3. MECANISMOS DE ENOVELAMENTO E O MODELO EM REDE

Muitos modelos minimalistas em redes regulares e no espaço contínuo têm sido propostos com o objetivo de estudo do mecanismo do enovelamento das proteínas. O primeiro destes modelos[36] foi desenvolvido na década de 70 e atualmente é conhecido como modelo de Gō[37]. Já com o propósito de obter representações simplificadas de cadeias polipeptídicas, um dos primeiros modelos simples de proteína em rede, foi introduzido por Lau e Dill[38] (1989) e foi chamado de modelo HP. Neste modelo, os vinte aminoácidos naturais (Figura 1.1, página 3) são agrupados em duas categorias chamadas polar, representada por P, e não polar, representada por H (hidrofóbico). A vantagem de se estudar modelos minimalistas em rede é principalmente devido à possibilidade de se efetuar análises mais rigorosas que produzem respostas detalhadas, pois o espaço das configurações pode ser totalmente conhecido, principalmente para cadeias de tamanho reduzido.

As proteínas reais enovelam-se em um tempo muito pequeno, comparativamente ao tempo necessário para uma busca aleatória pelo estado de menor energia global, como já citado no Capítulo 2. Com o objetivo de mimetizar, entre outros parâmetros, a taxa de enovelamento das proteínas reais, diferentes modelos em rede para o mecanismo de enovelamento têm sido adotados (Figura 3.1). Entre eles, destacam-se[39]:

- modelo estrutural (*framework model*)[40] – o enovelamento da proteína inicia-se com a formação hierárquica de elementos da estrutura secundária independentemente da estrutura terciária, ou pelo menos antes que a estrutura terciária esteja em sua conformação definitiva. Estes elementos, por sua vez, se agrupam compactamente na estrutura terciária nativa por difusão e colisão, ou por propagação passo a passo;

- colapso hidrofóbico (*hydrophobic collapse*)[41] – o evento inicial da reação acontece com um relativo colapso uniforme da molécula da proteína, principalmente dirigido pelo efeito hidrofóbico. A estabilidade da estrutura secundária começa aumentar somente no estado colapsado, o que, embora possuindo já alguns dos contatos nativos, apresenta ainda uma conformação indefinida. O processo que finalmente levará o glóbulo colapsado à nativa, pode ser mais ou menos rápido, dependendo de seus atributos topológicos; e,

- mecanismo de nucleação-condensação (*nucleation-condensation mechanism*)[42] – a formação inicial de um núcleo difuso catalisa o enovelamento. O núcleo primário consiste de poucos resíduos adjacentes que possuem algumas interações corretas da estrutura secundária mas só permanece estável quando há alguma interação correta da estrutura terciária.

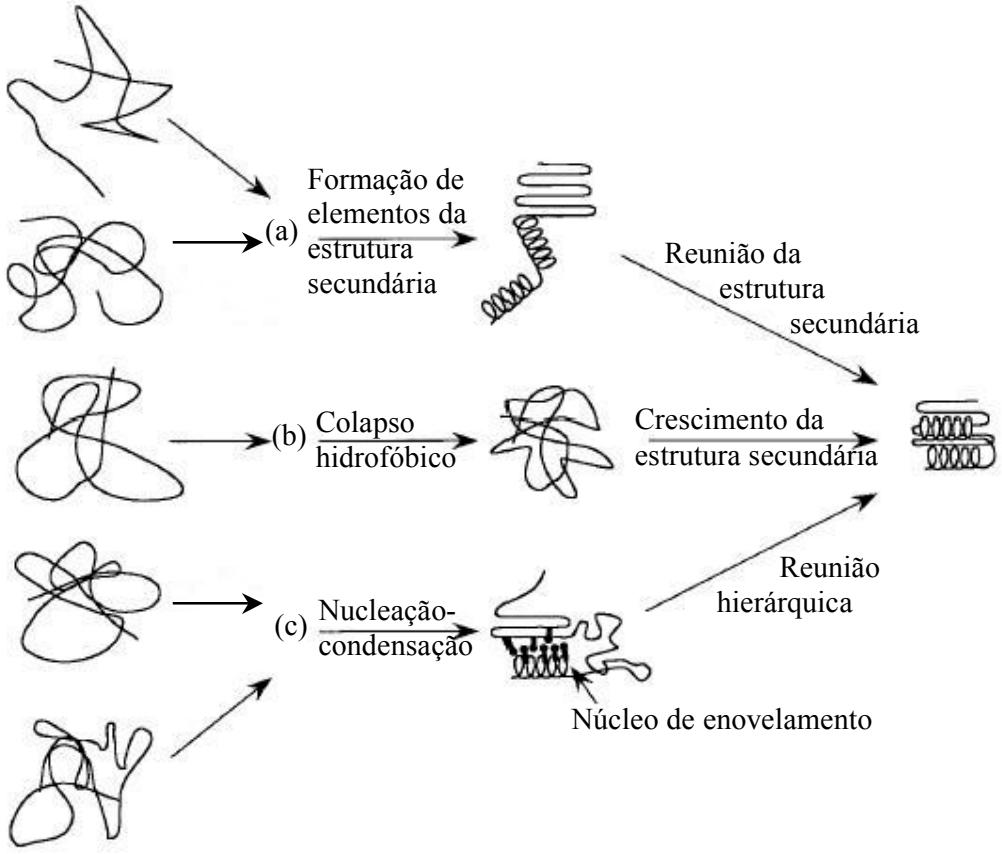


Figura 3.1. Modelos minimalistas. (a) modelo estrutural, (b) colapso hidrofóbico, (c) mecanismo de nucleação-condensação.

Estes modelos se distinguem, basicamente, pela identificação *a priori* de um determinado mecanismo. Evidentemente, os mesmos foram propostos baseados em observações experimentais que, devido à peculiaridade de cada experimento, enfatizam-se um ou outro mecanismo. Isto mostra, mais uma vez, a complexidade do problema do enovelamento protéico.

3.1. *Design* da proteína: especificidades químicas e estéricas

Uma proteína globular no estado funcional apresenta-se de forma altamente compacta, cuja densidade é comparável a de um sólido. Por isso, esta característica geométrica é usualmente representada nos modelos computacionais por uma estrutura maximamente compacta em uma rede regular. Apesar destas e outras simplificações adotadas, modelos minimalistas podem representar satisfatoriamente várias propriedades das proteínas reais, desde aquelas características seqüência-dependente, quanto outras relativas à estrutura nativa. Uma das propriedades das cadeias polipeptídicas que são seqüência-dependente, relaciona-se à degenerescência do estado de menor energia do sistema, que é representado pelo número de configurações distintas com a mesma energia mínima global[43].

Outra propriedade seqüência-dependente é descrita pelo conceito de *designability*, introduzido por Li e colaboradores[44], como sendo o número de seqüências distintas que têm como configuração nativa (a de menor energia potencial global) a mesma estrutura tridimensional compacta. A existência de estruturas com alto grau de *designability* sugere que as estruturas protéicas não são acidentais, mas uma conseqüência da seleção natural, pois mais *designability* pode conferir maior estabilidade termodinâmica e maior estabilidade contra eventuais mutações.

Os modelos minimalistas em rede têm sido utilizados com sucesso para mostrar que regularidades estruturais, do tipo das estruturas secundárias de proteínas reais, apresentam alta *designability*, e mais recentemente, têm sido empregados também para se estudar até que ponto certos parâmetros topológicos da estrutura nativa, como a ordem de contato relativo, se correlaciona com a taxa de enovelamento, principalmente para aquelas proteínas menores, e cuja cinética de enovelamento pode ser descrita como de dois estados (nativo e desnaturado). Com o aumento progressivo do número de

proteínas com suas estruturas conhecidas, torna-se cada vez mais claro a importância das interações daqueles monômeros próximos ao longo da cadeia para aumentar a rapidez do enovelamento.

O presente modelo emprega dezenas de estruturas com semelhanças e distinções topológicas. Por isso são caracterizadas todas as 51.704 configurações maximamente compactas de uma cadeia de 27 monômeros, não relacionadas por rotações, reflexões ou simetria de rotulação reversa[33]. A cadeia polipeptídica é representada por monômeros de igual tamanho, tem suas unidades situadas em vértices consecutivos de uma rede cúbica de dimensão infinita. Cada cadeia enovelada em uma das 51.704 possíveis configurações maximamente compactas auto excludentes (*Compact Self-Avoiding*, CSA) é considerada uma possível configuração nativa. Na configuração CSA, cada monômero assume um dentre três atributos topológicos: recebe *S* (*Straight*) se suas duas ligações covalentes estão alinhadas; recebe *T* (*Turn*) se suas duas ligações covalentes formam um ângulo reto; e recebe *E* (*End*) se o monômero estiver na posição final ou inicial da cadeia, conforme ilustra a Figura 3.2. No Capítulo seguinte, onde se tratará da classificação topológica das configurações CSA, estes atributos básicos serão retomados.

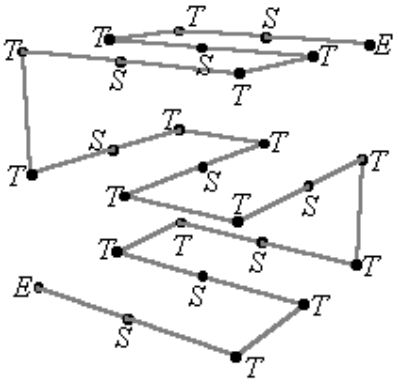


Figura 3.2. Atributos topológicos básicos dos monômeros, a saber, *S*, *T* e *E*, numa configuração CSA

As forças hidrofóbicas são adotadas como as forças dominantes no processo que direciona a cadeia para a conformação nativa. Assim, as especificidades químicas dos aminoácidos são representadas por um conjunto de dez diferentes unidades também denominado como “alfabeto de dez letras”, desenvolvido por R. da Silva e

colaboradores[10]. Através deste alfabeto foi desenvolvida uma sintaxe específica para se obter a seqüência de monômeros para cada estrutura nativa. Este alfabeto, embora tenha sido estabelecido heuristicamente através de vários experimentos computacionais, tem exatamente o tamanho do menor alfabeto necessário para reproduzir aspectos do processo de enovelamento, como determinado teoricamente por Fan e colaboradores[45].

No *design* da proteína, i.e., determinação da seqüência para cada estrutura alvo dada, os monômeros são identificados inicialmente pelos contatos que efetuam com o meio, definindo o grande conjunto $R(i)$, $i = 0, 1, 2, 3$, onde i identifica o número de contatos que o respectivo monômero numa configuração CSA faz com o meio solvente, imitando a superfície de exposição ao meio num sistema real. Cada uma das configurações CSA possui um único monômero $R(0)$ no centro do cubo $3\times3\times3$; seis monômeros $R(1)$ nos centros das faces; doze monômeros $R(2)$ nos meios das arestas; e oito monômeros $R(3)$ nos vértices do cubo $3\times3\times3$. A Figura 3.3 mostra o exemplo de uma estrutura maximamente compacta com a classificação por grupos e subgrupos hidrofóbicos. Os grupos $R(i)$ não discriminam uma configuração CSA de outra, ou seja, toda configuração CSA possui os mesmos monômeros $R(i)$, ou classes, embora em seqüências distintas. Cabe aos subgrupos $r(i,j)$ a tarefa de especificar os sub-níveis de hidrofobicidade. Note que os monômeros dos conjuntos $R(0)$ e $R(3)$ não possuem subgrupos, e para fins computacionais, são simbolizados, respectivamente, pelas letras *R* e *C*. Já os monômeros do conjunto $R(1)$ possuem dois subgrupos. O primeiro subgrupo é identificado por $r(1,1)$, e é aplicável sempre que o número de vizinhos topológicos na forma *S* é maior ou igual ao número de vizinhos topológicos na forma *T* (exceto o monômero do centro do cubo $R(0)$, o qual não entra na conta); por sua vez o subgrupo $r(1,2)$ é então determinado de forma automática. Na codificação do programa computacional, estes dois subgrupos são representados, respectivamente, pelas letras *A* e *H*. De forma análoga, os monômeros do conjunto $R(2)$, com seis subgrupos, serão identificados de forma hierárquica, como:

– $r(2,1)$, se os dois vizinhos topológicos são monômeros do tipo $r(1,1)$ e $r(1,1)$; representado pela letra *B*;
– $r(2,2)$, se os dois vizinhos topológicos são monômeros do tipo $r(1,1)$ e $r(1,2)$; representado pela letra *G*;

– $r(2,3)$, se os dois vizinhos topológicos são monômeros do tipo $r(1,1)$ e $R(3)$; representado pela letra *F*;

– $r(2,4)$, se 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*; representado pela letra *I*;

– $r(2,5)$, se os dois vizinhos topológicos são monômeros do tipo $r(1,2)$ e $R(3)$ na forma *E*; representado pela letra *E*;

– $r(2,6)$, se os dois vizinhos topológicos são monômeros do tipo $R(3)$; representado pela letra *D*.

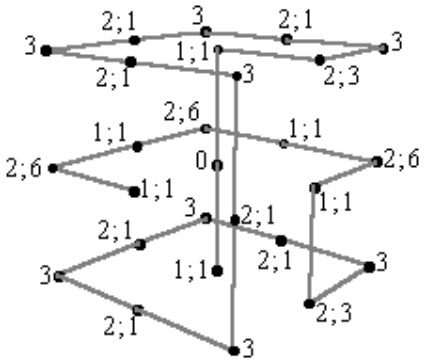


Figura 3.3. Ilustração dos grupos e subgrupos hidrofóbicos aplicados numa CSA específica.

Para o cálculo da energia de cada configuração CSA, é considerado o formalismo da energia de contato entre os monômeros *i* e *j*, pois no caso de uma rede regular e de densidade uniforme, existe uma identidade entre a energia de contato intra-cadeia e a energia hidrofóbica que envolve explicitamente a interação entre o solvente e os elementos da cadeia protéica. O tamanho da cadeia considerada é de $N = 27$ monômeros, dispostos sequencialmente nos sítios de uma rede cúbica, e assim, em cada configuração CSA possível, cada monômero em particular faz parte de seus contatos (primeiros vizinhos) com outros monômeros e o restante com o meio solvente. Todos os sítios da rede cúbica infinita são assumidos ocupados ou por monômeros da cadeia ou por moléculas do solvente (água)[46], de forma que no desenvolvimento da simulação, as posições espaciais ocupadas pelos elementos da cadeia e moléculas do solvente se intercambiam sistematicamente.

A variação da energia ΔE do sistema é determinada sempre que ocorre uma mudança configuracional, motivada pelas interações cadeia-solvente; todas as outras interações, i.e., solvente-solvente e monômero-monômero, são reduzidas a interações do tipo caroço duro (volume excluído). O potencial inter-resíduo é então estabelecido como interação de pares:

$$h_{i,j} = h_i + h_j \,, \tag{3.1}$$

onde h_i é o nível de hidrofobicidade do i-ésimo resíduo[47] ao longo da cadeia. O potencial hidrofóbico descrito por esta simplificação é eficiente para "encontrar" a estrutura nativa, mas não garante estabilidade configuracional ao sistema. Porém, adicionando-se ao potencial hidrofóbico um conjunto de restrições estéricas para pares (i, j) específicos, isto é,

$$e_{i,j} = h_i + h_j + c_{i,j} \tag{3.2}$$

foi possível verificar um grande aumento na estabilidade da cadeia na estrutura nativa[46].

Assim, o modelo utilizado neste trabalho é definido energeticamente pela equação 3.2. Porém, o conjunto $\{c_{i,j}\}$ desempenha um papel adicional importante que será detalhado em seguida. De fato, o conjunto de especificidades estéricas $\{c_{i,j}\}$ determina qual é o par de monômeros que podem fazer contato topológico de primeiros vizinhos e o que não pode, seja qual for a configuração na qual a cadeia se encontre ao longo da simulação; ver Tabela 3.1. De fato, este conjunto é composto de permissões e impedimentos estéricos que representam, de certa forma, a diversidade das especificidades estéricas encontradas no conjunto dos aminoácidos naturais, como diferentes formas e tamanhos. O conjunto $\{c_{i,j}\}$ exprime, assim, especificidades extras dos pares de aminoácidos, e seu principal efeito é o de selecionar caminhos através do espaço configuracional para enovelamento e desnaturação. Em outras palavras, as rotas de enovelamento para a configuração alvo (nativa) são significantemente reduzidas pelas restrições estéricas, principalmente com o glóbulo já mais compactado[11].

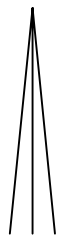
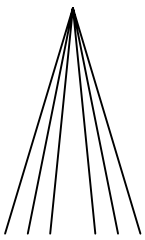
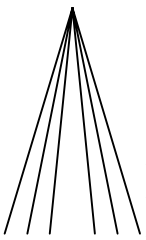
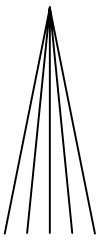
h	-2,1	-2,0	-1,9	-1,2	-1,1	-1,0	-1,0	-0,9	-0,8	+0,8
Especificidades estéricas	*R*	*A*	*H*	*B*	*G*	*F*	*I*	*E*	*D*	*C*
										
	R A H	*R A H B G F*	*R A H G I E*	*A*	*A H*	*A C*	*H C*	*H C*	*D C*	*F I E D C*
Classes	0	1		2						3

Tabela 3.1. Representação do conjunto das restrições estéricas e escala hidrofóbica: alfabeto de dez letras. Na linha rotulada por *h*, estão os níveis hidrofóbicos dos 10 tipos de monômeros identificados pelas letras *R*, *A*, *H*, *B*, *G*, *F*, *I*, *E*, *D* e *C*, na parte superior do campo das especificidades estéricas. Estes grupos, por sua vez, são identificados por linhas que conectam cada uma das letras superiores (tipos de monômeros) à um conjunto específico de outras letras. Como exemplo, o monômero designado pela letra *R* pode fazer contato de primeiro vizinhos com os monômeros designados por *R*, *A*, ou *H*. Já o monômero designado pela letra *B* somente poderá fazer contato com o monômero designado pela letra *A*. As especificidades dadas se referem somente às posições de primeiros vizinhos.

Assim, a energia configuracional é dependente somente da escala hidrofóbica, pois os termos $\{c_{i,j}\}$ efetivamente contribuem para a energia com "zero" (permissão) ou "infinito" (proibição). Uma sub-rotina específica (tenl) desenvolvida neste trabalho, sequencia automaticamente a cadeia, para cada uma das 51.704 configurações CSA. Assim, a energia hidrofóbica de determinada configuração é formalmente escrita como:

$$E\{(k,l)\} = \sum_{\{i,j\}} (h_{i,j} + c_{i,j}).\delta_{(i,j),\{k,l\}} , \qquad (3.3)$$

onde a soma ocorre sobre todos os pares $\{i,j\}$; o fator $\delta_{(i,j),\{k,l\}} = 1$, se o par (i,j) pertence ao conjunto $\{k,l\}$ de todos os pares em contato e $\delta = 0$, de outra forma.

3.2. Simetrias do cubo: enumeração das CSA

Cada configuração CSA pode ser considerada como um caminho Hamiltoniano em um grafo, ou seja, é um caminho que passa por todos os vértices (sítios) deste grafo uma única vez, sem cruzar os caminhos. Na rede cúbica e para uma cadeia de 27 monômeros, todo caminho Hamiltoniano começa por um vértice ou pelo centro de qualquer face do cubo. A prova deste lema foi comunicada oralmente a Shahknovich e Gutin[48], que, em 1990, fizeram a enumeração exaustiva de todas as CSA para uma cadeia com 27 monômeros, e encontraram que o número total de conformações não relacionadas por simetria é 103.346. Em 1995, dos Reis[33] mostra de forma explícita esta totalização. Utiliza um cubo numerado (Figura 3.4) para introduzir a matriz de conectividade, de tamanho 27×27, de primeiros vizinhos, onde todas as informações de interações estão explícitas. Utilizando o lema mencionado, foram escolhidos os monômeros de número 1 e 11 para encontrar os caminhos iniciais sem as restrições das simetrias. Na seqüência, compôs-se o grupo de simetrias do cubo, num total de 48 elementos, que inclui as rotações, reflexões e rotações impróprias, que são as combinações de rotações e reflexões. Dessa forma, foi possível construir dois grandes arquivos de dados: um com as coordenadas genéricas de um cubo, como mostrado na Figura 3.4, e outro com a seqüência de conectividade destes monômeros, que montam as 103.346 configurações CSA.

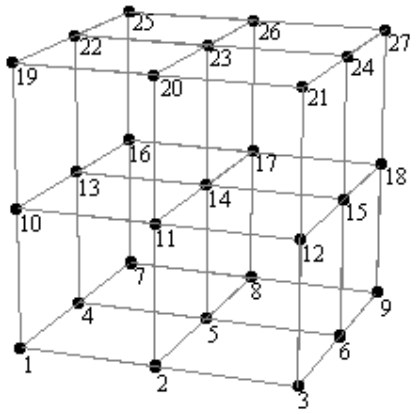


Figura 3.4. Cubo 3×3×3 com seus sítios (27) identificados por inteiros consecutivos.

3.3. Programa computacional implementado

Uma das finalidades do programa computacional inicialmente desenvolvido por dos Reis[33] era a de especificar as coordenadas de todas as (103.346) CSA, não relacionadas pelas operações de simetrias do cubo, pois as configurações que se obtém por estas simetrias não se diferenciam fisicamente. Na simulação do processo de enovelamento foram empregados os movimentos de canto, manivela e final de cadeia, e no modelo empregado usava-se um potencial randômico para interações intra-cadeia. Posteriormente, o programa computacional foi grandemente generalizado por R. da Silva[47], mas ainda utilizava um arquivo externo, composto à mão, cujos dados de entrada continham coordenada e contatos nativos de cada CSA, as sementes para cada simulação Monte Carlo, a seqüência de letras da configuração CSA e as restrições estéricas entre os monômeros.

Em 2003, com a colaboração do Dr. Marco Antonio Alves da Silva, as CSA não relacionadas pela simetria de rotulação reversa[44] (62 casos) foram então eliminadas, totalizando 51.704 CSA ($103.346 = 2 \times 51.704 - 62$). Uma implementação do programa "tempoMCS.f" foi realizada para que, fornecendo apenas o número da seqüência da CSA desejada, ou um arquivo de números de CSA desejadas, mais o arquivo das coordenadas genéricas dos monômeros do fragmento de rede de tamanho $3\times3\times3$ ("coord.dat") e o arquivo com a seqüência de uniões dos monômeros ("cfg51704.dat"), as seguintes etapas são performadas automaticamente:

- Lê o arquivo das restrições entre os monômeros, "repulsao.dat";
- Lê o arquivo das hidrofobicidades (letras do alfabeto utilizado), "hidrof.dat";
- Gera, à partir de uma sub-rotina (gsnr), e guarda em um vetor (isemt) as sementes necessárias para efetuar a quantidade de simulações Monte Carlo independentes desejadas;
- Monta, em uma sub-rotina (natcont), o vetor dos contatos nativos (kvntv) de cada CSA desejada; e finalmente,
- Performa o *design* da proteína, ou seja, compõe, utilizando outra sub-rotina (tenl), o vetor da seqüência de letras (itl) de cada CSA;

Durante a determinação da seqüência de letras para cada CSA, no programa computacional implementado, os monômeros dos vértices do cubo recebem uma entre duas letras, ao invés de somente a letra *C*, como mostrado no artigo de R. da Silva e colaboradores[10] e na Tabela 3.1. Com esta modificação, se o monômero na posição do vértice é final de cadeia, recebe a letra *M*, caso contrário, recebe a letra *C*. Apesar desta alteração, originalmente introduzida também por R. da Silva[47], a hidrofobicidade destas duas letras é a mesma. Há pequena modificação nos contatos permitidos, anteriormente mostrado na Tabela 3.1. O monômero correspondente à letra *C* pode fazer contato topológico de primeiro vizinho com os monômeros *F*, *D*, *I* e *C*, enquanto que o correspondente à letra *M* pode fazer contatos com os *F*, *E*, *D* e *C* (note a única troca de *I* por *E*). Qual o efeito desta troca? O monômero *I* é mais hidrofóbico que o *E*, e assim o monômero do extremo da cadeia fica um pouco mais "livre".

3.4. Ordem de Contato relativo

O parâmetro topológico global, denominado "ordem de contato relativo", χ, foi inicialmente introduzido por Plaxco e colaboradores[49], e tem sido considerado um atributo configuracional importante no estudo da taxa de enovelamento protéico. Este parâmetro mede o grau de contato intracadeia da estrutura nativa, isto é, a distância média entre todos os pares (*i*,*j*) de resíduos contactantes, normalizada pelo comprimento *L* da seqüência:

$$\chi = \frac{1}{N_c} \cdot \sum_{contatos} \frac{\Delta l_{i,j}}{L} \tag{3.4}$$

onde $\Delta l_{i,j}$ é a separação ao longo da cadeia entre os resíduos contactantes (*i*,*j*). A soma é sobre todos os pares (*i*,*j*) contactantes e N_c é o número total de contatos da estrutura nativa. Em geral, tem sido considerado que dois resíduos distintos estão em contato quando a menor distância entre seus átomos pesados for menor que λ = 6 Å. Contudo, outros valores de λ têm sido usados, como 5,5 e 7,5 Å[50].

A ordem de contato relativo χ do conjunto das 51.704 estruturas analisadas se distribui no intervalo $0{,}238095 < \chi < 0{,}494709$: há somente 97 valores diferentes de

ordem de contato, que correspondem a diferentes múltiplos de $2/(27\times28)$, i.e., $\chi = (2.n)/(27\times28)$, com n = 90, 91, ..., 187 e $n \neq 185$. A curva em forma de sino, Gráfico 3.1, mostra que há uma pequena fração de configurações que compõem suas caudas. Esta propriedade da distribuição está em harmonia com o fato de as proteínas terem χ pequeno, e possuírem muita semelhança nos padrões de enovelamento.

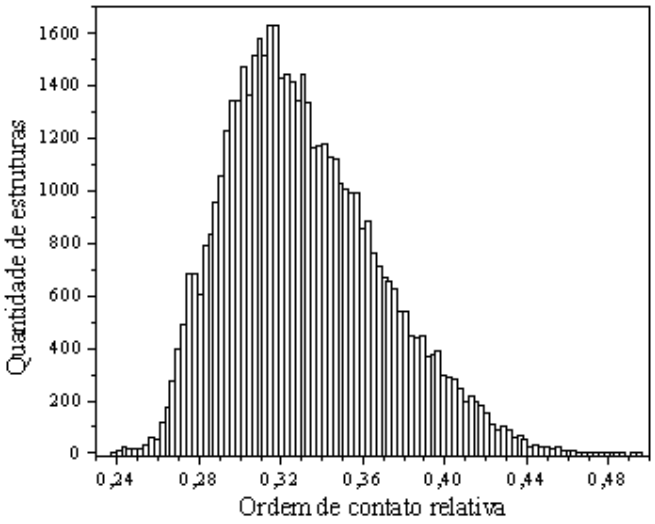


Gráfico 3.1. Distribuição dos 97 valores de ordem de contato entre as 51.704 estruturas analisadas. Há grande concentração de estruturas ao redor de $\chi \approx 0{,}31$ e poucas estruturas nas caudas.

Várias generalizações para o cálculo da ordem de contato relativo têm sido feitas e confirmam que o logaritmo da taxa de enovelamento de proteínas depende fortemente dos parâmetros topológicos estruturais. Contudo, trabalhos experimentais mostram que mutações afetam a taxa de enovelamento em até três ordens de magnitude[51], indicando que a topologia não é o único determinante da taxa de enovelamento. Adicionalmente, sabe-se que muitas proteínas com mesma estrutura topológica enovelam com diferentes valores de taxa de enovelamento, enquanto que proteínas de estrutura topológica distinta enovelam com valores de taxa de enovelamento similares[52, 53, 54, 55], e assim estas variantes precisam ainda ser explicadas.

3.5. Taxa de enovelamento

Com o progressivo aumento no número de proteínas estudadas, tem sido possível verificar que proteínas pequenas (do tipo dois estados termodinâmicos), em solução, apresentam taxa de enovelamento log k_f com variação de muitas ordens de magnitude conforme o valor da ordem de contato relativo χ varia de 5 a ~30%[56, 57, 58]. Os modelos simplificados têm sido também utilizados para investigar a relação entre parâmetros topológicos da estrutura nativa e a taxa de enovelamento (log k_f) de pequenas proteínas, cujo comportamento cinético pode ser aproximado pela dinâmica de sistemas com somente dois estados termodinâmicos[49, 59], como observado experimentalmente[51, 60, 61].

A abordagem deste problema, mesmo por modelos minimalistas, não é imediata, pois em geral se necessita uma regra para o *design* da proteína. Apesar de que para modelos que usam alfabeto de duas letras, como modelo HP[35] (hidrofóbico-polar) e modelo de Gō[62, 63], o *design* seja inerente, tais modelos não apresentam uma correlação consistente entre a taxa de enovelamento (log k_f) e a ordem de contato relativo (χ), assim como outros mais sofisticados e com alfabetos maiores mas que utilizam esquemas simples de interação de pares[64, 65, 66]. Na intenção de investigar esta correlação, um certo grau de cooperatividade no esquema das interações tem sido proposto com sucesso[51, 67] em modelos nativa-dirigidos.

No caso deste trabalho, o modelo utilizado não é nativa-dirigido e possui um alfabeto de dez letras, impondo, assim, a existência de uma regra para o *design* da seqüência. A taxa de enovelamento foi obtida através da simulação do método Monte Carlo. Para cada estrutura nativa foram efetuadas 15 simulações independentes usando uma janela de tempo $t_w = 3\times10^7$ passos MC. Para cada simulação, o tempo t_i necessário para encontrar a estrutura nativa pela primeira vez[67] foi anotado. A simulação do enovelamento é tida como de sucesso se $t_i \leq t_w$. Foi utilizado um conjunto inicial de 52 estruturas nativas escolhidas de forma a cobrir todo o range dos possíveis valores de χ e fornecer uma rica variedade de padrões estruturais. Adicionalmente, várias dezenas de outras estruturas nativas foram simuladas com o objetivo de validar as proposições deste trabalho e suas conclusões.

A taxa de enovelamento tem sido representada pelo inverso do tempo médio requerido para a primeira passagem pela nativa (FPT – *first passage time*) em diversos trabalhos[51, 66, 67]. Contudo, neste trabalho, utilizou-se para a estimativa da taxa de enovelamento, a média geométrica do tempo para encontrar a nativa, pois algumas estruturas nativas têm o valor de t_i variando até duas ordens de magnitude, o que contrasta com o pequeno tamanho do conjunto $\{t_i\}$ empregado aqui. A taxa de enovelamento log k_f, representada pelo recíproco da média aritmética do tempo, isto é, $1/\langle t \rangle$, é obtida de:

$$\langle t \rangle = (1/N)\sum_{i=1}^{N} t_i \,. \tag{3.5}$$

De fato, para proteínas pequenas, a distribuição dos tempos de enovelamento pode ser aproximada por uma exponencial, $P(t) \approx \exp(-t/t_0)$. Para um conjunto $\{t_i\}$ grande suficiente, $\langle t \rangle$ pode ser aproximado pela média aritmética, a qual representa bem o cálculo rigoroso, dado por:

$$\langle t \rangle = \int_0^{\infty} tP(t)dt = t_0 \,, \tag{3.6}$$

com $\int P(t)dt = 1$. O logaritmo comum (base 10) da taxa de enovelamento é então obtida de $1/\langle t \rangle$, ou seja,

$$\log(k_f) = \log(1/\langle t \rangle) = -\log\langle t \rangle \,. \tag{3.7}$$

Pode ser utilizada, com vantagem, uma propriedade da média de ln(t) para as distribuições exponenciais, para o caso de um pequeno conjunto de simulações efetuadas (de fato um total de 15 simulações independentes apenas foram efetuadas para cada um dos 52 casos). Esta possibilidade vem do fato de que $\langle \ln(t) \rangle$ está linearmente relacionada com $\ln(\langle t \rangle)$, isto é:

$$\langle \ln(t) \rangle = \int_0^\infty \ln(t) P(t) dt = \ln(\langle t \rangle) - \gamma \,, \tag{3.8}$$

que é diretamente verificado usando o resultado: $\int_0^\infty \ln(t) \exp(-t) dt = -\gamma$, onde $\gamma \approx 0{,}5772$ é a constante de Euler. Logo, $\log(\langle t \rangle) = \langle \log(t) \rangle + \gamma . \log(e)$, onde $\log(e) \approx 0{,}4343$. Esta propriedade é particularmente útil para pequenas amostras. O efeito da dispersão, sempre presente em pequenas amostras, é então minimizado efetuando-se a média de $\log(t_i)$, ao invés de $\log(\langle t \rangle)$. A função logarítmica aplicada no conjunto de simulações reduz o efeito das flutuações, sempre presentes no pequeno conjunto de amostras. Assim, $\log k_f$ é representado neste trabalho por:

$$\log k_f = \frac{1}{15} . \sum_{i=1}^{15} \log t_i^{-1} \,, \tag{3.9}$$

isto é, a média geométrica utilizada segue de uma propriedade matemática imediata:

$$\langle \ln(t) \rangle = (1/N)(\ln(t_1) + \ln(t_2) + ... + \ln(t_N)) = (1/N) \ln(t_1 . t_2 ... t_N) \,, \tag{3.10}$$

e então:

$$\langle \ln(t) \rangle = \ln(t_1 . t_2 ... t_N)^{(1/N)} = \ln(\langle t \rangle_g) \,. \tag{3.11}$$

De forma a validar a aproximação usada para representar a taxa de enovelamento k_f, obtida através da equação 3.9 acima, o histograma do número cumulativo dependente do tempo $N(t)$ de proteínas enoveladas foi feito, utilizando-se a expressão:

$$N(t) = N_0 . (1 - \exp(-t / t_0)) \,, \tag{3.12}$$

com $N_0 = 15$. O número característico t_0, estimado através do ajuste do mesmo na equação acima, para cada estrutura que apresentou sucesso de 100% na taxa de enovelamento – a maioria delas; foi então utilizado para determinar a discrepância entre

$\log(t_0^{-1})$ e o correspondente $\log k_f$ obtido pela equação 3.9. Os casos nos quais o método do histograma não pode ser usado incluem aquelas estruturas nativas que apresentam sucesso de enovelamento menor que 66% e aquelas com espectro de taxa de enovelamento muito amplo (alguns com mais do que duas ordens de magnitude).