observar que o conteúdo relativo de estruturas tipo-secundárias (número e comprimento das linhas pontilhadas paralelas à diagonal principal e paralela) diminui com o aumento progressivo de  $\chi$  mas sempre estão presentes. Ver também Tabela 5.1. . . . .61



## LISTA DE GRÁFICOS

Gráfico 3.1. Distribuição dos 97 valores de ordem de contato entre as 51.704 estruturas analisadas. Há grande concentração de estruturas ao redor de $\chi \approx 0,31$ e poucas estruturas nas caudas . . . . .	.32
Gráfico 4.1. Total depositado no ano e total acumulado de estruturas conhecidas de moléculas biológicas no PDB até 31/05/2005. . . . .	.38
Gráfico 4.2. Distribuição do número $N(n_T)$ de configurações com $n_T$ unidades do tipo $T$ . . . . .	.39
Gráfico 4.3. Distribuição do padrão [*STTS*] em função dos valores da ordem de contato relativa das estruturas . . . . .	.42
Gráfico 4.4. Distribuição da seqüência $\{T...T\}$ para as 51704 configurações CSA.	43
Gráfico 4.5. Distribuição do número $N(n_{snail})$ das 51.704 configurações tendo $n_{snail} = 0, 1, 2$ ou 3 padrões <i>Snail</i> na mesma estrutura. O <i>inset</i> mostra o mesmo para configurações com padrões <i>Staple</i> . . . . .	.44
Gráfico 4.6. Distribuição dos valores da energia em função da ordem de contato .	45
Gráfico 5.1. Comportamento do logaritmo da razão de enovelamento $\log k_f$ e sucesso de enovelamento (--*--) como função da ordem de contato relativa. Potencial hidrofóbico utilizado na simulação $\{e_{i,j} = h_i + h_j\}$ . A fraca correlação entre $\log k_f$ e $\chi$ foi calculada com as estruturas nativas com sucesso de enovelamento maior que 66%. Note que o intervalo de $\chi$ foi segmentado em três faixas complementares de Baixa, Intermediária e Alta ordem de contato. Estas faixas são representadas pelas Letras L, I e H, respectivamente. . . . .	.52
Gráfico 5.2. Comportamento do logaritmo da razão de enovelamento $\log k_f$ na simulação com o conjunto de vínculos estéricos $\{c_{ij}\}$ como função da ordem de contato relativa. Os resultados correspondem à completa interação dos pares de aminoácidos $\{e_{i,j} + c_{i,j}\}$ . . . . .	.54

## LISTA DE SIGLAS

1BY0 – <i>wild-type plastocyanin from silene</i>	. . . . .	.10
1CSP – <i>universal nucleic acid-binding domain</i>	. . . . .	.11
1HRC – <i>horse heart cytochrome c</i>	. . . . .	.11
1L2Y – <i>Trp-Cage Miniprotein Construct Tc5B</i>	. . . . .	.11
1LMB – <i><math>\lambda</math> repressor-operator complex</i>	. . . . .	.11
3-D – Espaço tridimensional	. . . . .	.02
BPTI – Inibidor de Tripsina Pancreática de Bovino	. . . . .	.14
BSE – Encefalopatia Espongiforme Bovina	. . . . .	.03
CSA – <i>Compact Self-Avoiding</i>	. . . . .	.04
DM – Dinâmica Molecular	. . . . .	.12
DNA – Ácido Desoxirribonucléico	. . . . .	.01
FGF – <i>Fibroblast Growth Factor</i>	. . . . .	.02
FPT – <i>first passage time</i>	. . . . .	.34
HP – Hidrofóbico Polar	. . . . .	.21
MC – Monte Carlo	. . . . .	.15
NMR – Ressonância Magnética Nuclear	. . . . .	.03
PDB – <i>Protein Data Bank</i>	. . . . .	.01

## LISTA DE SÍMBOLOS

$\chi$ – Ordem de contato relativo	. . . . .	.05
$k_f$ – Taxa de enovelamento	. . . . .	.05
$C\alpha$ – Carbono-alfa	. . . . .	.07
COOH – Grupo Carboxila	. . . . .	.07
NH <sub>2</sub> – Grupo Amino	. . . . .	.07
H – Átomo de Hidrogênio	. . . . .	.07
R – Radical	. . . . .	.07
H <sub>2</sub> O – Molécula de água	. . . . .	.07
$\Phi$ – (phi) Ângulo de rotação entre o nitrogênio e o carbono-alfa	. . . . .	.08
$\Psi$ – (psi) Ângulo de rotação entre o carbono alfa e o carbono	. . . . .	.08



## SUMÁRIO

<b>1 INTRODUÇÃO</b>	.01
1.1 Motivação.	.01
1.2 Abordagem do problema .	.04
1.3 Especificando o problema deste trabalho	.05
<b>2 MÉTODOS DE SIMULAÇÃO E O MÉTODO MONTE CARLO</b>	.07
2.1 O sistema protéico e suas representações estruturais	.07
2.2 Dinâmica Molecular	.12
2.3 Método Monte Carlo	.15
2.4 Método Monte Carlo e o modelo computacional utilizado neste trabalho	.19
<b>3 MECANISMOS DE ENOVELAMENTO E O MODELO EM REDE.</b>	.21
3.1 <i>Design</i> da proteína: especificidades químicas e estéricas.	.23
3.2 Simetrias do cubo: enumeração das CSA	.29
3.3 Programa computacional implementado.	.30
3.4 Ordem de contato reativo.	.31
3.5 Taxa de enovelamento	.33
<b>4 CARACTERIZAÇÃO TOPOLÓGICA DAS 51.704 CONFIGURAÇÕES CSA</b>	.37
4.1 Elementos topológicos básicos	.39
4.2 Tipos de finais de cadeia .	.40
4.3 Padrões estruturais	.41
4.4 Energia da cadeia na configuração CSA.	.44
4.5 Tabela resumo	.45
<b>5 RESULTADOS E CONCLUSÕES</b>	.50
5.1 Análise dos diferentes grupos de estruturas notáveis	.55
5.2 Comentários e conclusões	.62
5.3 Trabalhos futuros .	.66
<b>6 BILIOGRAFIA</b>	.67
<b>ANEXOS</b>	.75