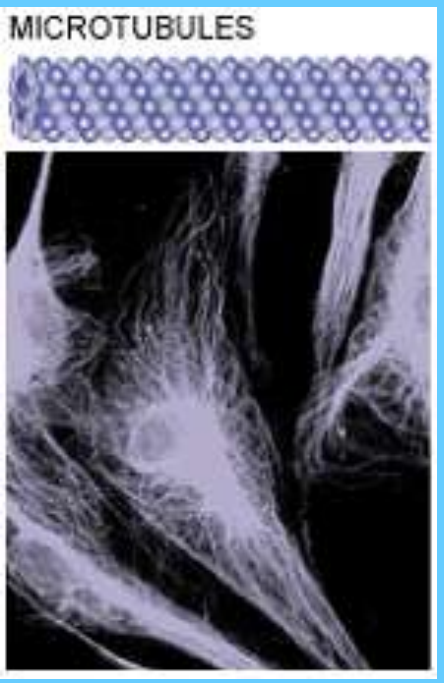
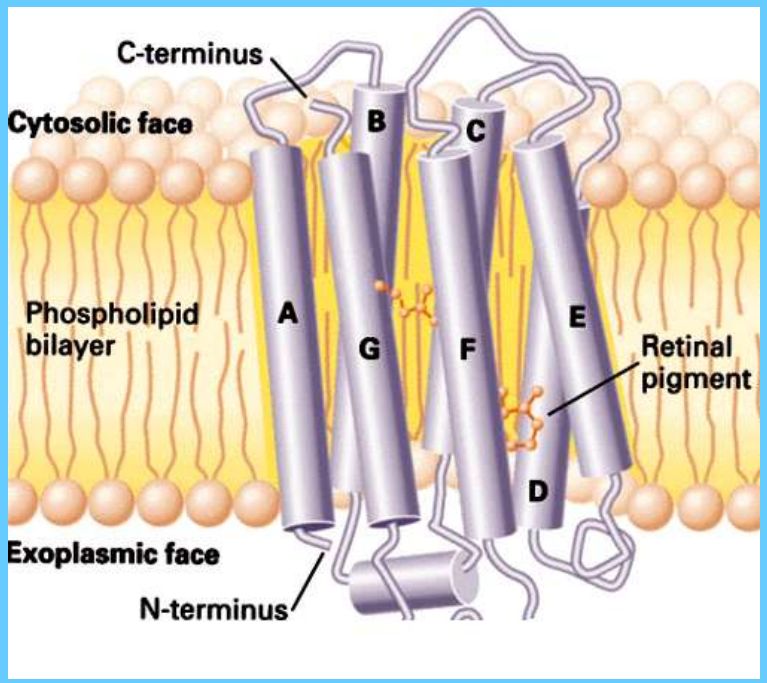
Efeito das flutuações térmicas locais na cinética do processo de *folding* de proteínas

Grupo de Física Biológica
Departamento de Física e Química
FCFRP – USP



Proteína Estrutural



Proteína de Membrana

Proteína Globular

O problema sob a abordagem da MENE

Folding de proteína

- Características do *folding*
- Modelos de *folding*
- O modelo Estereoquímico:
  - objetivos
  - dificuldades

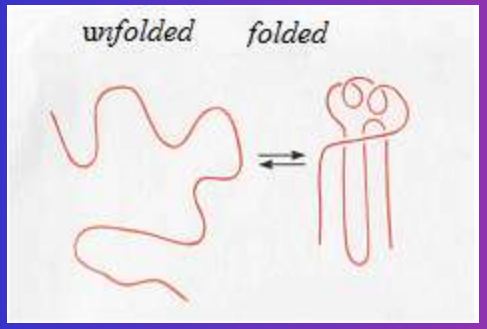
Mecânica Estatística Não-Extensiva

- Origem: generalização da entropia .

$$\ln W = \lim_{q \to 1} (W^{q-1} - 1)/(q-1)$$

$W > 0;\ q \in \Re$

- Peso de Tsallis como uma média do peso de Boltzmann devido à flutuação térmica local.



$$S_q = k \frac{1 - \sum_{i=1}^{W} p_i^q}{q - 1}$$

Proteínas Globulares

Proteínas são heteropolímeros lineares e flexíveis, cujas unidades provém de um repertório típico de 20 diferentes aminoácidos (aa) naturais.

Sequência de aa: {Ser, Tir, Met, Glu, ..., Glu, Fen} ≡ Estrutura Primária

ID da proteína → sequência de aminoácidos

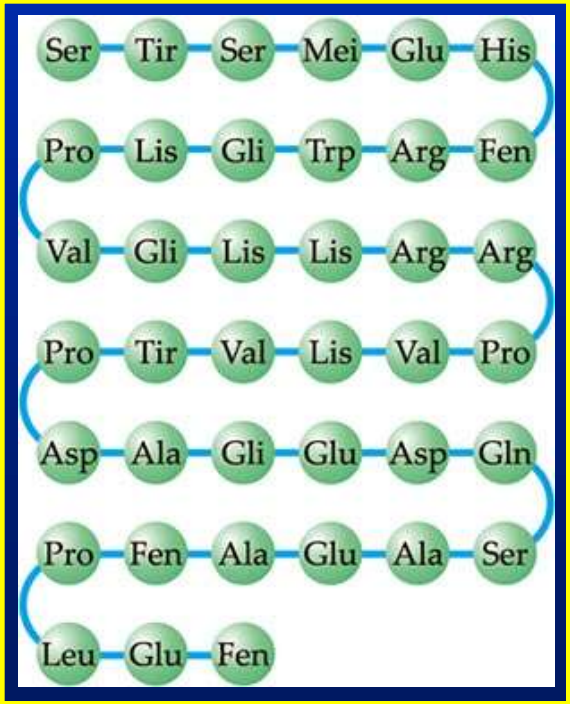
Então, o número N_L de proteínas diferentes, de tamanho L, é muito grande: $N_L = 20^L$.

Exemplo: $L = 62$ (proteína relativamente pequena)

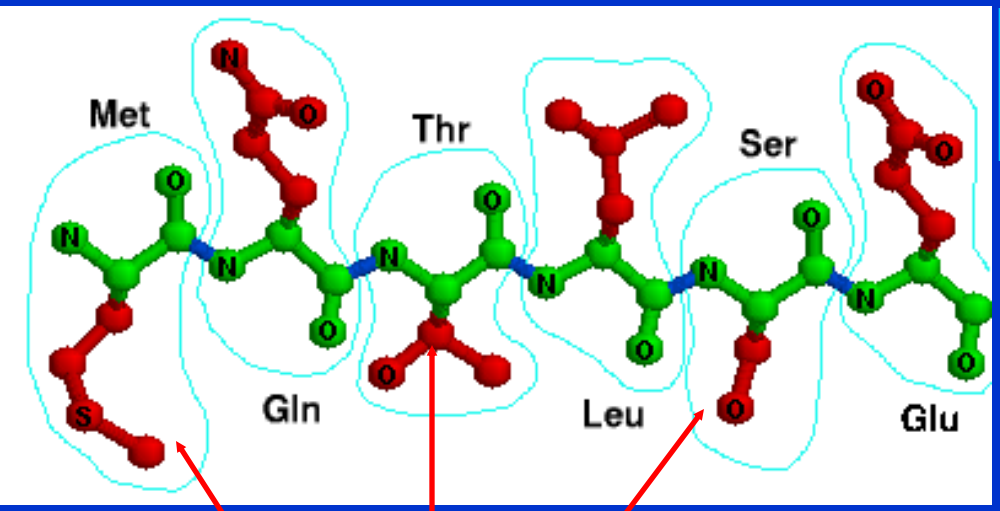
$$N^{62} \sim 10^{80}$$

Algo como o número de átomos visíveis do Universo

Proteína de tamanho L = 39 aa

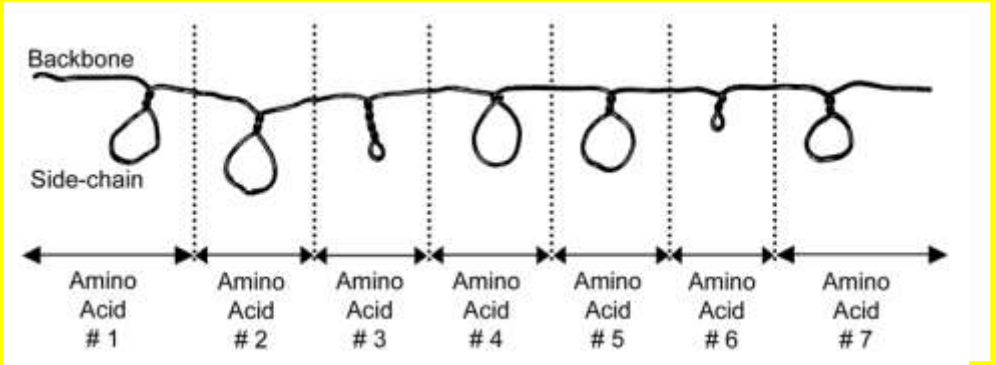


Repertório: tamanhos, formas e função química

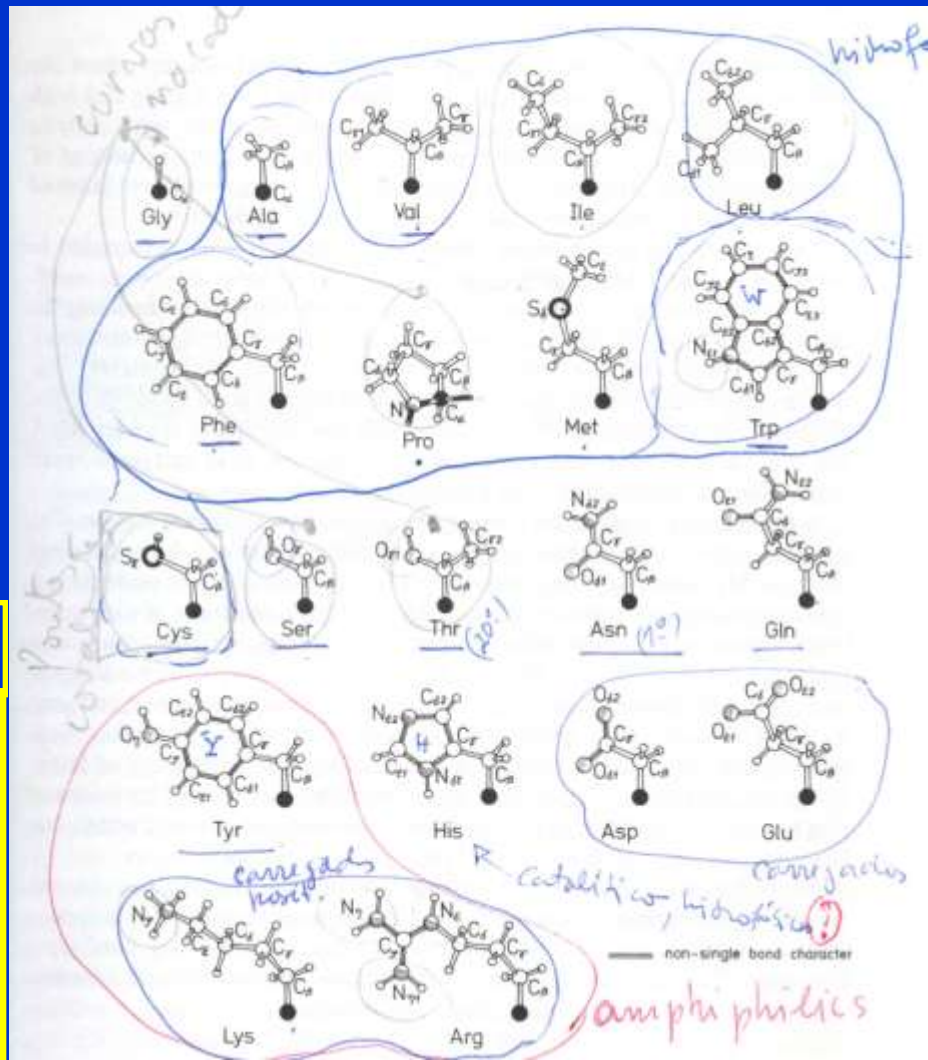


cadeias laterais

ID: sequência de aa (apelidos: função, ...)



Alfabeto de 20 amino ácidos distintos



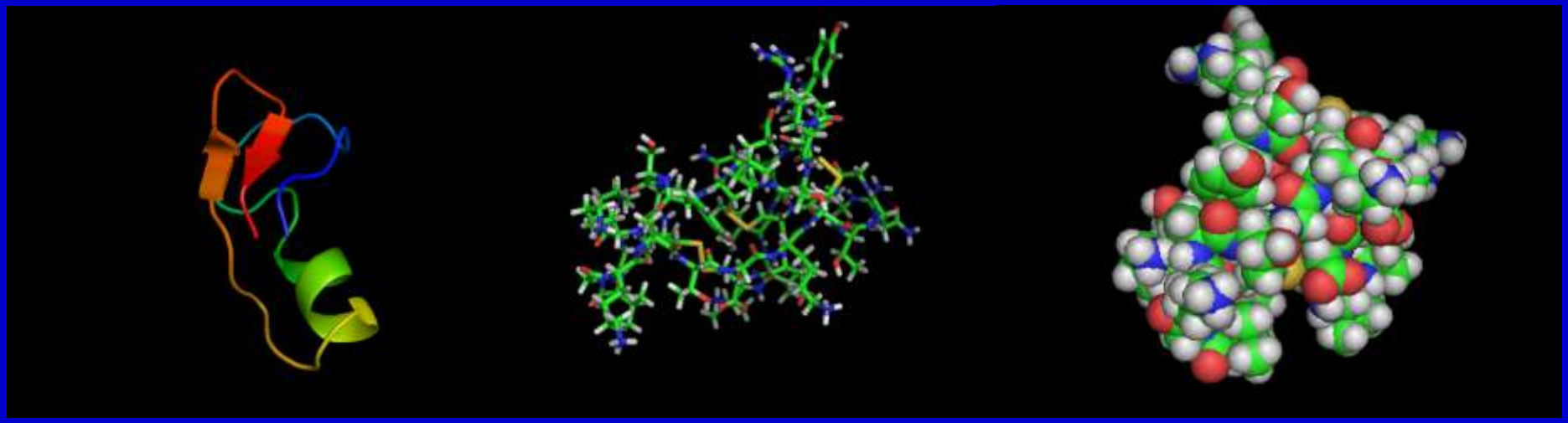
Representação da estrutura nativa de uma proteína

PDB ID: 1TSK

SEQUÊNCIA: Met - Lys – Val – Leu - ...

MKVLYGILIIFILCSMFYLSQEVVIGQRCYRSPDCYSACKKLVGKATGKCTNGRCDC

Estrutura secundária (*cartoon*) ***Sticks*** **Esferas**

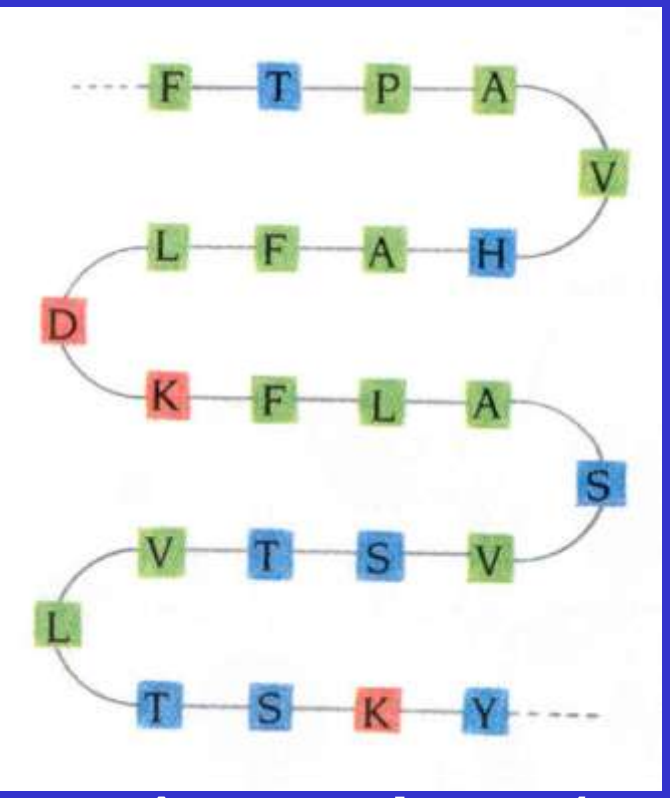


O problema de *folding*

– *FOLDING* direto (Polissemia): Dada a estrutura primária DETERMINAR sua(s) correspondente(s) estrutura(s) terciária(s)

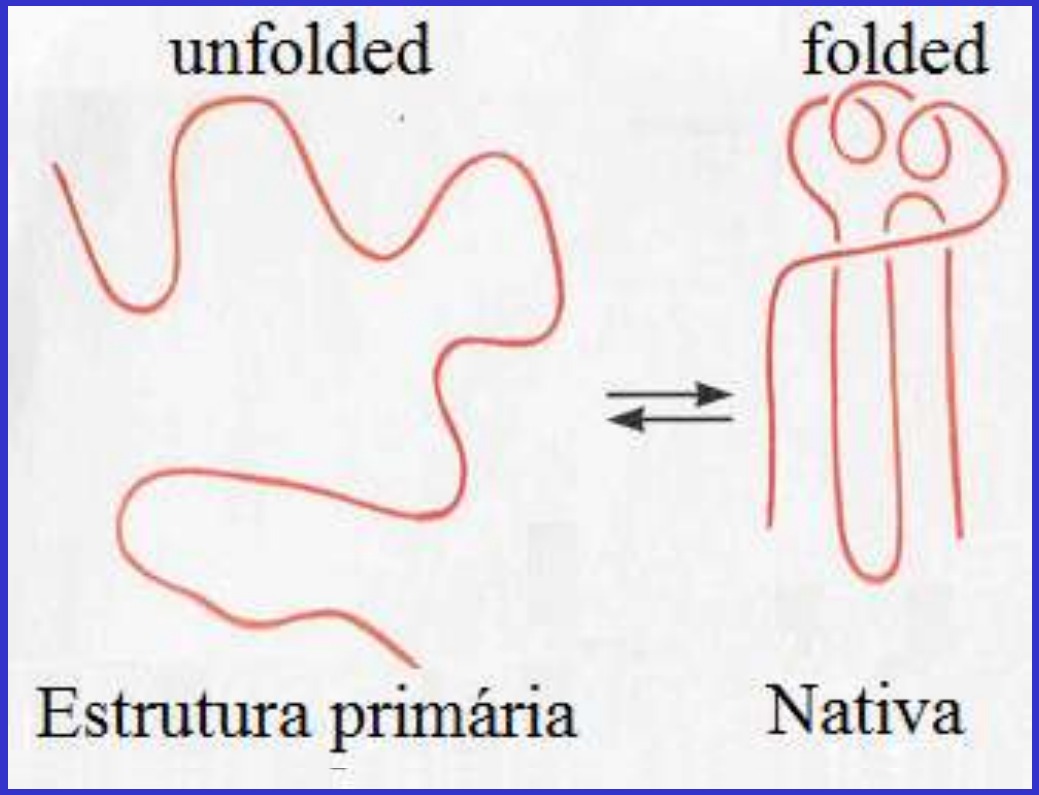
– *FOLDING* inverso (Sinonímia): Dada a estrutura terciária DETERMINAR sua(s) correspondente(s) estrutura(s) primária(s)

Representação de uma cadeia polipeptídica



Sequência: ID da proteína

Processo de *folding* - evolução esquemática: estrutura primária → secundária → terciária



Nativa: estrutura (3D); precisa, estável, única

folding como gramática

Considere uma **sentença** constituida de uma sequência linear de **letras** (selecionadas de um "alfabeto" de 26 **letras**); pequenas regiões da **sentença** podem ter existência autônoma (chamamos tais regiões de "**palavras**") mas o **significado da sentença** como um todo somente emerge quando todas estas **palavras** são colocadas juntas –frequentemente modificando as **palavras** quando isoladas

Trocando: setença → proteína; letra → aminoácido;

palavra → elementos de estrutura secundária; significado → função.

Considere uma **proteína** constituída de uma sequência linear de **aminoácidos** (selecionados de um "alfabeto" de 20 **aminoácidos**); pequenas regiões da **proteína** podem ter existência autônoma (chamamos tais regiões de "elementos de estrutura secundária") mas a **função da proteína** como um todo somente emerge quando todos estes **elementos de estrutura secundárias** são colocados juntos –frequentemente modificando os **elementos de estrutura secundária** quando isolados.

O processo de *folding*

O *folding* das proteínas globulares é pensado aqui como um processo em dois estágios temporais distintos:

1. Estágio de Busca: a cadeia polipeptídica, guiada pelo efeito hidrofóbico, evolui erraticamente e é compactada (barreira entrópica) – etapa lenta;
2. Estágio de Refinamento Estrutural: próximo o suficiente da estrutura nativa, e somente nesta condição, interações intramoleculares e interações molécula-solvente são otimizadas; a partir daí, ajustes espaciais levam à precisão conformacional e à estabilidade da proteína – etapa ultra rápida.

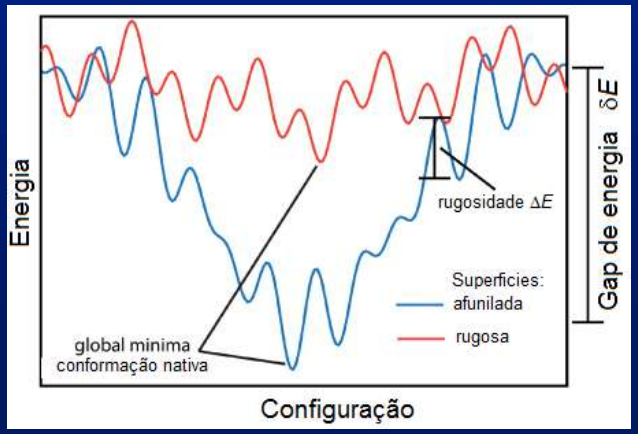
Ênfase no primeiro estágio

O processo: taxa de *folding*

Para pequenas proteínas globulares (N ~100) *de um único domínio, sem prolinas, pontes disulfeto ou grupos prostéticos a taxa de folding cobre seis ordens de grandeza*

Insights a partir de simulações de modelos simplificados

- diferentes sequências apresentam tempos muito distintos de *folding;*
- é difícil encontrar sequências que levam a *folding* rápidos;
- *folding* rápido estaria associado a superficies de energia suave (espaço configuracional), e na forma de funil (rugosidade < $k_B T$);
- um pronunciado mínimo de energia entre o estado fundamental e o segundo estado de mais baixa energia potencial garantiria um *folding* rápido;
- taxa de *folding* estaria relacionada com colapso cooperativo da cadeia: $\sigma = 1\text{-}T_m/T_\theta$



Limitação destes entendimentos para *folding*: experimentos computacionais são feitos, em geral, com sequências rápidas.

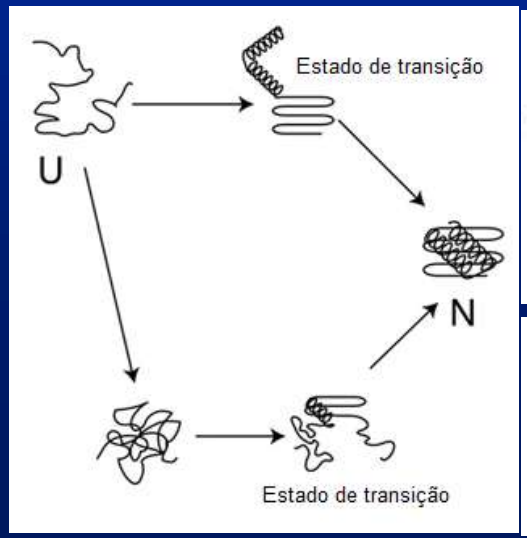
O processo: taxa de *folding*

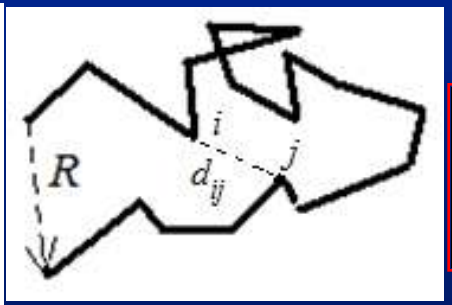
Para pequenas proteínas globulares (N ~100) de um único domínio, sem prolinas, pontes disulfeto ou grupos prostéticos a taxa de folding cobre seis ordens de grandeza

Insights a partir de experimentos

- Modelo Nucleação-condensação: estado de transição para o *folding* contém um núcleo de contatos nativos; o tempo de formação de tal núcleo poderia contribuir para a taxa *de folding*.

- Topologia da Nativa como um determinante das taxas de *folding*: Ordem de Contato (CO ≡ média da distância ao longo da cadeia, entre pares de resíduos contactantes da estrutura nativa), e parâmetro Ω (média da probabilidade dos pares de contatos nativos (baseado numa cadeia guassiana).





Cadeia ideal

$P(R) = (3/2\pi Nl^2)^{3/2}\exp(-3R^2/2Nl^2)$;

Seja: $l = 1$; $R = 1$; $N = d_{ij}$; $\rightarrow$

$$CO = (1/N_C)\sum_{\{i,j\}} d_{i,j}$$

$$\Omega \propto \sum_{\{i,j\}} (L - d_{ij}) \frac{\exp(-3/2d_{ij}^2)}{d_{ij}^{3/2}}$$

Características gerais do processo de *folding* de proteínas globulares

1. Rapidez: *folding* é um processo geral e muito rápido;
2. Robustez: *folding* é similarmente processo em um grande intervalo de temperatura: dependendo do organismo vivo, ocorre em todas temperaturas de 0^{o} a $100^{o}C$;
3. Mano-máquina independente : como em partículas coloidais, o glóbulo protéico é sistematicamente agitado e deformado por forças não-balanceadas.

1. O processo de *folding* é muito rápido!

→ … dificuldades para cálculos computacionais.

- O processo é pelo menos 10 ordens de grandeza mais rápido do que um mecanismo randômico de busca no espaço das configurações.

- Estima-se que, para proteínas globulares relativamente pequenas e de um único domínio, a **velocidade limite** de *folding* demanda um tempo τ mínimo igual ao tempo necessário para formação correta de todos os elementos estruturais da proteína (estrutura secundária). Grosseiramente,

$$\tau \sim N/100 \ \mu s$$

onde N é tamanho da proteína (número de amino ácidos).

Conjectura de Levinthal

Tempo característico de folding

Proteína de tamanho $N = 50$ (50 aminoácidos)

Γ : espaço configuracional contendo $(4,5)^{50} \sim 4,6 \times 10^{32}$ config.

Então, se $\tau = 10^{-15}$ s é tempo para uma visita ao acaso cada config.

$$\rightarrow \quad t_0 \sim 4,6 \times 10^{32} \times \tau = 4,6 10^{17} \text{ s},$$

onde t_0 é o tempo para varrer Γ, o que equivale, aproximadamente, a 10 × Idade do Universo[1] !

Caminho(s) preferencial(is) em Γ !

[1] Idade do universo ≅ 13,5 bilhões de anos ≅ $4,3 \times 10^{16}$ s

Velocidade limite

Para cada tipo de elemento estrutural (estrutura secundária: α-hélices; β-*strands*; *loops*), o tempo característico de *folding*, τ, é necessariamente maior do que o tempo para formação dos elementos estruturais. Em geral, $\tau_\alpha < \tau_\beta$; $\tau_\alpha < \tau_{\alpha\beta}$ (para proteínas de tamanhos comparáveis)

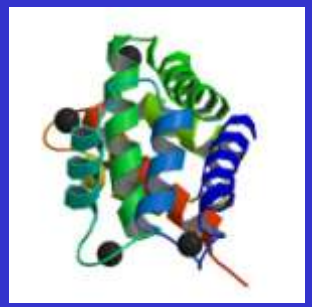
$\tau_{loop} \sim (0.07 - 0.4)\ \mu s$

$\tau_\alpha \sim 0.5\ \mu s$ (peptídeos ricos em alanina)

$\tau_\beta \sim 6\ \mu s$

Oito ordens de grandeza!

Proteínas Globulares



(Calgranulina)

$\tau \sim (10^{-5} - 10^{3})\ s$

Simulação do processo de *folding*

Dinâmica Molecular

Principal dificuldade - tempo de máquina para simular 1 μs é ainda muito grande: para se determinar o tempo característico de *folding* de proteínas de tamanho relativamente pequeno, $N \sim 100$ amino ácidos, o tempo de máquina pode ser de semanas de CPU nas mais rápidas máquinas disponíveis na atualidade.

Solução (precária, mas disponível): computação distribuída:

Se a evolução do número M de proteínas enoveladas evolui exponencialmente, isto é, $M_N(t) = M_0[1- \exp(-t/\tau)]$, então para tempos t pequenos em comparação a τ, a razão $M(t)/M_0 \cong (-t/\tau)$, isto é: ao final de um tempo t, temos que o número esperado de proteínas enoveladas será

$$M(t) \cong (M_0/\tau)t.$$

Assim, se τ for da ordem de 1 μs, $M_0 = 100$ *jobs* independentes submetidos em M_0 computadores diferentes, ao cabo de $t = 0{,}1$μs (tempo real de simulação) teremos $M(t) \sim 10$ simulações bem sucedidas!

Alternativa

Simulação Monte Carlo (algoritmo de Metropolis)

Vantagens: varredura rápida do espaço conformacional

Desvantagens:

1- perda de detalhes (modelos minimalistas);

2- retenção do processo em armadilhas energéticas (mínimos locais).

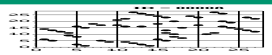
Remédio para a dificuldade 2

Mecânica Estatística não Extensiva: $\exp(-\beta\Delta E) \rightarrow \exp_q(-\beta\Delta E)$

onde

$$\exp_q(-\beta\Delta E) = [1-(1-q)\beta\Delta E]^{1/(1-q)}.$$

Demonstra-se que



é uma média do fator de Boltzmann, $\exp(-\beta\Delta E)$, ponderada pela distribuição χ-quadrado $f(\beta)$. Isto significa que $\exp_q(-\beta_0\Delta E)$ leva em conta as flutuações locais da temperatura.

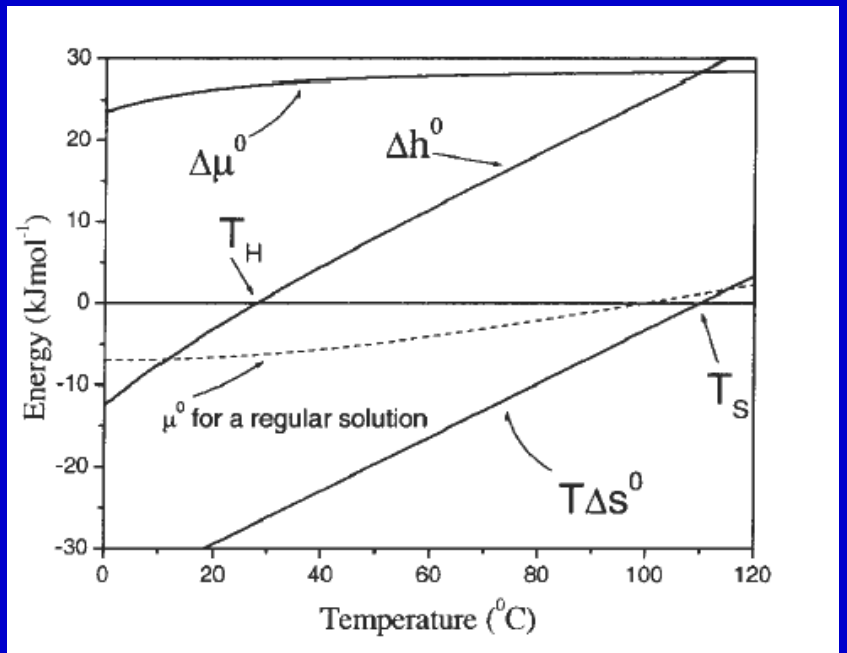
2. Robustez do processo de *folding*

O *folding* de proteínas ocorre em ambientes **extremos** (temperatura, pressão pH), além de ambientes menos radicais.

Alguns organismos como bactérias, *archea* e mesmo animais superiores (Verme de Pompéia), são termófilos, podendo se desenvolver, em alguns casos, em temperaturas acima de 100º C.

Outros organismos são psicrófilos (bactérias, fungos e algas), existindo casos em que se desenvolvem em temperaturas abaixo de 0º C.

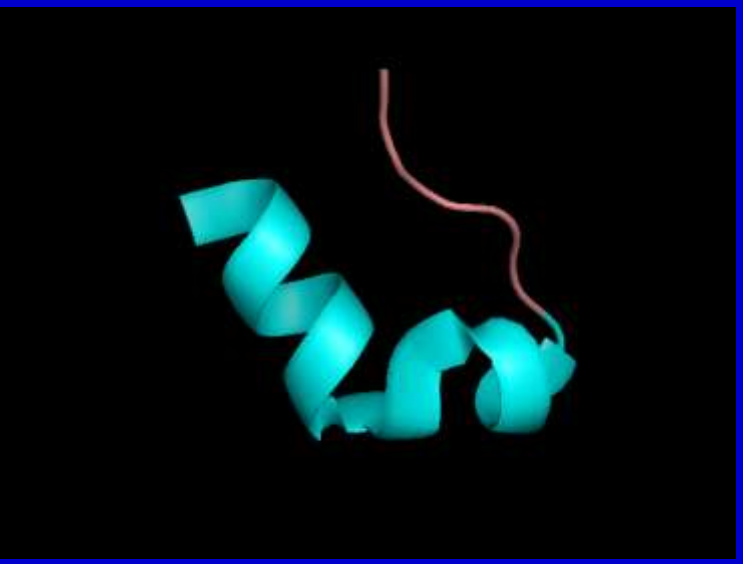
Bactérias: Yellowstone



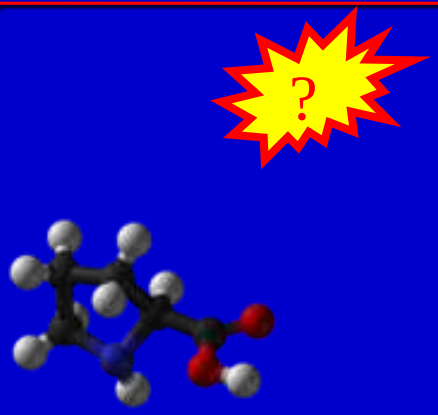
Tardigrada

3. Proteína: uma nano-máquina

Proteínas globulares apresentam diferentes tamanhos, que pode variar de **~ 1 a 20 nm** (de um a 30.000 aminoácidos, o que equivale; tipicamente, de 1 a 3.000 Kdaltons). Nesta escala espacial, forças locais não-balanceadas agitam e deformam sistematicamente o glóbulo protéico.
Como flutuações térmicas locais poderiam afetar o processo de *folding*?



Trp-Cage: 20 aa - designed



Prolina - a menor proteína: < 1 nm

Titin: 30.000 aa ~ 20 nm

Propriedades ⟶ 3 premissas

1. O processo de *folding* é composto de dois estágios mecânicamente e temporalmente independentes; efeito hidrofóbico governa o primeiro estágio do *folding*.
2. Para proteínas globulares de dois estados e um único domínio, a instrução codificada em sua sequência de amino ácidos provê uma cinética do processo de *folding* tão rápido quanto possível.
3. Flutuações térmicas locais emergem como atributo intrínseco do sistema proteína-solvente, promovendo um processo de *folding* eficiente.

Desafio: incorporação destas premissas num MODELO

JUSTIFICATIVA:

A energética do processo de *folding* é ainda pobremente compreendia: continua sendo muito controverso, pois ainda produz resultados teóricos e experimentais diametralmente opostos.

Assim, concentrar-se em algumas questões específicas é imperativo. E uma questão recorrente, no tocante à simulação computacional, é:

Teria a água papel determinante no processo de *folding*?

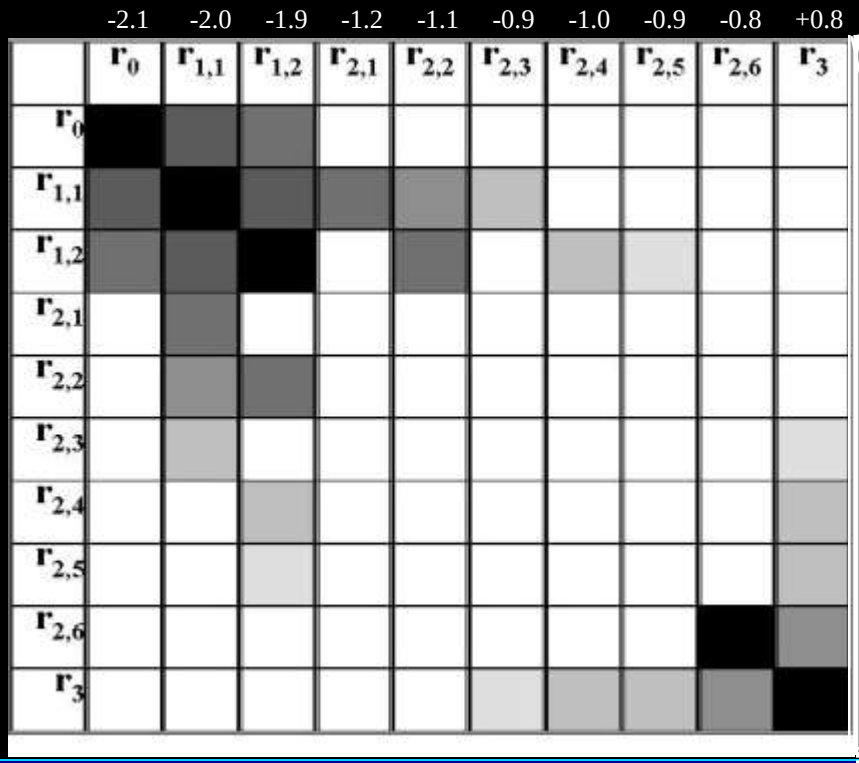
A resposta tem importantes consequências na elaboração de modelos minimalistas; e estes ainda são indispensáveis para se obter *insights* sobre o processo de *folding*.

MODELO ESTEREOQUÍMICO

Modelo Estereoquímico: Uma cadeia protéica embebida numa rede cúbica infinita

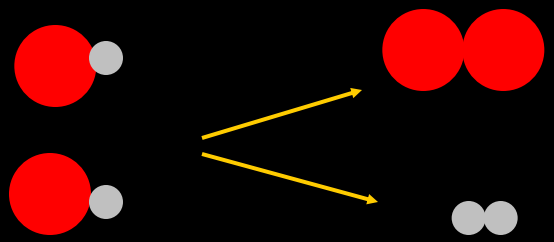
Efeito do solvente incluído → $e_{i,j} = (h_i + h_j) + c_{i,j}$

Cadeia "27-mer" e alfabeto de 10-letras: pares de restrições espaciais $\{c_{i,j}\}$ e níveis de hidrofobicidade $\{h_i\}$



$h_i \equiv$ nível de hirofobicidade do residuo *"i"*

O princípio da segregação, a saber, $2h_{ij} \geq (h_{ii} + h_{jj})$, é marginalmente satisfeito pelo potencial hidrofóbico, $h_{ij} = h_i + h_j$, isto é: $2h_{ij} = (h_{ii} + h_{jj})$.

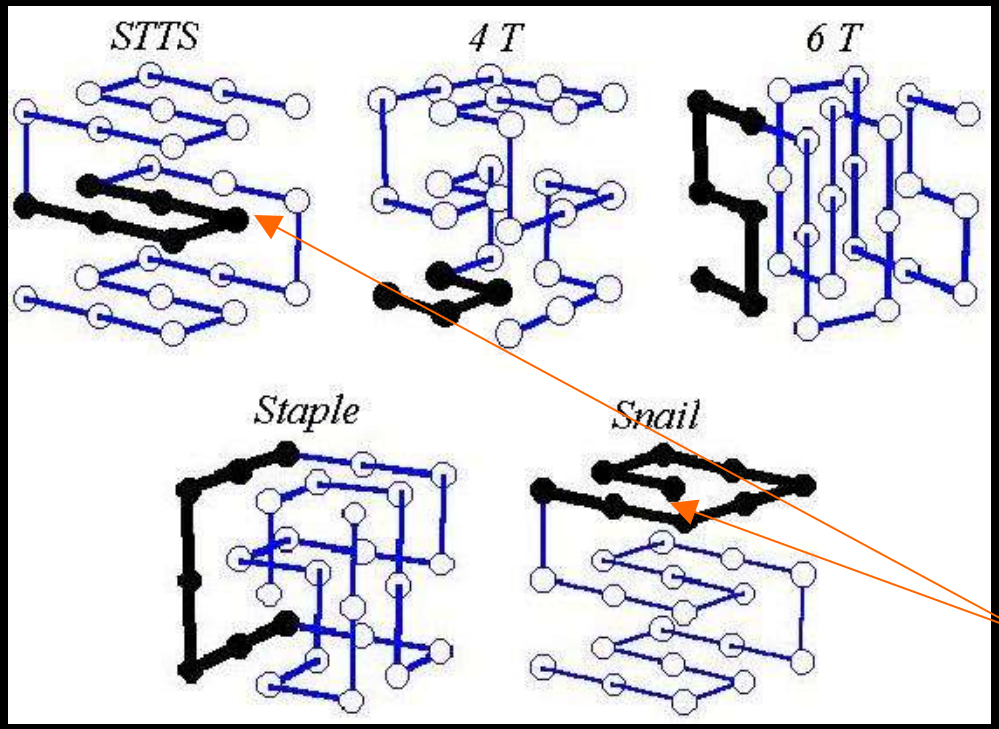


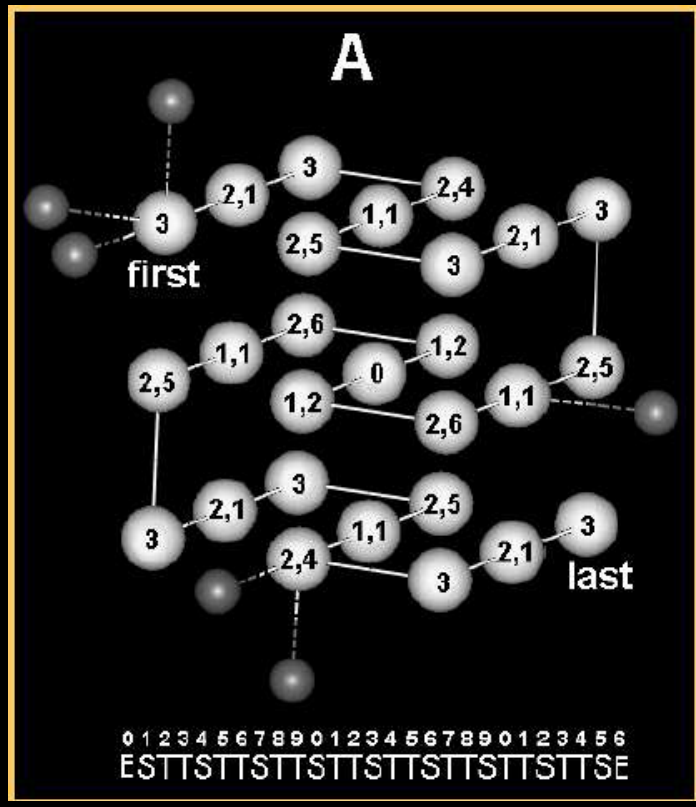
As restrições espaciais $\{c_{i,j}\}$ reduzem a flexibilidade da cadeia e aumentam a especificidade intra-cadeia

O modelo também provê a "sintaxe" para o *design* da sequência, dada qualquer estrutura 3-D

Energia livre configuracional

$$G(\{k,l)\} \propto \sum_{\text{all}\{i,j\}} (h_{i,j} + c_{i,j})\delta_{(i,j),\{k,l\}}$$





Detalhes configuracionais importantes das ≅ 50.000 CSA configações.

Estruturas Nativas são representadas por configurações *Compact Self Avoiding* (CSA) – formando "cubos" 3 3 3

Questões específicas decorrentes

Se compreender completamente o processo de folding fosse já um fato, poderíamos então responder todas as questões específicas, como:

1. Dada uma sequência de aminoácidos, sob especificadas condições (pH, temperatura, pressão, ...), qual a sua estutura 3-D correspondente?
2. Alternativamente, conhecida a estrutura de uma proteína, qual seria(m) a(s) possível(is) sequência(s) de amino ácidos correspondentes?
3. Ou ainda, conhecida a sequência e/ou a estrutura nativa de uma proteína, qual seria o tempo característico de *folding* da mesma?

Simulação Monte Carlo

sistema, parâmetros, objetivo, método

1. Rede cúbica infinita;
2. Cadeia de 27 monômeros {← repertório de 10 # resíduos};
3. Estrutura alvo: uma das CSA;
4. Sequência: desenhada a partir da configuração CSA selecionada;
5. Condição Inicial: cadeia aberta (sem contatos topológicos);
6. Temperatura: fixa;
7. Objetivo: obtenção do tempo característico de *folding* (tempo de primeira passagem pela Nativa -configuração CSA escolhida);
8. Método: Metrópolis com peso de Boltzmann generalizado.

Peso de Tsallis

$$\exp_q(-\beta_0\varepsilon) = \int_0^\infty \exp(-\beta\varepsilon)\, f(\beta) d\beta,$$

onde $$f(\beta) = \frac{1}{\Gamma(n/2)}\left(\frac{n}{2\beta_0}\right)^{n/2} \beta^{-1+n/2} \exp(-\frac{\beta}{2\beta_0} n),$$

e impondo que a média da temperatura, $<\beta^{-1}> = \int_0^\infty \beta^{-1}\, f(\beta) d\beta$, coincida com a temperatura do reservatório $\beta_o{}^{-1}$, obtemos

$$\exp_q(-\beta_0\varepsilon) = 1 - (1-q)\beta_0\varepsilon \quad , \quad \text{onde}$$

$q = 1 + 2/n$, e n é o número de graus de liberdade da distribuição χ^2, $f(\beta)$, isto é, *n é* o número de variáveis independentes que especifica completamente o sistema.

Parâmetro entrópico q como uma variável

O índice entrópico q está essociado com o número de graus de liberdade da distribuição: $q = 1+2/n$. Desta forma, como q estaria relacionado com os graus de liberdade do sistema físico?

Esta questão se impõe pelo fato do índice entrópico q ser usualmente introduzido " à mão " –*fitting*; mas este não é o nosso caso.

Ensaios numéricos indicaram que:

1. existe $q = q^* > 1$ que minimiza o tempo característico τ de *folding*; q^* é diferente para cada estrutura alvo escolhida (CSA);
2. associando n (e então q) ao grau de compactação do glóbulo ($q = a/R_G^2$) foi possível estabelecer um valor máximo para q, a saber, $q_{max}=1+ 2/6$, de tal forma que τ é minimizado: então obtemos $q = 1+2(q_{max}-1)/R_G^2$. O mínimo de $R_G^2 = 2$, que para o caso, $q = q_{max}$
3. a média $< q >$ da distribuição de q coincide com q^*.

Este resultado satisfaz a primeira característica do processo de *folding:*

Rapidez – *folding* é um processo geral e muito rápido.

O índice entrópico q como uma variável

Assim, o valor de q varia ao longo da simulação, sendo que

$q_{min} = 1$, correspondendo a $n \rightarrow \infty$ (cadeia aberta, número máximo de graus de liberdade, e

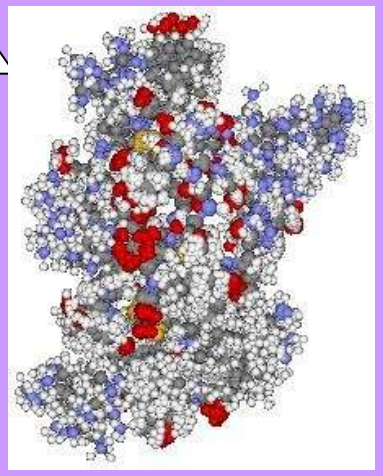
$q_{max} = 4/3$, correspondendo a $n \rightarrow 6$. Mas, por que $n = 6$?

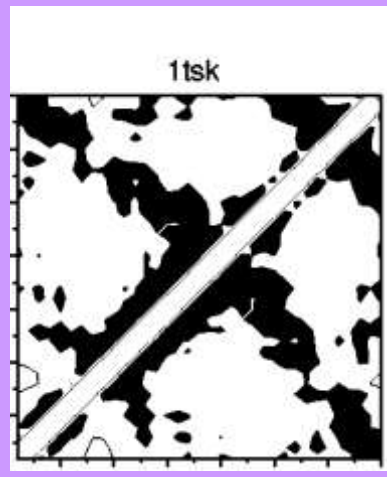
Representação de uma proteína

DETALHES

Ana Paula Arósio

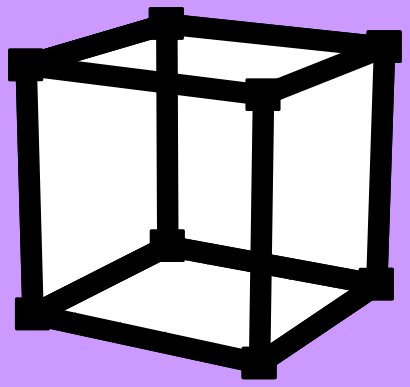
PERFEIÇÃO

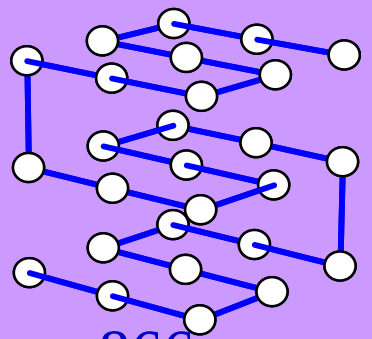




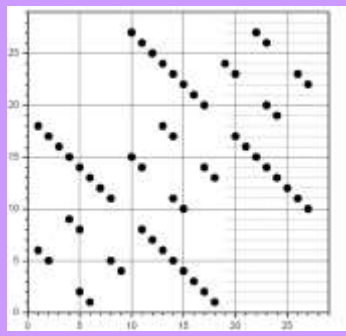
1TSK

Falcão



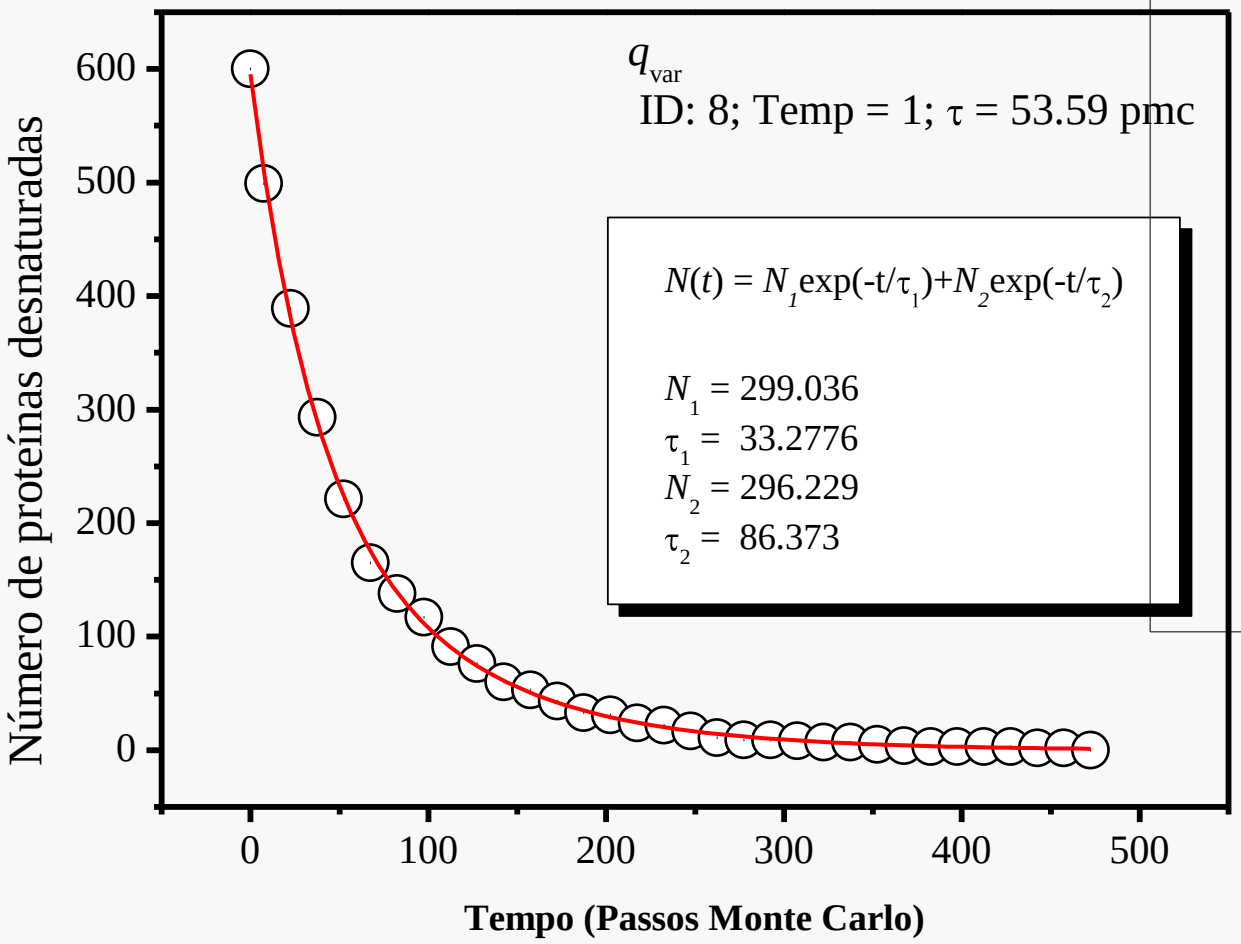


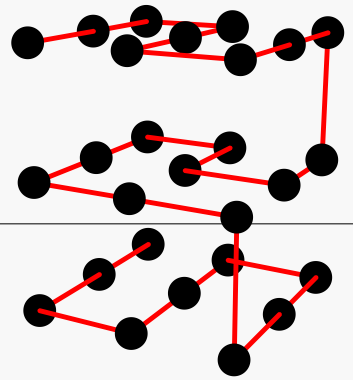
866



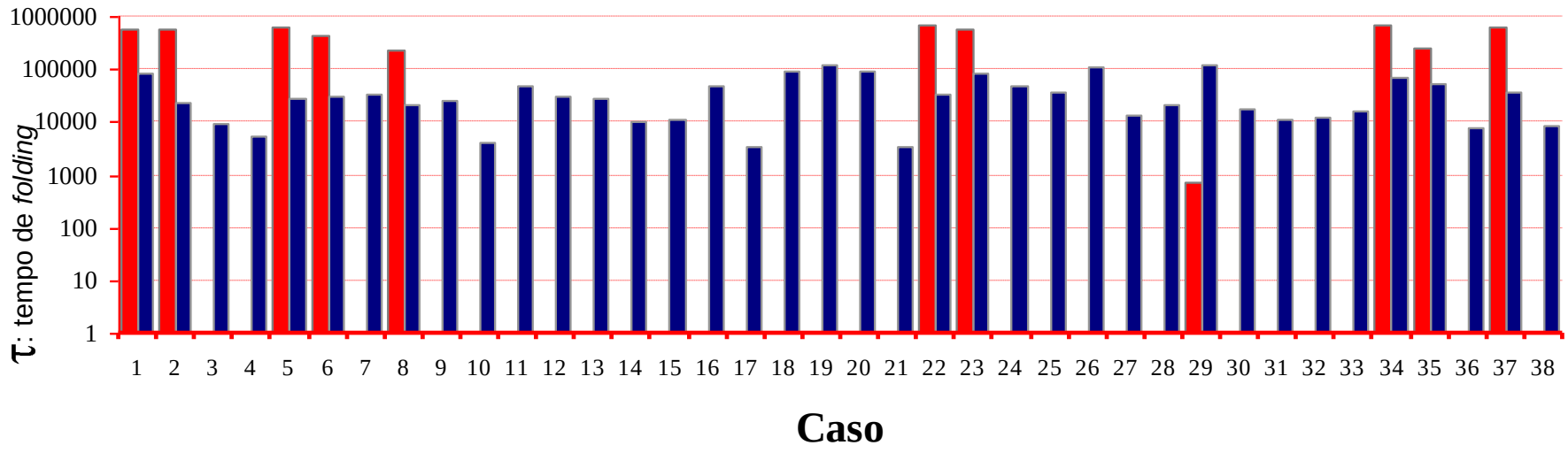
Resultados: tempo característico de *folding* τ

600 simulações independentes

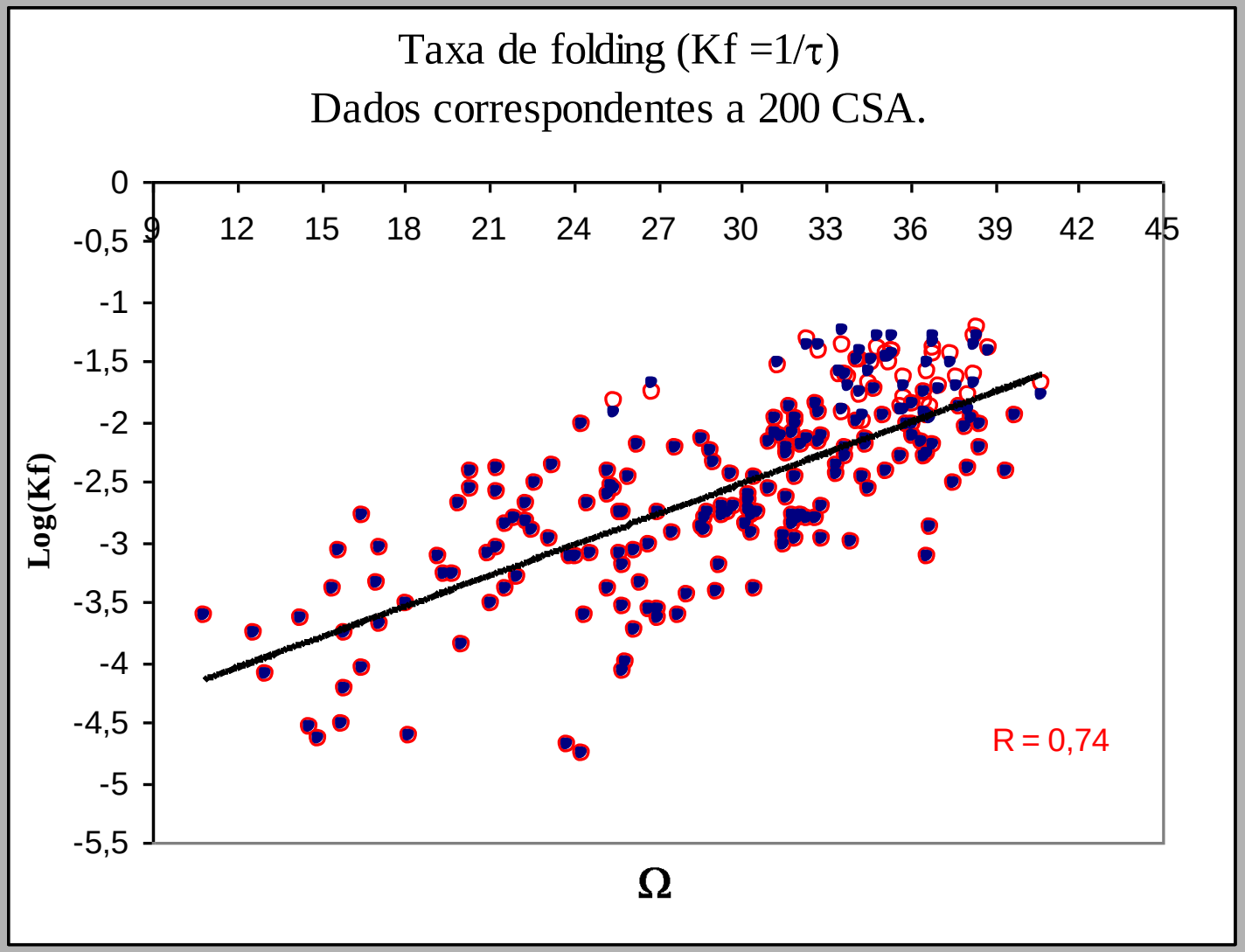




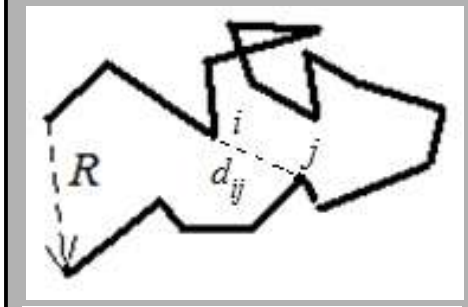
Comparação



Barras (vermelhas) ausentes correspondem aos casos de insucesso na busca pela estrutura nativa, na janela de tempo estipulada (~8×10^8 passos MC).



$$CO = (1/N_C)\sum_{\{i,j\}} d_{i,j}$$



$d_{i,j}$: distância ao longo da cadeia

Cadeia Ideal:

$P(R) = (3/2\pi Nl^2)^{3/2}\exp(-3R^2/2Nl^2)$

Para R =1 (contato) e l = 1 →

$$\Omega \propto \sum_{\{i,j\}} (L - d_{ij}) \frac{\exp(-3/2d_{ij}^2)}{d_{ij}^{3/2}}$$

FIM

Gradecimentos:

João Paulo Dal Molin

Marco A.A. Da Silva

Grupo de Física Biológica

Ref.: PHYSICAL REVIEW E 84, 041903 (2011)