

## 1. INTRODUÇÃO

### 1.1 Motivação

As proteínas compõem uma das classes de moléculas biológicas mais estudadas atualmente<sup>†</sup>. Isto é devido, entre várias outras razões, às descobertas das últimas décadas que mostram um número crescente de doenças identificadas como resultantes de falhas no enovelamento, ou *folding*, das proteínas. Um esforço mundial é devotado no elucidamento dos “mistérios” que envolvem as proteínas, como bem ilustra a criação em 1971 de um depósito mundial<sup>1</sup> para processamento e distribuição de estruturas tridimensionais destas macromoléculas biológicas, o *Protein Data Bank* (PDB – [www.pdb.org](http://www.pdb.org)). Outro empreendimento internacional relacionado às proteínas e iniciado formalmente em 1990 é o Projeto Genoma, que tem como objetivos identificar e mapear os genes existentes no DNA (ácido desoxirribonucléico) do ser humano, determinar suas seqüências de bases químicas e armazenar estas informações em bancos de dados e torná-las acessíveis para novas pesquisas.

O marco inicial do estudo científico das proteínas data de 1838, quando Gerardus Johannes Mulder, químico holandês, descobriu que algumas substâncias orgânicas tinham em comum um comportamento estranho: quando aquecidas, ao contrário das outras substâncias, mudavam do estado líquido para o estado sólido (como: a clara do ovo; a caseína, substância do leite; e a globulina, um componente do sangue). Descobriu também que todas estas substâncias continham carbono, hidrogênio,

---

<sup>†</sup> Este fato pode ser evidenciado quando comparada a quantidade de *sites* da *Internet* relacionados com proteínas em relação a outras moléculas biológicas, como DNA, vitaminas, carboidratos, etc, numa pesquisa em [www.google.com.br](http://www.google.com.br). As entradas, datadas de 05/2005, para cada uma das seguintes moléculas eram: proteínas:  $33,7 \times 10^6$ ; DNA:  $40,6 \times 10^6$ ; vitaminas:  $9,41 \times 10^6$ ; carboidratos:  $3,49 \times 10^6$ . Os interesses pelas proteínas abrangem desde estudos acadêmicos, teóricos e práticos, até interesses médicos/farmacológicos, nutricionais, cosméticos e industriais.

nitrogênio e oxigênio. Denominou então esta “fórmula” de “proteína”<sup>2</sup> (proteína, do grego *porteía* = primazia, do latim *primatia* = primeiro plano<sup>3</sup>).

As proteínas têm sido classificadas em três grandes grupos: Proteínas Estruturais, Proteínas de Membrana e Proteínas Globulares. As proteínas estruturais, ou fibrosas, são constituídas de feixes ou placas de longas cadeias lineares de aminoácidos<sup>††</sup>, formando micro filamentos e micro tubos, que constituem a matéria prima, por exemplo dos cabelos e unhas, e estão presentes também nos ossos. Já as proteínas de membrana encontram-se ancoradas na membrana celular, ou as interpenetram (uma ou mais vezes). Parte das funções deste grupo de proteínas inclui a constituição de canais na membrana celular, pois são responsáveis pela intermediação de substâncias que entram e saem da célula, são transportadoras de produtos específicos através da membrana, agem como receptoras de substâncias importantes para a função celular, atuam como enzimas para catalisar reações na superfície da membrana e agem como marcadoras das células provenientes de outros organismos<sup>4</sup>. E finalmente, as proteínas globulares, as mais estudadas, constituem a quase totalidade das estruturas espaciais conhecidas e depositadas no PDB, e é o objeto central deste trabalho.

As proteínas globulares desempenham funções diversificadas no organismo: ação enzimática (catalisador biológico), de transporte (exemplo: hemoglobina, mioglobina), função reguladora (hormônios) e como fator de crescimento FGF (*Fibroblast Growth Factor*)<sup>5</sup>. Funcionam também como anticorpos contra antígenos externos, atuam na coagulação e na produção de energia e fazem parte do material cromossômico. Assim, a forma, a regulação, preservação e a reprodução dos seres vivos são controladas pelas proteínas globulares.

Fundamentalmente, todas as proteínas naturais são macromoléculas lineares formadas a partir de um repertório de 20 aminoácidos, os chamados aminoácidos naturais (Figura 1.1). A seqüência de aminoácidos que constitui uma proteína é chamada estrutura primária<sup>4</sup>, e um dos grandes desafios atuais da Biologia e áreas científicas relacionadas, é relacionar a seqüência de aminoácidos de uma proteína com a

---

<sup>††</sup> As palavras aminoácido, resíduo, unidade e monômero serão utilizadas neste estudo como sinônimos, dependendo do contexto.

sua estrutura tridimensional (3-D). Este problema é conhecido como “problema do enovelamento de proteína”. Usualmente, duas abordagens distintas são utilizadas: (i) Enovelamento direto: dada uma seqüência de aminoácidos, a tarefa é então prever a sua única e precisa estrutura 3-D; (ii) Enovelamento inverso: dada uma estrutura 3-D, a tarefa agora é encontrar uma seqüência de aminoácidos que leve a cadeia conformar-se e estabilizar-se naquela estrutura. Na primeira abordagem a idéia é principalmente a de “resolver” o problema, isto é, dada a seqüência de aminoácidos da proteína, prever a sua estrutura nativa<sup>†††</sup>; para isso, todo e qualquer método disponível é bem vindo. Mas para a segunda abordagem do problema, utilizada neste trabalho, “entender” o mecanismo do *folding* é pré-requisito indispensável. Como nos dias atuais é muito mais fácil, tecnologicamente, sequenciar uma proteína do que determinar sua estrutura (Raios-X, NMR – Ressonância Magnética Nuclear), uma enorme quantidade de estruturas a serem determinadas está se acumulando. Cerca de somente 10% das proteínas já seqüenciadas têm suas estruturas determinadas. Por isso, o desenvolvimento teórico para se entender o processo, e métodos computacionais de determinação estrutural, são prementes.

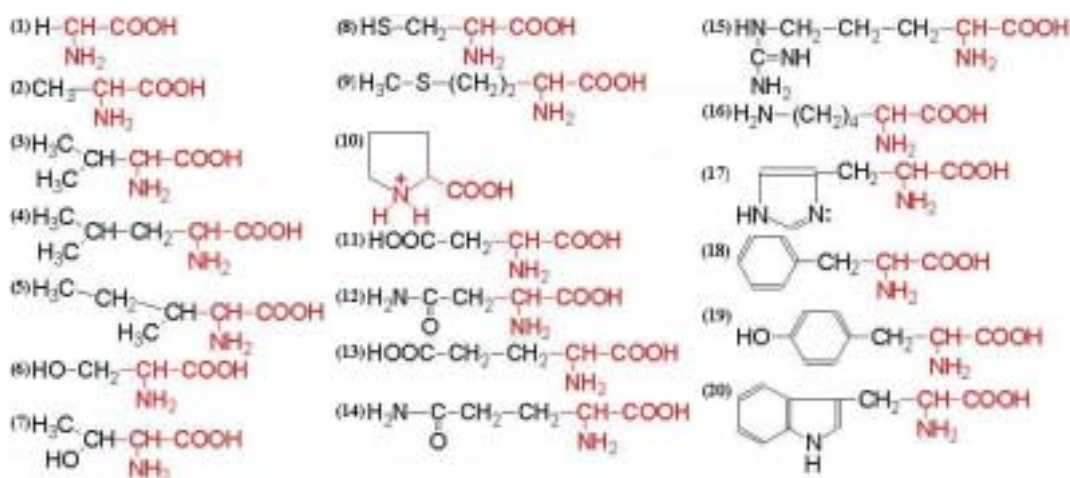


Figura 1.1. Os 20 aminoácidos naturais das proteínas: (1) Glicina, (2) Alanina, (3) Valina, (4) Leucina, (5) Isoleucina, (6) Serina, (7) Treonina, (8) Cisteína, (9) Metionina, (10) Prolina, (11) Aspártico, (12) Asparagina, (13) Glutâmico, (14) Glutamina, (15) Arginina, (16) Lisina, (17) Histidina, (18) Fenilalanina, (19) Tirosina, (20) Triptofano.

<sup>†††</sup> No presente contexto, a estrutura chamada nativa corresponde a uma conformação espacial da cadeia que inclui a configuração de menor energia potencial.

Particularmente, o problema do enovelamento inverso é de muito interesse médico e farmacológico, principalmente porque a função de uma proteína é dependente de sua conformação estrutural. De fato, uma grande motivação para se estudar o enovelamento de proteínas é a possibilidade do desenvolvimento e produção de novos medicamentos. Estudos recentes revelam um número crescente de doenças que resultam de falhas no enovelamento das proteínas<sup>6</sup>. Estas falhas podem provocar um déficit funcional e levar a sérias conseqüências, como ocorre na anemia falciforme. Erros no enovelamento também podem provocar o agregamento das proteínas causando doenças neuro-degenerativas, como Mal de Alzheimer e mal de Parkinson, diabetes tipo II, bem como doenças raras, como a Amilóide Polineuropática Familiar. Inclui-se também entre as doenças causadas por falhas no enovelamento, a doença da vaca louca (Encefalopatia Espongiforme Bovina–BSE), certo tipo de enfisema pulmonar, alguns tipos de câncer<sup>6,7</sup>.

## 1.2 Abordagem do problema

Uma das grandes dificuldades no tratamento científico do problema do enovelamento de proteínas reside no fato da irredutibilidade do sistema cadeia-solvente. De fato, múltiplos ingredientes estão envolvidos, como interações químicas intra-cadeia e cadeia-solvente, interações estéricas (devido a formas e tamanhos distintos dos aminoácidos) e questões da unicidade conformacional da estrutura nativa. Assim, devido à complexidade envolvida no processo do enovelamento das proteínas, os modelos minimalistas têm sido um importante recurso, pois a principal limitação atual no estudo de sistemas protéicos<sup>††††</sup> é o tempo requerido nas técnicas de simulação computacional, como dinâmica molecular, principalmente quando da inclusão dos detalhes do envolvimento da proteína com o solvente. Tais modelos minimalistas também são capazes de reproduzir aspectos característicos do processo do enovelamento das proteínas, como o tempo de enovelamento, identificar os caminhos para a conformação nativa<sup>8</sup>, e descrever propriedades termodinâmicas com detalhes<sup>9</sup>. Este trabalho utiliza um modelo computacional simplificado (modelo minimalista) para estudar o processo de enovelamento da proteína, através da representação da cadeia

---

<sup>††††</sup> Há divergência quanto à acentuação desta palavra pelos diversos autores. Neste trabalho esta palavra será utilizada com acento agudo e pronunciada com ‘e’ aberto, segundo a orientação descrita em CEGALLA, D. P. Dicionário de dificuldades da língua Portuguesa. Ed. Nova Fronteira, 2<sup>a</sup>. ed., Rio de Janeiro: 1999.

polipeptídica de 27 aminoácidos, por meio de “27 contas” conectadas linearmente e restritas a ocuparem 27 sítios exclusivos de uma rede cúbica. As estruturas nativas são representadas por configurações maximamente compactas (*Compact Self-Avoiding* – CSA) e as interações intra-cadeia são definidas por um modelo estéreo-químico<sup>10</sup>, que combina um conjunto de especificidades estéricas com energia hidrofóbica<sup>11</sup> (detalhes no Capítulo 3). Este modelo é tratado por simulação computacional utilizando um programa em código Fortran implementado com diversas sub-rotinas especificamente desenvolvidas.

### 1.3 Especificando o problema deste trabalho

Uma das questões atuais do problema do enovelamento protéico é identificar os fatores determinantes da taxa de enovelamento, ou seja, entender as razões que determinam as diferentes “velocidades” com que as proteínas se enovelam. Resultados experimentais referentes a pequenas proteínas revelam que a taxa de enovelamento  $k_f$  se correlaciona com parâmetros geométricos globais de suas respectivas conformações nativas, como por exemplo, a ordem de contato relativo  $\chi$ . Assim, o objetivo principal deste trabalho é identificar parâmetros topológicos da estrutura nativa que são determinantes da cinética do enovelamento de proteínas globulares. Para isto, procurou-se: (i) caracterizar as estruturas maximamente compactas em relação a certos parâmetros topológicos relevantes da estrutura nativa; (ii) verificar a influência das especificidades estéricas no processo do enovelamento de proteínas; (iii) estudar a correlação entre parâmetros estruturais globais, como ordem de contato relativo e a taxa de enovelamento; (iv) estudar a correlação entre as características topológicas específicas da estrutura nativa e a cinética do processo de enovelamento.

No Capítulo 2 são apresentados os métodos de simulação molecular computacional mais populares, a saber, Dinâmica Molecular e Monte Carlo. São também apresentadas as características básicas do programa computacional desenvolvido e utilizado neste trabalho, para simulação do processo de enovelamento de proteínas. O Capítulo 3 faz uma breve recapitulação dos principais modelos em rede estudados e descreve os mecanismos de enovelamento utilizados neste trabalho. São apresentados detalhes sobre as restrições estéricas, o conceito da ordem de contato

relativo e a forma empregada de cálculo da taxa de enovelamento das proteínas. Introduzem-se, no Capítulo 4, os diversos parâmetros topológicos estudados para o modelo em rede utilizado, entre eles, elementos topológicos básicos, tipos possíveis de extremidades da cadeia, padrões estruturais (certas combinações de elementos topológicos básicos que lembram hélices, *loops*, etc, de proteínas reais), o cálculo da energia da estrutura nativa.

O Capítulo 5 trata de uma análise detalhada da influência dos atributos topológicos das configurações nativas na cinética do enovelamento, levando a concluir que a taxa de enovelamento é fortemente dependente do conteúdo de padrões estruturais tipo-secundárias da estrutura nativa. O interessante é que esta não depende propriamente do valor da ordem de contato relativo e é muito influenciada pelos padrões configuracionais componentes da configuração nativa e suas combinações. A dependência linear de  $\log k_f$  com  $\chi$  é determinada por aquelas configurações em que há uma quantidade equilibrada de padrões estruturais que mesclam contatos efetivos de curto alcance com outros de longo alcance. E neste caso, o conteúdo de estruturas tipo-secundárias da nativa e o seu correspondente valor de  $\chi$  são equivalentes. Porém, 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. Também neste Capítulo é abordada a questão do mecanismo físico básico que relaciona o conteúdo de estruturas tipo-secundárias e a taxa de enovelamento, através do conceito de cooperatividade.

Na contracapa deste trabalho encontra-se um *CD rom* com vários documentos: *i*–cópia desta tese (.pdf), *ii*–arquivo com a seqüência dos monômeros das 51.704 configurações CSA (.pdf), *iii*–arquivo com as posições padrões no cubo em rede para os 27 monômeros (.dat), *iv*–arquivo com a relação dos monômeros que podem ser vizinhos topológicos (dependendo das restrições estéricas) (.dat), *v*–arquivo com os valores da escala de hidrofobicidades (.dat), *vi*–tabela-resumo das diversas estruturas secundárias analisadas contendo dados topológicos energéticos e cinéticos (.pdf), *vii*–programa computacional utilizado para o cálculo do tempo do enovelamento de cada configuração alvo (com histórico de seu desenvolvimento; .pdf).