UNIVERSIDADE FEDERAL DE PERNAMBUCO
DEPARTAMENTO DE FÍSICA – CCEN
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

# TESE DE DOUTORADO

## ASPECTOS ESPACIAIS E TEMPORAIS DO PROBLEMA DO ENOVELAMENTO PROTÉICO

por

**Pedro Hugo de Figueirêdo**

Tese apresentada ao Programa de Pós-Graduação em Física do Departamento de Física da Universidade Federal de Pernambuco como parte dos requisitos para obtenção do título de Doutor em Física.

**Banca Examinadora:**

Prof. Sérgio Galvão Coutinho (Orientador-UFPE)
Prof. Edvaldo Nogueira Júnior (Co-orientador - UFBA)
Prof. Marcelo Albano Moret S. Gonçalves (Co-orientador - FVC)
Prof. Paulo Mascarello Bisch (IBCCF - UFRJ)
Prof. Jerson Lima Silva (IBM - UFRJ)
Prof. Marcelo Andrade de Filgueiras Gomes (DF-UFPE)
Profª. Rita Maria Zorzenon dos Santos (DF – UFPE)

Recife - PE, Brasil
Setembro – 2006

Figueirêdo, Pedro Hugo de
Aspectos espaciais e temporais do problema do enovelamento protéico / Pedro Hugo de Figueirêdo. – Recife : O autor, 2006.
xii, 124 folhas : il., fig., tab.

Tese (doutorado) – Universidade Federal de Pernambuco. CCEN. Física, 2006.

Inclui bibliografia e apêndice.

1. Mecânica estatística. 2. Enovelamento protéico. 3. Caminhantes aleatórios. 4. Séries temporais 5. Multifractais I. Título.

530.13 CDD (22.ed.) FQ2006-0019

Universidade Federal de Pernambuco
Departamento de Física – CCEN
Programa de Pós-Graduação em Física
Cidade Universitária - 50670-901 Recife PE Brasil
Fone (++ 55 81) 2126-8449/2126-8450 - Fax (++ 55 81) 3271-0359
http://www.df.ufpe.br/pg e-mail: posgrad@df.ufpe.br

## Parecer da Banca Examinadora de Defesa de Tese de Doutorado

**Pedro Hugo de Figueirêdo**

**ASPECTOS ESPACIAIS E TEMPORAIS DO PROBLEMA DO ENOVELAMENTO PROTÉICO**

A Banca Examinadora composta pelos Professores Sérgio Galvão Coutinho (Presidente e Orientador), Marcelo Andrade de F. Gomes, Rita Maria Zorzenon dos Santos, todos da Universidade Federal de Pernambuco, Edvaldo Nogueira Júnior (Co-orientador), da Universidade Federal da Bahia, Marcelo Albano Moret S. Gonçalves (Co-orientador), da Fundação Visconde de Cairu, Paulo Mascarello Bisch, da Universidade Federal do Rio de Janeiro e Jerson Lima Silva, da Universidade Federal do Rio de Janeiro, consideram o candidato:

( ) Aprovado com Distinção (X) Aprovado ( ) Reprovado

Secretaria do Programa de Pós-Graduação em Física do Departamento de Física do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco aos quinze dias do mês de setembro de 2006.

Prof. Sergio Galvão Coutinho
Presidente e Orientador

Prof. Marcelo A. Moret S. Gonçalves
(Co-orientador)

Prof. Edvaldo Nogueira Júnior
(Co-orientador)

Profª. Rita Maria Zorzenon dos Santos

Prof. Marcelo Andrade de F. Gomes

Prof. Paulo Mascarello Bisch

Prof. Jerson Lima Silva

*Em comunhão com aqueles que me ajudaram a desvencilhar os novelos da vida.*

# Agradecimentos

Certo dia ao longo de uma das crises acerca de nossos investimentos na carreira de físicos um amigo perguntou-me:

- Pedro, o que será que nos lembraríamos daqui a dez anos se abandonassemos a física ?

Ele me questionou se saberíamos por exemplo o que seriam bósons e férmions[1]. Terminamos a conversa rindo e dizendo que afirmaríamos que um era um vírus e outro era uma bactéria, mas não saberíamos identificar qual era o que. Hoje passados quase oito anos dessa conversa olho para trás e vejo um rastro formado por livros e artigos lidos, por listas de exercícios e exames cumpridos. Também é possível vislumbrar companheiros de farras, noites em praias e violões, comidas saborosas e discussões acaloradas. De certo é necessário que se registrem aqueles que contribuiram para que o estudante secundário que ingressou nessa instituição a dez anos atrás, se convertesse nesse candidato a doutor, e se eles não o tornaram um melhor físico, definitivamente foram decisivas em sua formação como um ser humano melhor.

Embora a enumeração possa ser perigosa, arrisco-me a citar alguns nomes mesmo que a memória me traia. Agradeço aos meus pais Maria e João, e a minha irmã Elisa que durante anos investiram paciência e apostaram em minha formação, mesmo sem entenderem muitas vezes o que eu fazia. Aos amigos do início da caminhada que faziam da biblioteca sua casa: Wilton, Paulo Henrique, Patrícia Façanha, Daniella Collier, Cibelle Nascimento, Chico Vieira e Mardson. Aos amigos de turma: Eric Parteli, Helinando Pequeno, Ana Maia, Clésio Leão, Frederico Brito e Jonas

[1]Se você não é físico não se preocupe, bósons e férmions são classificações técnicas para as partículas fundamentais constituintes da matéria. Bósons são partículas de spin inteiro enquanto que férmions são partículas de spin semi-inteiro.

Campelo. As garotas Karlla Adriana e Priscila Silva responsáveis por boa parte da alegria do ano de 2005. Aos matemáticos baianos Alex Ramos e Calitéia. Aos colegas "das turmas da frente" Laércio Dias, Mário Henrique, Mércia Liane, Fernando Parísio, Antônio de Pádua, Renê e Clécio Clemente; e "das turmas posteriores" Caio Veloso (encosto), Guga, Márcio Heráclito, Marcelo Alencar, Leonardo Cavalcanti, Felipe Fernando, Gerson Cortês. Ao meu afilhado e amigo de tantas escaladas: Ailton Fernandes, com quem compreendi que competência e humildade podem coexistir. Ao amigo José Ferraz, pelas tantas conversas e pensamentos compartilhados. À Cássia Donato, uma grande amiga, que de perto ou de longe sempre se soube fazer presente de forma positiva. Aos companheiros de grupo Hallan Silva, Washington Lima e Lenira e aos ex-companheiros Gustavo Camelo Neto e Alexandre Rosas.

Em especial a Marcelo Miranda e Josie Rabelo por terem colaborado para que o ano de 2006 fosse um ano de feliz aprendizado e cumplicidade no ingresso à vida adulta.

Aos mestres do ciclo básico Cláudio Furtado e Carlos Alberto. Aos primeiros mestres de pesquisa Marco Gameiro, Marcília Andrade Campos e Fernando Morais, pelos exemplos de ética, bom humor e profissionalismo.

Ao professor Sérgio Coutinho que ao longo destes seis anos de convivência mostrou-se um orientador atento, propiciando a liberdade, a confiança e a experiência necessárias para a consolidação desse trabalho. Aos professores Marcelo A. Moret e Edvaldo Nogueira, pelas colaborações e pelo competente aprendizado científico acerca dos sistemas protéicos.

Aos funcionários Joaquim, Ivo, Humberto, Cristina, Ricardo (in memorian), Paulo Pinto, João, Ana e Joana os quais deram e dão suporte para que a estrutura desse departamento funcione eficientemente.

Ao CNPq e a FACEPE que forneceram o suporte financeiro para esta pesquisa.

Por fim, acredito que daqui a dez anos bósons e férmions continuarão nos livros e poderão ser consultados, mas definitivamente as pessoas citadas acima continuarão em minha memória.

**Resumo**

A maneira na qual uma proteína se enovela a partir de uma espiral aleatória para um estado nativo *único*, em um intervalo de tempo relativamente curto, é um dos problemas fundamentais da biofísica molecular. É bem aceito que esta estrutura tridimensional única, característica de cada proteína e de sua seqüencia de amino ácidos, determina as funções da proteína. Nesta Tese, duas abordagens distintas serão empregadas para estudar aspectos gerais deste problema: 1) uma modelagem estocástica da cadeia principal e de formação de estruturas secundárias, para explorar os aspectos espaciais do estado nativo; 2) uma modelagem de dinâmica molecular, para analisar e caracterizar estatisticamente a evolução temporal da energia conformacional, durante o processo de enovelamento.

Na primeira abordagem, o modelo proposto gera uma cadeia principal com uma fração de estruturas secundárias através de um *caminhante aleatório angular* no espaço tridimensional, cuja trajetória, com passo de tamanho fixo e os ângulos diedrais ($\Phi$ e $\Psi$) das ligações peptídicas escolhidos por duas distribuições de probabilidades gaussianas, cujas médias estão associadas com as estruturas secundárias e a variância $\delta^2$ como um parâmetro ajustável do modelo. Este modelo permite construir uma grande variedade de cadeias distintas, desde aquelas totalmente aleatórias às condizentes com dados experimentais. Algumas propriedades geométricas de proteínas globulares compostas por uma fração $f$ de hélices-$\alpha$ e/ou fitas-$\beta$ são particularmente estudadas. O comportamento de escala obtido para o raio de giração ($R_g$) em função do tamanho da cadeia ($N$); o grau de compactação ($\gamma$); a distribuição do número de coordenação ($z_c$) de carbonos $C_\alpha$ na estrutura e a energia total

envolvida, nestes contatos, foram explorados e comparados com dados extraídos de centenas de proteínas depositadas do *wwPDB* (*worldwide Protein Data Bank*). Os resultados encontrados mostram que, para a fração média $f \sim 0.6$ de estruturas secundárias (hélice-$\alpha$ e/ou fita-$\beta$), as cadeias geradas com distribuições de desvio padrão finito e próximo de $\delta \simeq 0.15\pi$ são mais compactas do que aquelas construídas com outros pares de ângulos diedrais. Independente dos detalhes dos mecanismos físico-químicos subjacentes, a construção de cadeias principais de proteínas com método geral proposto nesta Tese sugere que tais estruturas são governadas por distribuições de probabilidades estreitas e a estocasticidade desempenha um papel fundamental na sua compactação.

Na segunda abordagem, investigamos as propriedades multifractais de séries temporais da energia conformacional de pequenas estruturas em hélice-$\alpha$, especificamente de uma família das polialaninas. Através do método de análise multifractal de flutuações destendenciadas (MF-DFA, do inglês *multifractal detrended fluctuation analysis*), estimamos o expoente de Hurst generalizado $h(q)$ e os associados expoentes de escala multifractal $\tau(q)$, para diversas séries, geradas numericamente por simulações de dinâmica molecular de sistemas em diferentes conformações iniciais. As simulações foram realizadas utilizando-se o campo de força GROMOS implementado no programa THOR. Os resultados mostram que todas as séries analisadas exibem um comportamento multifractal que depende do número de resíduos e da temperatura do sistema. Além disso, as propriedades multifractais das séries revelam aspectos importantes sobre a evolução temporal do sistema e sugerem que o processo de nucleação de estruturas secundárias, durante às visitas da proteína à sua hiper-superfície de energia potencial conformacional, são essenciais no processo de enovelamento.

Palavras-chave: Enovelamento Protéico, Caminhantes Aleatórios, Séries Temporais, Multifractais.

**Abstract**

The manner in which a protein folds from a random coil into a unique native state in a relatively short time is one of the fundamental puzzles of molecular biophysics. It is well accepted that a unique native three-dimensional structure, characteristic of each protein and determined by the sequence of its amino-acids, dictates protein functions. In this Thesis two distinct approaches are considered to study general aspects of such problem: 1) a stochastic modeling of the backbone chain of the protein secondary structures to explore the general spacial aspects of the native state; 2) an analysis of the time evolution of the protein conformational potential energy calculated during the folding process mimicked by methods of molecular dynamics.

In the first approach the proposed model generates a general backbone chain with a fixed fraction $f$ of secondary like structures by means of a three-dimensional off-lattice random walk with fixed steps and the $\Phi$ and $\Psi$ dihedral angles within the peptide bonds chosen by Gaussian probability distributions. Such probability distributions have their mean value corresponding to the angles associated with the chosen secondary structures and the variance $\delta^2$ left as a free parameter to be determined. This model allows the construction of a great variety of backbone chains running from full random structures up to the biological ones observed in proteins. Some geometrical properties of globular structures, composed by a fraction $f$ of $\alpha$-helix and/or $\beta$-strands, were particulary studied. The scaling behavior of the ratio of gyration ($R_g$) with the chain size ($N$); the degree of compactness ($\gamma$); the distribution of coordination number ($z_c$) of the Carbon $C_\alpha$ atoms and energy

involved on such contacts were explored and compared with data of hundreds of proteins extracted from the *wwPDB* (*worldwide Protein Data Bank*). The results indicate that simulated structures are more compact when a fraction of $f \sim 0.6$ of secondary portions ($\alpha$-helices and/or $\beta$-strands) are present than those built with other sets of dihedral angles, whenever the standard deviation of the probability distributions are finite and close to $\delta \sim 0.15\pi$. Independent of the details of all underlying physical chemistry mechanisms, building protein backbones with the method proposed in the present Thesis suggests that these structures are driven by narrow distributions leading to the conclusion that stochasticity has a fundamental role on the its compactness.

The second approach investigate the multifractal properties of the time-series of the conformational energy of small $\alpha$-helix structures, in particular that of poly-alanine family. Using the multifractal detrended fluctuation analysis method (MF-DFA) the generalized Hurst exponent $h(q)$ and its associated multifractal scaling exponent $\tau(q)$ were estimated for several time-series numerically generated by molecular dynamic simulations considering distinct initial configurations. Such simulations were done using the force field GROMOS implemented by the software THOR. In general, the analyzed time series exhibit a multifractal behavior, which depends on the number of residues $N$ and the temperature $T$ of the system. Furthermore, whenever represented by the $h(q)$ or $\tau(q)$ spectra, the time-series multifractal properties reveal important aspects of the time evolution of the system. In particular, suggesting that the nucleation process of secondary structures, which should occurs during the *walk* of the protein on the corresponding portion of the conformational potential energy hyper-surface landscape, is essential for the folding process.

Keywords: Protein Folding, Random-Walks, Time Series, Multifractals.

*Tese de Doutorado*

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

*"Não devemos nos sentir desencorajados pela dificuldade de interpretar a vida a partir das leis comuns da física."*

Erwin Schrödinger - O que é vida ?

Mais notadamente nas duas últimas décadas, a física, a mais "natural" de todas as ciências e a de menor contato com o público leigo em geral, tem participado de um processo de interação com as demais áreas do conhecimento. Tal processo de intercâmbio que vai das ciências sociais aplicadas como a economia [1] até as ciências biológicas [2], passando pela imunologia [3, 4], pela psicofísica [5], pela sociologia [6], pela linguística [7] e pelo urbanismo [8], entre outras; tem permitido avanços não só na descrição e previsão dos fenômenos destas áreas, como também tem ajudado no desenvolvimento de uma série de ferramentas teóricas e experimentais que possiblitam o desbravamento de novos e velhos problemas físicos.

Com métodos comprovadamente eficazes desde os limites do mundo microscópico, como as dimensões dos motores moleculares [9] ao mundo regido pelas

escalas astronômicas [10], as contribuições da física mostram-se perceptíveis e robustas todas as vezes em que se faz necessário abordarmos os elementos constituintes de um sistema e suas interações, através de uma linguagem matemática que possa fornecer informações não só qualitativas como quantitativas acerca do comportamento macroscópico do sistema.

Neste contexto as colaborações ocorridas entre a biologia e a física tem sido um dos exemplos mais profícuos e emblemáticos. Contribuições historicamente impactantes da física à biologia remontam às suas próprias origens com as discussões de Galileu a respeito da altura característica dos seres humanos [11], passa pelas observações de Newton acerca da visão [12], chegando às sugestões de Erwin Schrödinger que abririam o caminho para a biologia molecular [13].

Do século XVI até os dias atuais [2] ambos os campos desenvolveram-se gerando uma grata interação que tem reunido, matemáticos, epidemiologistas, físicos, virologistas, cientistas da computação, químicos, neurologistas, teóricos evolucionários, dentre outras tantas áreas; todos em última análise movidos por perguntas fundamentais a cerca da vida como: Qual a sua origem? De que forma ela evoluiu até as formas que conhecemos hoje? Quais as possibilidades de sua existência sob outras condições distintas das observadas na Terra? De que maneira e em que nível dá-se o processamento de sinais neurológicos? Nesta tese abordaremos um dos aspectos centrais de um dos constituintes básicos de todas as formas de vida conhecida: as proteínas.

Diversas justificativas podem ser dadas na motivação de tal estudo. A primeira trata-se fundamentalmente do impacto que a determinação das estruturas protéicas tem para a indústria farmacêutica, uma vez que a ação dos chamados medicamentos inteligentes está reladionada à ancoragem dos agentes medicamentosos

às proteínas, como no caso dos inibidores de protease do HIV utilizados na terapia de tratamento para a AIDS. A segunda é que se credita à ocorrência de diversas doenças [14, 15] como anemias falciformes, fibroses, catarata, mal de Alzheimer e de Parkinson, bem como distúrbios como o da encefalopatia espongiforme ("mal da vaca louca"), à má-formação das proteínas ("miss-folding"), o que impediria sua correta função no organismo [16]. Desta forma a elucidação de tais doenças passa por uma profunda compreenssão do que ocorre estruturalmente com as proteínas em questão.

Além disso as analogias entre o problema biológico e seu mapeamento em problemas de cunho físico, bem como a utilização de ferramentas físicas tornam o assunto por si só interessante. É importante lembrar que nas duas últimas décadas a física, e em particular a denominada física da matéria condensada, responsável pelo estudo das propriedades estruturais e de transporte em sistemas físicos macroscópicos, tem dado grandes contribuições na caracterização dos chamados sistemas macios [17]. Exemplos de sistemas desse tipo são: os cristais líquidos; os polímeros; os colóides; e as emulsões, que embora descritos por interações distintas, são caracterizados por uma intensa resposta quando estimulados por campos externos. Da mesma forma, as proteínas como heteropolímeros flexíveis, são candidatas potenciais a serem investigadas pelos métodos físicos.

A primeira identificação das proteínas, como unidade fundamental biológica, é creditada ao químico Jöns Jacob Berzelius (1779-1848) (ver Figura 1). Quando nos seus estudos acerca da alimentação constatou que um óxido orgânico parecia ser básico para a nutrição animal, daí o nome originário do grego $\pi\rho\omega\tau\epsilon\iota o\xi$ que significaria primevo, primitivo. Posteriormente, no início do século XX, o químico alemão Emil Fischer (1825-1919) descobriu que as proteínas eram formadas por

cadeias peptídicas, compostas por aminoácidos. Desde então, muitos passaram a atribuir ao sucesso da vida no planeta Terra, o trabalho conjunto de dois grupos de macromoléculas: o ADN (ácido dessoxiribonucléico) e as proteínas. O primeiro responsável pelas instruções de construção e operação das estruturas e o segundo responsável pela construção *per se* [18], estabelecendo-se assim um esquema de "software" (programa) e "hardware" (máquina).

Figura 1.1: Jöns Jacob Berzelius propositor da existência das proteínas - 1838.

Em última análise seriam as proteínas as unidades responsáveis pelas principais atividades enzimáticas, como no caso da lactase responsável pela digestão da lactose; pelo transporte e armazenamento de substâncias, como no caso da hemoglobina; pela regulação e controle do sistema imunológico, tarefa dos anticorpos; pela contração dos músculos, ações da actina e miosina; pela transmissão dos impulsos nervosos no interior das células, como por exemplo, a rodopsina receptora dos bastonetes na retina; bem como no caráter estrutural dos organismos nos ligamentos e tendões constituídos por colágeno e queratina.

Para realizar adequadamente essas diferentes tarefas, acima citadas, as proteínas adotam uma única e bem definida configuração tridimensional, ditada por uma dada sequência de aminoáciodos, que recebe a denominação de estado enovelado ou nativo

[19]. A determinação e classificação dessas estruturas têm sido um dos grandes desafios aos pesquisadores da área, pois exigem a utilização de técnicas experimentais de boa resolução. Duas técnicas têm dado grandes contribuições neste sentido:a cristalografia de raios-X ou e a Ressonância Magnética Nuclear(RMN)[19]. Aplicações dessas técnicas tem possibilitado o aumento no número de proteínas catalogadas no WWPDB[1] (do inglês *Worldwide Protein Data Banking*). De fato esse número cresceu de 250 em 1988, para mais de 37392 na atualidade[2], graças sobretudo às contribuições oriundas de centros de pesquisa e universidades espalhadas em todo o mundo[3]. No entanto, hoje, um paradigma se apresenta em relação às pesquisas na área de biologia molecular, que pode ser resumidamente exposto da seguinte forma: como dar um passo além?, ou seja, como não apenas catalogar, mas, acima de tudo, prever a estrutura das proteínas e construí-las para um propósito específico?

Nesta Tese, apresentaremos duas abordagens para discutir a importante questão do enovelamento das proteínas. Na primeira, proporemos uma metodologia para construção de cadeias proteícas com um controlado processo de formação de estruturas secundárias. Na segunda, combinaremos recursos da dinâmica molecular e da análise estatística multifractal, para analisarmos propriedades estatísticas associadas às flutuações, presentes em séries temporais de energia potencial de proteínas. A Tese está organizda da seguinte maneira: No Capítulo 2, descreveremos as principais características dos sistemas protéicos, procurando relacioná-las ao problema do *enovelamento protéico* [20] e discutiremos os principais modelos propostos para abordagem desse problema. No Capítulo 3, apresentamos uma abordagem espacial simples, para a construção de polímeros que possuem várias das características

[1] http://www.wwpdb.org
[2] 27 de Junho de 2006
[3] Research Collaboratory for Strutctural Bioinformatics - http://www.rcsb.org

estruturais de proteínas reais. Diversas grandezas, tais como: o raio de giração, o número de contatos e a energia de interação entre os mesmos, serão obtidas e analisadas [21, 22]. No Capítulo 4, estudamos uma proposta para a caracterização multifractal das séries temporais, da energia conformacional de proteínas, obtidas por meio da dinâmica molecular de sistemas protéicos [23]. No Capítulo 5, apresentamos nossas conclusões gerais e perspectivas. No Apêndice A, derivamos resultados analíticos associados ao expoente de escala $\nu$ do modelo de Flory e no Apêndice B, apresentamos o sistema de coordenadas utilizado no modelo do caminhante Gaussiano.

# Capítulo 2

# Características Gerais dos Sistemas Protéicos

*"... começar pelo princípio, como se esse princípio fosse a ponta sempre invisível de um fio mal enrolado que bastasse puxar e ir puxando até chegarmos à outra ponta, a do fim, e como se, entre a primeira e a segunda, tivéssemos tido nas mãos uma linha lisa e contínua em que não havia sido preciso desfazer nós nem desenredar estrangulamentos, coisa impossível de acontecer na vida dos novelos e, se uma outra frase de efeito é permitida, nos novelos da vida."*

José Saramago - A caverna

## 2.1 Características gerais dos sistemas protéicos

A despeito de toda variedade de organismos vivos existentes, o processo da vida tal como conhecemos na Terra e como procuramos encontrar em outros pontos do Cosmo, baseia-se numa unidade química denominada aminoácido. Do ponto de

vista funcional cada aminoácido é um elemento com características próprias como: massa, carga, tamanho e hidrofobicidade que determinam as diferentes e possíveis estruturas denominadas proteínas. Essas "composições" de aminoácidos são determinantes na execução de funções que provêem a manutenção da grande estrutura que pode definir um organismo vivo.

Do ponto de vista químico um aminoácido é uma molécula constituída por um carbono central $C_\alpha$ conectado a um átomo de hidrogênio ($H$); um grupo amina ($NH_2$); um grupo ácido carboxílico ($COOH$) e um radical orgânico **R**, como ilustrado na Figura 2.1. A principal propriedade que determina as características estruturais dos aminoácidos e, portanto, das proteínas é a hidrofobicidade (apolaridade) ou hidrofilicidade (polaridade) do radical orgânico **R**, que está conectado ao carbono central denominado **carbono-$\mathbf{C}_\alpha$**, o qual mantém, por sua vez, duas ligações simples: uma com o nitrogênio (**N**) do grupo amina e outra com o carbono $\mathbf{C}_c$ do grupo carboxílico. Um **resíduo** é composto por um radical **R** e pelos átomos que fazem parte da cadeia principal da proteína. Finalmente devemos ressaltar que embora sejam conhecidos cerca de 100 aminoácidos, apenas 20 destes são geneticamente codificados, sendo portanto encontrados em todos os seres vivos.

A síntese protéica ocorre quando um resíduo conecta-se a outro por meio de uma **ligação peptídica**; tal ligação consiste de um condensamento do grupo carboxílico $\mathbf{C}_c$ de um aminoácido ao grupo amina **N** do aminoácido seguinte, liberando assim uma molécula de água ($H_2O$). A repetição de tal processo por meio de sucessivas ligações peptídicas cria uma estrutura denominada cadeia principal ou "esqueleto" ("*backbone*") da proteína. Na Figura 2.2 apresentamos os vinte tipos de aminoácidos existentes, enquanto que na Tabela 2.1 listamos sua nomenclatura, separados em três grupos: o primeiro constituído por aminoácidos hidrofílicos (po-

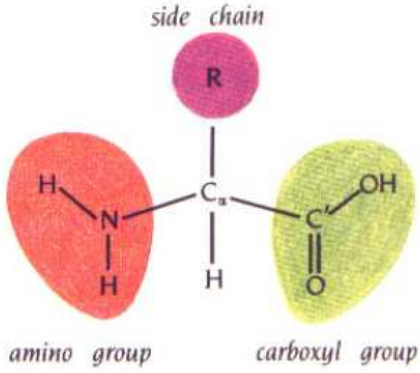


Figura 2.1: Estrutura química dos aminoácidos.

lares), o segundo por grupos eletricamente carregados e o terceiro por aminoácidos hidrofóbicos (apolares)[1].

Os aminoácidos formados pelos três grupos, acima citados, são moléculas quirais, ou seja, podem existir em duas diferentes formas, denominadas levógira (forma-**L**) e dextrógira (forma-**D**), em que uma é a imagem especular da outra. Na natureza apenas as formas levógiras são encontradas. Este fato é um dos grandes enigmas[2] que ainda permeiam a química e que parece possuir conexão estreita com a origem da vida em nosso planeta [18, 24]. Uma vez que o funcionamento dos sistemas biológicos está conectado com a especificidade estrutural das proteínas a existência dos dois tipos **L** e **D** poderia acarretar em deficiências no organismo.

### 2.1.1 Ligações químicas fundamentais

A princípio todos os potenciais e, consequentemente, as forças envolvidas

[1]O aminoácido Glicina (G=Gly), omitido na tabela 2.1 é constituido por apenas um único átomo de hidrogênio e possui características especiais de forma que muitas vezes é considerado um quarto grupo.

[2]Do ponto de vista das forças que unem as moléculas não há distinção entre uma forma ou outra

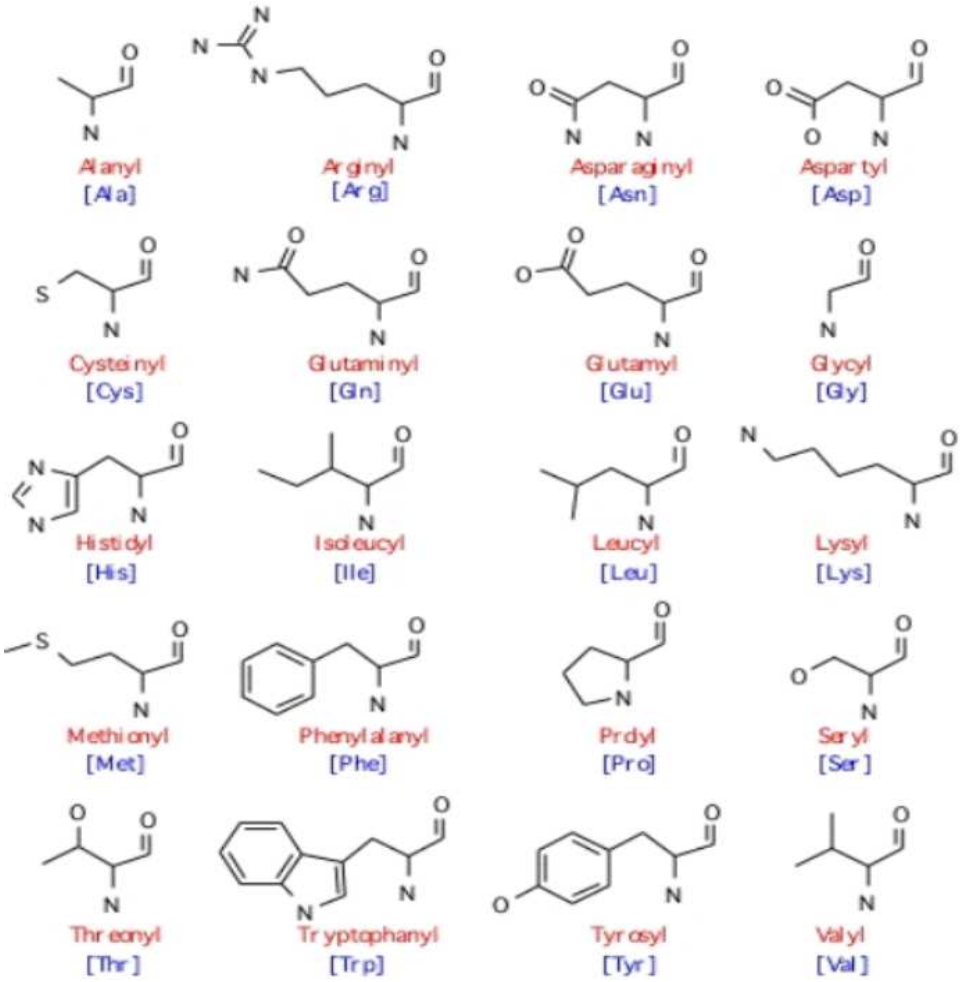


Figura 2.2: Relação dos 20 aminoácidos codificados pelos organismos vivos.

na estabilização conformacional das proteínas são de natureza eletrostática e encontram descrição quântica, de modo que o problema poderia, em princípio, ser solúvel exatamente. No entanto, esse procedimento torna-se impraticável do ponto de vista químico, uma vez que o número de resíduos constituintes de uma proteína pode alcançar a ordem de centenas, neste contexto falamos de ligações químicas efetivas. Para entendermos melhor essa questão; inicialmente vamos estabelecer a

| Classificação | Nomenclatura |
|---|---|
| **Hidrofílicos** | Serina (S=Ser), Cisteína (C=Cys), Tirosina (Y=Tir), Asparagina (N=Asn) Glutamina (Q=Gln), Triptofano (W=Trp) e Teronina (T=Thr) |
| **Hidrofóbicos** | Alanina (A=Ala), Valina (V=Val), Fenilalanina (F=Phe), Prolina (P=Pro), Metionina (M=Met), Isoleucina (I=Ile) e Leucina (L=Leu) |
| **Eletricamente carregados** | Ácido Aspártico (D=Asp), Ácido Glutâmico (E=Glu), Lisina (K=Lis), Arginina (R=Arg) e Histidina (H=His) |

Tabela 2.1: Classificação e nomenclatura dos 20 aminoácidos, sintetizados pelos organismos vivos, por hidrofobicidade, hidrofilicidade e carga elétrica.

escala de energia e as unidades utilizadas nesses sistemas. Frequente, trabalha com unidades de kilocaloria por mol (**kcal/mol**)[3]. Uma boa maneira de termos uma idéia da ordem de grandeza com estes valores é observarmos que o corpo humano a uma temperatura de 310K (37C°) corresponderia a uma energia de 0,642kcal/mol e que o banho térmico de uma sala à 298K (25C°) corresponderia a uma energia de 0,617kcal/mol, esta equivalência pode ser feita por meio da expressão $k_BT$ onde $k_B$ é a constante de Boltzmann e $T$ é a temperatura absoluta do sistema. Estabelecido o padrão passaremos agora a examinar as forças [25, 26] e os tipos de ligações químicas envolvidas, bem como seus valores típicos. As ligações químicas e interações associadas a estabilidade estrutural das proteínas são:

- **Ligações covalentes**: Dentre as interações presentes na estabilização estrutural das proteínas as ligações covalentes são as mais intensas[4], com energias abrangendo variações entre 50 - 250 kcal/mol. Estas ligações são as responsáveis

[3]para converter em elétron-volts utilizar o fator $1eV = 23.06\,kcal/mol$

[4]para quebrar ligações dessa ordem por meio de radiação necessitaríamos incidir ondas eletromagnéticas na faixa de ultra-violeta ($\sim 10^{15}$Hz)

pela interação entre resíduos que não se encontram próximos na sequência dos aminoácidos, mas que se aproximam no estado enovelado, como no caso das pontes dissulfídicas que são formadas por dois átomos de enxofre presentes em diferentes partes da estrutura tridimensional. Tipicamente originadas de dois resíduos de Cisteínas, as pontes dissulfídicas são construídas a partir do processo de oxidação em que átomos de enxofre adjacentes "perdem" seus átomos de hidrogênio como descrito esquematicamente abaixo:

$$2(-CH_2SH) + 1/2O_2 \rightarrow -CH_2S - S - CH_2 + H_2O. \tag{2.1}$$

- **Efeitos Eletrostáticos:** Na Tabela 2.1 observamos que cinco dos vinte aminoácidos geneticamente codificados possuem carga líquida. Estas cargas estão expostas à ação de solventes, comumente a água, e de momentos de dipolo de outras moléculas, ou até mesmo de outros aminoácidos como, por exemplo, os hidrofílicos. Para se ter uma idéia dos valores típicos desses momentos de dipolo vale a pena observar que a molécula da água possui um momento de dipolo de 1.85D[5]. Outra fonte de carga líquida na estrutura protéica provém do desbalanceamento espacial de elétrons, nas ligações covalentes, este desequilíbrio tem sua origem nas diferentes eletronegatividades dos elementos que formam os aminoácidos. Os dipolos peptídicos são uma das grandes contribuições para a estabilidade local de macromoléculas biológicas. As energias associadas a estes efeitos eletrostáticos são descritos pelo potencial Coulombiano:

$$V_{ele} = \frac{1}{4\pi\epsilon_r} \sum_{i<j} \frac{q_i q_j}{r_{ij}}, \tag{2.2}$$

[5]Um próton de carga +e separado de um elétron de carga -e por uma distância de 1Å possui momento de dipolo de aproximadamente 4.8 Debye.

onde o somatório inclui as contribuições entre pares de átomos $i$ e $j$, nas posições $r_{ij}$ representa suas distâncias relativas, $q_i eq_j$ suas respectivas cargas e $\epsilon_r$ é a constante dielétrica do meio. Esta constante pode ter valores muito distintos dependendo do meio em que as cargas se encontrem. A água por exemplo, principal solvente biológico, possui uma constante dielétrica próxima a 80, enquanto que no interior de uma proteína esse valor cai para cerca de 4. Tal variação faz com que o potencial Coulombiano entre dois átomos separados por uma distância de 4Å tenha um valor de 1kcal/mol na água e de 20kcal/mol no interior de uma proteína. Em abordagens tipo dinâmica molecular para descrever o movimento de uma proteína [27], o valor da constante dielétrica desempenha fator decisivo e deve ser minuciosamente modelada afim de evitar divergências numéricas.

- **Ligações ponte de hidrogênio:** Como pode ser observado na estrutura do radical orgânico **R**, que determina os aminoácidos presentes na proteína, os átomos de hidrogênio **H** encontram-se ligados por meio de ligações covalentes a elementos fortemente eletronegativos como nitrogênio (**N**), oxigênio **O** e enxofre **S**. Devido a relação entre a eletronegatividade destes elementos e a eletropositividade do hidrogênio **H**, se estabelece um desbalanceamento espacial de carga que se traduz num momento de dipolo efetivo. A interação entre este momento de dipolo e um átomo parcialmente negativo em suas proximidades é denominada ligação de hidrogênio, ou ponte de hidrogênio. O alcance desta interação, ou seja, a distância característica entre um átomo doador e outro aceitador de carga é tipicamente da ordem de 3.5 Å e sua intensidade é da ordem de 1-7 kcal/mol.

- **Interações de van der Waals:** São exemplos típicos de ligações químicas

efetivas, compostas pela combinação de duas interações: uma primeira repulsiva de curto alcance e uma segunda atrativa de longo alcance. Do ponto de vista fundamental a primeira interação é de origem quântica, e resulta da repulsão das nuvens eletrônicas, devido ao Princípio da Exclusão de Pauli e a da força eletrostáticas entre os núcleos atômicos. A segunda interação resulta das flutuações quânticas do momento de dipolo dos átomos. Interações de van der Waals têm um comprimento característico [6] da ordem de 1.2 Å- 2.2 Å e um valor mínimo de energia em torno da energia térmica, ou seja, da ordem de décimos de kcal/mol e são descritas por:,

$$V_{vdw} = \sum_{i<j} \left[ \frac{C_{12}(ij)}{{r^{12}}_{ij}} - \frac{C_6(ij)}{{r^6}_{ij}} \right], \tag{2.3}$$

onde $r_{ij}$ representa a distância entre os átomos $i$ e $j$ dentro da estrutura e o somatório leva em conta todos os pares de átomos. Embora menos intensas que as interações discutidas anteriormente, as interações tipo van der Waals possuem um importante ingrediente de exclusão, que restringe o número de configurações acessíveis às proteínas[7]. Tal efeito pode se acentuar caso estejamos tratando de muitos contatos formados, o que ocorre quando superfícies moleculares complementares interagem, num mecanismo tipo chave-fechadura.

### 2.1.2 Formações estruturais típicas

Como discutido acima, as interações de van der Waals levam a uma restrição no número de possibilidades dos arranjos espaciais entre aminoácidos consecutivos. Aliado a este fato devemos observar que as ligações peptídicas são planares,

[6] Denominado raio de van der Waals.

[7] Este efeito de exclusão possui o nome técnico de efeito estérico.

de modo que rotações em torno dos grupos **N-C**$_\alpha$ e **C**$_\alpha$**-C**$_c$ de aminoácidos consecutivos, possuem valores específicos. Estas rotações são descritas por meio dos chamados ângulos diedrais ou de torção associados com a rotação do plano formado por 4 átomos (O,N,C,H), em relação à direção da cadeia principal da proteína. O ângulo diedral que corresponde à rotação em torno da ligação **N-C**$_\alpha$ é denominado $\Phi$, enquanto que o correspondente à ligação **C**$_\alpha$**-C**$_c$ é denominado $\Psi$, conforme a Figura 2.3. A distribuição dos valores desses ângulos, como exposta na Figura 2.4, apresenta um caráter universal, na medida em que todas as proteínas conhecidas podem ser classificadas dentro de determinados grupos ou formas típicas. Primeiramente estabelecida pelo biofísico indiano G. N. Ramachandran, a distribuição dos valores de $\Phi e \Psi$, denominado *mapa de Ramachandran* [28, 29, 30] constitui-se hoje numa das caracterizações quantitativas mais importantes no estudo conformacional de proteínas.

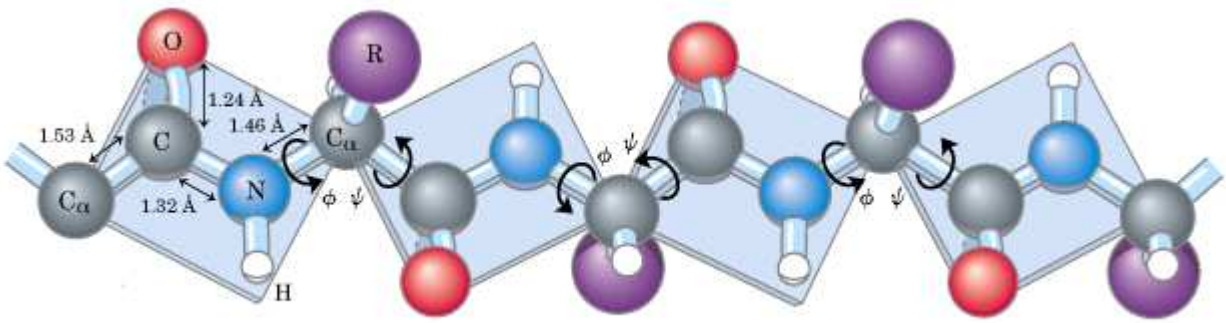


Figura 2.3: Valores dos ângulos de torção $\Phi$ e $\Psi$ envolvidos nas ligações peptídicas.

Morfologicamente as proteínas podem ser classificadas como: **globulares** (estruturas esféricas) e **fibrosas** (morfologia tipo bastão, formada por várias cadeias proteícas). Nesta Tese trataremos basicamente de cadeias de proteínas globulares, cujo número de aminoácidos está entre 50 e 500 resíduos, o que corresponderia a uma

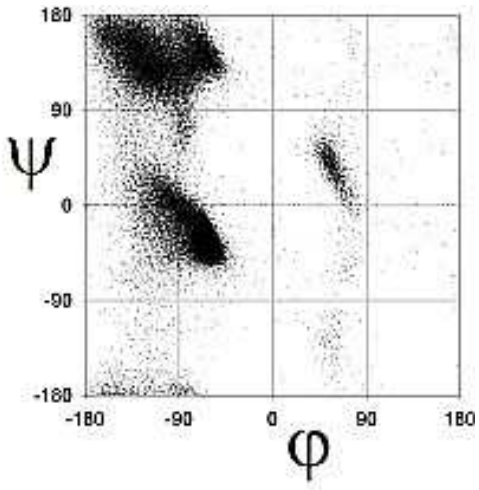


Figura 2.4: Mapa de Ramachandran exibindo regiões permitidas para os valores de $\Phi$ e $\Psi$ nas estruturas protéicas

escala de comprimento que varia entre 25Å e 100Å. As proteínas globulares podem existir em quatro diferentes conformações: nativa (ordenada), "agregado globular", "pré-agregado globular" e desenovelada. Embora possamos observar uma distinção morfológica entre globulares e fibrosas em ambos, podemos identificar "motivos" ou níveis estruturais comuns a todas as proteínas:

1. **Estrutura Primária:** corresponde a sequência de aminoácidos que defines as proteínas. Proteínas podem ser comparadas com relação às suas sequências primárias, por uma técnica denominada de *homologia*, utlizada para determinar quão similar uma estrutura protéica é de outra. Similaridades entre polipeptídeos podem existir graças a permanência de estruturas selecionadas pela evolução biológica dos organismos.

2. **Estrutura Secundária:** devido à interação hidrofóbica entre a água, que envolve a proteína, e grupos que a compõem, certos grupos de aminoácidos agrupam-se de forma a criarem domínios os quais são caracterizados por

possuírem um núcleo hidrofóbico, uma superfície hidrofílica e curta dimensão espacial[8]. Estes padrões regulares são formados por ligações tipo ponte de hidrogênio entre a cadeia principal e os grupos **NH** e $\mathbf{C}_c\mathbf{O}$; correspondendo a valores específicos para o par de ângulos $\Phi$ e $\Psi$ no mapa de Ramachandran. Estes arranjos espaciais recebem nomes particulares:

- hélice-$\alpha$: estruturas clássicas da biologia molecular, são elas as formas preditas por Linus Pauling (em 1951) e, independentemente, pela dupla James Watson e Francis Crick (em 1953) como sendo os padrões energeticamente favoráveis à existência do DNA [31]. As hélices-$\alpha$ foram experimentalmente comprovadas, em 1958, por meio de técnicas de cristalografia de baixa resolução para a mioglobina, por Max Perutz e, posteriormente, com alta resolução por John Kendrew em 1958. De acordo com o mapa de Ramachandran, as hélices-$\alpha$ corresponderiam a região localizada no terceiro quadrante ($\Phi = -60°$ e $\Psi = -40°$) da Figura 2.4. Outros parâmetros que caracterizam as hélices são: seu comprimento, de aproximadamente 15 Å, seu número de resíduos, em torno de 3.6 (resíduos/passo) e o passo da hélice[9], que é da ordem de 5.4 Å. Como todos os átomos de hidrogênio dentro da hélices-$\alpha$ estão alinhados na mesma direção, os momentos de dipolo formados em cada aminoácido somam-se, de forma que toda a estrutura se comporta como um grande momento de dipolo. Devido a hidrofobicidade e a eletronegatividade característica de cada aminoácido, apenas alguns dos 20 listados na Tabela 2.1 são preferencialmente observados na formação de uma estrutura tipo $\alpha-$hélice; estando a alanina, o ácido glutâmico, a leucina e a metionina no grupo

[8]Quando comparados ao tamanho da proteína como um todo

[9]Tecnicamente a altura característica de uma volta (passo) é recebe o nome em inglês de *pitch*.

dos melhores formadores. A Figura 2.5 ilustra representações azimutais (c) e laterais (a,b,d) de uma estrutura secundária tipo $\alpha-$hélice.

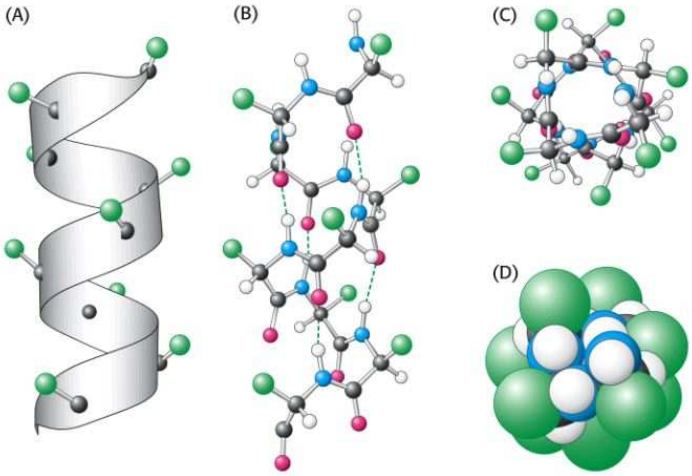


Figura 2.5: Diferentes representações de uma estrutura secundária em hélice-$\alpha$

- fita-$\beta$: estruturas compostas de 5 a 10 resíduos e que são determinadas por ângulos presentes no segundo quadrante do mapa de Ramachandran tipicamente cujos valores estão entre $\Phi = -117°$ e $\Psi = 142°$. Comparativamente às hélices-$\alpha$, os motivos tipo fita-$\beta$ são bem menos compactos, apresentando-se com uma morfologia de fita, as quais podem se alinhar paralelamente ou anti-paralelamente, formando padrões, topologicamente similares a um plano, denominados folhas-$\beta$. Junto com as estruturas em hélice perfazem aproximadamente 60% do total de motivos tipicamente observados nas proteínas[10]. Na Figura 2.6 ilustramos esquematicamente, uma estrutura secundária tipo folha-$\beta$.

[10]Consultar http://en.wikipedia.org/wiki/Secondary_structure.

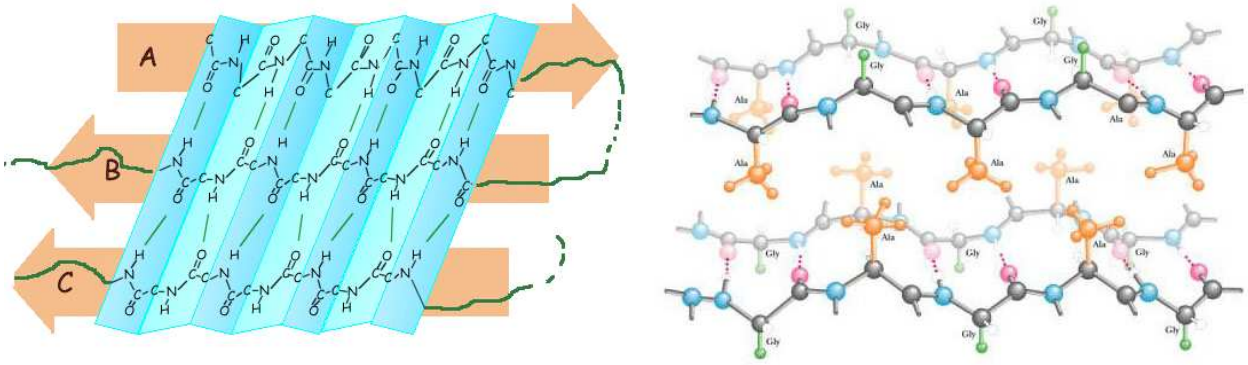


Figura 2.6: Estrutura secundária em folha-$\beta$ e suas configurações paralela e anti-paralela

3. **Estrutura Terciária**: a unidade fundamental da estrutura terciária é um domínio, que pode ser definido como uma cadeia polipetídica ou parte da mesma, ou seja, essencialmente um domínio é uma estrutura secundária. Em geral, o conjunto formado pelos domínios globulares são estabilizados pelo empacotamento de motivos, hélice-$\alpha$ e/ou folha-$\beta$, ligados por pontes dissulfídicas. Estruturas terciárias possuem tipicamente 200 aminoácidos e são de crucial entendimento no processo de enovelamento, uma vez que cada domínio pode se empacotar separadamente. As conexões entre estruturas secundárias são feitas pelas denominadas regiões de *loop*, de formas irregulares e de tamanhos diversos. A Figura 2.7 mostra uma estrutura terciária esquemática, composta de hélices-$\alpha$, fitas-$\beta$ e *loops*.

4. **Estrutura Quaternária**: num nível hierárquico crescente, chegamos à composição de várias estruturas terciárias determinando o que é denominado de estrutura quaternária. Estes complexos globulares consistem de dois (dímero), três (trímero), quatro (tetrâmero) ou mais proteínas individuais; e podem ser classificados como homomérico ou heteromérico, caso sejam constituídos por

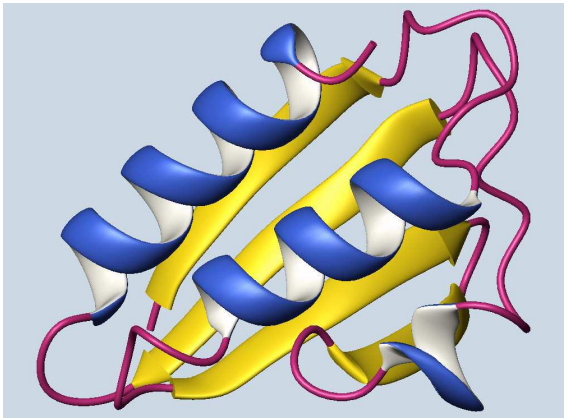

Figura 2.7: Estrutura terciária composta por hélices-$\alpha$, folhas-$\beta$ e loops

um único tipo ou mais tipos de cadeias de proteínas. Embora pareça estranho falar de uma proteína composta por proteínas, lembramos que esta classificação é esquemática, pois na realidade a proteína é a estrutura que desempenha uma função específica. Um caso clássico de estudo e bastante elucidativo é a hemoglobina, um hetero-tetrâmero composto por quatro "proteínas" distintas, duas hélices-$\alpha$, duas folhas-$\beta$ e seus *loops*, como mostrado esquematicamente na Figura 2.8.

Finalizando esta seção, destacamos ainda que à luz dos níveis estruturais, discutidos acima, vemos que o estado denominado "agregado globular", caracteriza-se pela ausência de uma cooperatividade da estrutura terciária da proteína, o que implica num raio hidrodinâmico, em média, 15% maior que o da estrutura nativa implicando num acréscimo de $\sim$ 50% em seu volume. Pelos mesmos argumentos o estado "pre-agregado globular" seria caracterizado por uma estrutura terciária "derretida" onde apenas aproximadamente 50% dos motivos secundários da estrutura nativa final estariam presentes. Observamos ainda que, existem indícios de que a chave para o entendimento da estrutura protéica encontra-se em sua estrutura

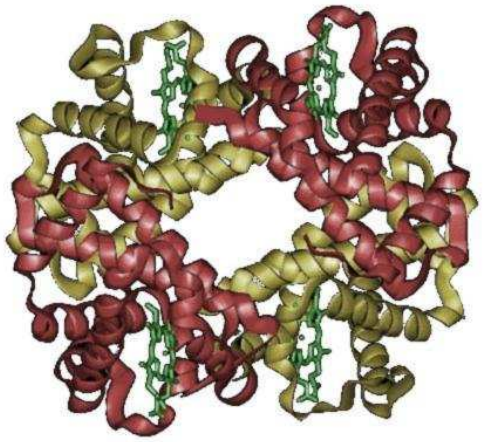

Figura 2.8: Representação de uma hemoglobina exibindo suas cadeias de proteínas constituintes.

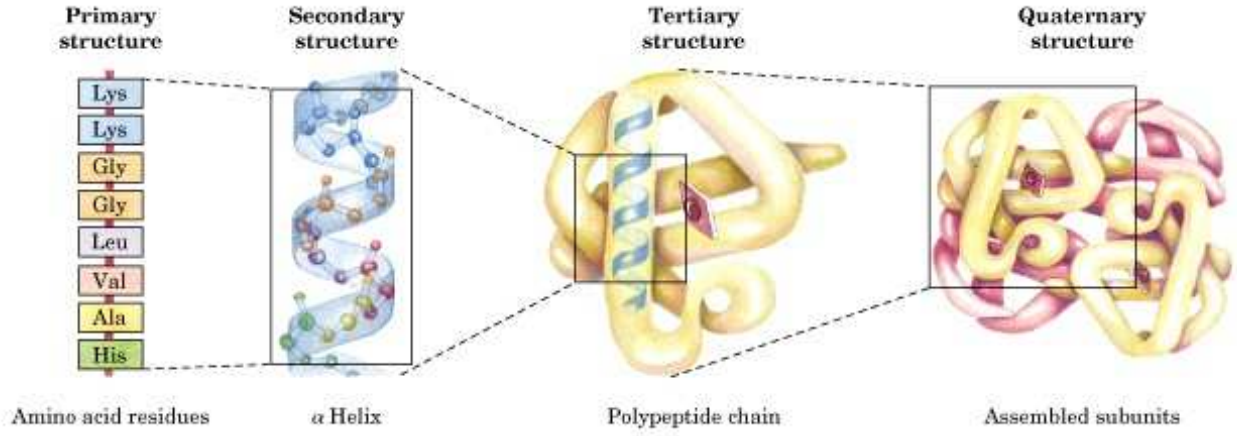


Figura 2.9: Diagramas esquemáticos das proteínas, exibindo os quatro níveis estruturais.

primária. Sob este aspecto cada proteína pode ser entendida como uma palavra construída a partir de um alfabeto de 20 letras (aminoácidos). Embora essa analogia pareça tentadora, é importante observar que cada palavra isoladamente não

traz informação relevante à semântica desse texto biológico, o que está intimamente ligada ao contexto (meio) em que se encontra [32, 33, 34]. Na Figura 2.9, resumimos a composição estrutural das proteínas mostrando sequencialmente sua hierarquia estrutural.

### 2.1.3 A hipersuperfície de energia e a hipótese termodinâmica

É importante observar que o padrão espacial da estrutura protéica é uma consequência direta das interações físicas de atração e repulsão entre as cargas que compõem a molécula e destas com as existentes no meio onde a mesma se encontra imersa. Neste processo cada átomo se compromete a satisfazer um princípio de minimização da energia da estrutura como um todo, mesmo que localmente se veja obrigado a estar submetido a um potencial mais elevado. Essa "frustração", típica de sistemas com muitas partículas, em outros problemas de interesse físico como nos vidros de spin [35, 36]. Nestes sistemas, os momentos magnéticos[11] com interações ferro e antiferromagnéticas, competem de modo a atingir uma configuração de menor energia, em contraposição às flutuações térmicas. Isto significa que, todos os átomos ou momentos magnéticos que os representam não podem localmente satisfazer uma configuração que minimize a energia com todos os elementos constituintes do sistema. No caso magnético o número de configurações estruturais possíveis de serem geradas cresce exponencialmente com o número de elementos envolvidos, de forma que, num processo de resfriamento, o sistema pode se encontrar preso em mínimos locais de energia do sistema. Estes estados, denominados meta-estáveis, encontram-se separados por barreiras de energia com altura infinita e entrelaçados em uma estrutura hierárquica.

[11] A origem de tais momentos magnéticos é puramente quântica.

Uma elucidativa analogia mecânica para tais tipos de problemas, consiste em imaginar a hipersuperfície de energia correspondente a cada uma das diversas configurações como sendo uma paisagem topográfica inundada formada por vales de diversas profundidades. Quando a estrutura encontra-se totalmente inundada a partícula não enxerga qualquer um dos vales. Contudo à medida que o nível da água baixa (correspondendo a uma diminuição da temperatura) a partícula pode ficar armadilhada num dos tantos vales que compõem a estrutura. Nesse quadro uma temperatura elevada pode ser identificada a um estado paramagnético do sistema magnético ou a uma molécula totalmente desenovelada. Enquanto que baixas temperaturas corresponderia a um estado ordenado ou ao "estado-nativo", ou enovelado da proteína.

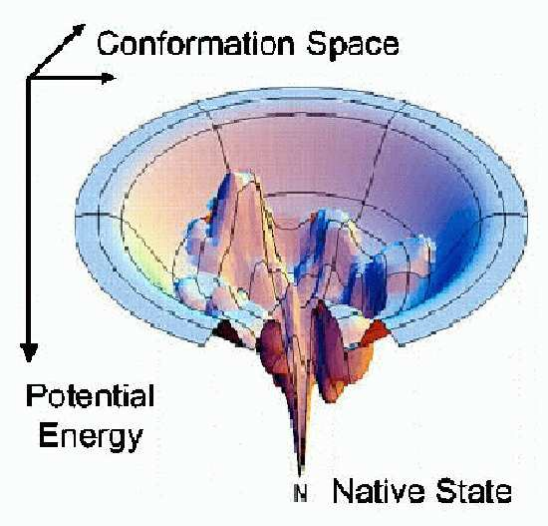


Figura 2.10: Representação da hipersuperfície de energia potencial característica das proteínas

Ao longo da Tese, utilizamos a palavra enovelada ou desenovelada referindo-se ao estado estrutural de uma proteína, mas sem indicarmos com precisão do que

se trataria. Tecnicamente define-se *enovelamento* como sendo o processo no qual uma proteína assume uma forma tridimensional específica, denominada também de *estado nativo*, que a permite realizar sua função biológica. Uma proteína, assim como outros tipos de heteropolímeros, podem enovelar-se ou desenovelar-se reversivelmente, por mudanças causadas pelo $pH$ do meio aquoso, onde se encontre, ou por variações da temperatura desse meio. Neste caso, as proteínas usualmente perdem suas atividades biológicas e são denominadas *denaturadas*. Embora existam estudos, para o problema da denaturação, desenovelamento parcial ou total da proteínas, desde a década de trinta [37, 38, 39], foi somente a partir da década de cinquenta que os trabalhos de Anfisen [40] e outros levaram à formulação da chamada "hipótese termodinâmica", que estabeleceria bases quantitativas para a explicação do enovelamento.

Em linhas gerais, essa hipótese afirma que a informação contida na sequência de aminoácidos (estrutura primária) determinaria sua estrutura enovelada e que esta corresponderia a um mínimo global da energia livre configuracional do heteropolímero. Aparentemente, existem algumas exceções a esta regra, mas estas parecem ser atribuídos a estados meta-estáveis, armadilhados cineticamente durante o enovelamento [41]. Um caso intrigante que difere conceitualmente da referência [41] consiste de uma família de proteínas nativas intrinsecamente desenoveladas que não possuem propriedades estruturais uniformes [42].

Termodinamicamente as contribuições relevantes para caracterizarmos a separação energética entre os estados enovelado (nativo) e desenovelado (denaturado) de uma proteína são: a **Entalpia** e a **Entropia**. A primeira deriva das energias envolvidas nas ligações não-covalentes entre os peptídeos (essencialmente as interações hidrofóbicas, ligações de hidrogênio e iônicas). É sabido que as ligações

covalentes diferem muito pouco entre o estado nativo e denaturado, as ligações não-covalentes exibem um comportamento bem diverso, sendo muito intensas e de longo alcance no caso nativo e quase que irrelevantes no caso denaturado. Enquanto que a entropia conformacional possui valores baixos no caso nativo, uma vez que a organização estrutural é elevada, e muito elevados no caso denaturado, devido a desordem, também denominado "random-coil". É importante ressaltar que para que a entropia diminua na proteína o valor desta grandeza no meio deve se elevar.

A ordem de grandeza da barreira de energia que separa estes dois estados, denominada **energia livre**, é da ordem de 5 - 15kcal/mol, um valor típico de algumas pontes de hidrogênio. Contudo esta pequena diferença energética é um resultado envolvendo dois grandes números, comumente da ordem de centenas de kcal/mol. É importante notar que na ausência de algum fator que compense a entropia, o sistema tenderia a estar no estado denaturado, entropicamente mais favorável, assim a estabilidade da estrutura protéica, é assegurada por um ajuste químico, envolvendo variações de $pH$ e temperatura do meio. Se do ponto de vista da previsão estrutural tal comportamento mostra-se tecnicamente complicado, é importante observar que isso assegura o sucesso da vida, uma vez que o funcionamento do organismo não deve envolver um elevado gasto energético para transformar proteínas de uma forma funcional para outra.

A questão que se coloca é a seguinte: como a proteína "encontraria" seu estado de mínima energia num tempo razoavelmente aceitável? Se partirmos da consideração, de que as proteínas realizam uma busca aleatória, entre as 3 regiões permitidas no mapa de Ramachandran ($\alpha, \beta$ e L) (uma estimativa bastante simplificada), até atingirem um estado "congelado", numa conformação de menor energia, então uma cadeia de 150 aminoácidos (uma proteína pequena) teria aproximada-

mente $3^{150} \simeq 10^{68}$ conformações distintas possíveis. Se o tempo requerido para converter uma conformação em outra for de $1ps = 10^{-12}$ s[12], então uma busca sistemática por todas as configurações levaria um tempo de aproximadamente $10^{56}$ segundos ou seja $10^{48}$ anos [13], conflitando com as observações experimentais, de que partindo de sua estrutura desenovelada, as proteínas mais rápidas atingem seu estado enovelado em tempos da ordem de 1ms e 1s *in vivo* e *in vitro*. Este comportamento aparentemente anacrônico denominado **Paradoxo de Levinthal** [43, 44, 45], em homenagem ao seu propositor o biofísico Cyrus Levinthal.

No final da década de 1960 Levinthal, mostrou que tal busca não seria possível, ou de outra forma, haveriam caminhos favorecidos nessa busca. As idéias iniciais para explicar o Paradoxo de Levinthal partem do pressuposto de que existem trajetórias específicas para o enovelamento, ou seja, haveriam moléculas que restringiriam o número de caminhos, assim as cadeias não necessitariam varrer todo o espaço configuracional [43]. A descoberta de estruturas intermediárias, parcialmente organizadas, como os agregados globulares e os domínios, ao longo do processo de enovelamento, parecem apoiar esta idéia e, embora ainda exista uma grande resistência entre os pesquisadores, alguns [46, 47, 48, 49] começam aceitar elucidações para o "paradoxo de Levinthal".

## 2.2 Abordagens para o problema do enovelamento protéico

As duas principais técnicas utilizadas para determinação da estrutura tri-

[12]Este seria o tempo em que uma ligação atômica se reorganizaria.

[13]Para efeito de comparação é importante notar que a idade do universo é estimada em 10-15 $10^9$ anos.

dimensional das proteínas são: a cristalografia de raios-X [50] e a Ressonância Magnética Nuclear (RMN) [51]. Aliadas a estas, outras como a espectroscopia de massa, a espectroscopia de infra-vemelho e a espectroscopia de ultra-violeta são fundamentais para caracterização dinâmica dos processos de enovelamento. Os primeiros experimentos, que inauguraram a forma de estudar o problema, do enovelamento foram realizados por Anfisen [40] seu principal objetivo nessas pesquisas era entender de que forma a conformação de equilíbrio de pequenas proteínas se alterava sob a variação cíclica de parâmetros como o $pH$ e a temperatura do solvente, onde a mesma estava mergulhada. Como as proteínas atingiam a mesma configuração, sempre que as condições fisiológicas eram reproduzidas independente da trajetória, estes experimentos ajudaram a fundamentar a conjectura básica de que a conformação da proteína só dependeria da sequência dos aminoácidos que a constituiam.

Um exame do grau de enovelamento de uma proteína, ou seja, que fração de sua estrutura encontra-se enovelada pode ser medida através de dicroísmo circular [52]. Essencialmente o que estes experimentos descrevem é de que maneira a polarização da luz incidente numa dada amostra é afetada. Tal método se baseia na propriedade das estruturas secundárias hélice-$\alpha$ e fita-$\beta$ girarem a polarização da luz[14]. Assim podemos acompanhar a população relativa de estruturas já enoveladas a medida que modificamos algum parâmetro externo, podendo então caracterizar a transição de fase estrutural dos heteropolímeros biológicos.

Um proteína típica consiste de algumas pontes salinas, uma centena de ligações de hidrogênio e milhares de interações de Van der Waals. Afim de simularmos este tipo sistema precisamos selecionar modelos apropriados que satisfaçam uma relação entre precisão e custo computacional. *A priori* poderíamos executar

[14] A onda eletromagnética é composta por campos elétrico e magnético, a direção de oscilação do campo elétrico é denominada de direção de polarização da luz.

cálculos com a precisão desejada a partir das posições e velocidades de todos os átomos que constituem a proteína, construindo um Hamiltoniano com todas as interações entre estas partículas, contudo tal abordagem computacionalmente não se mostra eficiente[15]. Primeiro, porque sabemos que algumas dessas interações são mais relevantes do que outras (a magnitude de determinadas interações pode chegar a ser de até duas ordens de grandeza maior do que outras); segundo, porque dependendo dos observáveis de interesse, um grau extremo de detalhamento pode ser totalmente irrelevante e por fim, pelo tempo e recursos computacionais que se dispõe. Assim, níveis de aproximação podem ser impostos aos sistemas de maneira a focalizarmos certos aspectos, como sua dinâmica temporal, sua geometria espacial ou aspectos da dinâmica evolucionária de um certo grupo de proteínas.

Diversos modelos têm sido propostos com diferentes metodologias e objetivos, a seguir descreveremos algumas classes desses modelos existentes, suas principais características, limitações e objetivos. A intenção aqui não é fornecer uma visão completa da "fauna" de modelos existente, mas sim discutir as principais idéias envolvidas em cada tipo de abordagem.

Nos **modelos de cinética química** diversas conformações são identificadas como alguns estados macroscópicos, comumente categorizados em três grupos. O primeiro grupo descreve os possíveis estados desenovelados (**U**, do inglês *unfolded*), correspondendo a todas as conformações não estruturadas da proteína. O segundo grupo consiste do estado nativo (**N**, do inglês *native*) da proteína que corresponderia a uma única conformação. O terceiro e último grupo caracteriza os estados intermediários (**I**, do inglês *intermediate*) que conectariam os estados **U** e **N**, por meio de uma cadeia de eventos descrita pela sequência:

[15]Neste caso estamos nos referindo à cálculos *ab initio*.

$$\mathbf{U} \rightarrow \mathbf{I}_1 \rightarrow \mathbf{I}_2 \rightarrow \ldots \rightarrow \mathbf{N}.$$

Cada estado encontra-se separado do outro por meio de uma barreira energética, de modo que as transições entre cada um dos estados são governadas por equações estocásticas, descritas pelas relações de Kramers [53, 54] utilizando-se de um perfil unidimensional de energia que dirige o caminho do enovelamento [55]. Este tipo de abordagem oferece uma descrição para as distriubições dos tempos de *folding*, podendo ser utilizado em conexão com potenciais de energia livre empíricos para predizer o efeito de mutações na estabilidade e na cinética de uma dada proteína [56]. Contudo, ele se mostra inadequado para descrever mecanismos moleculares na dinâmica de enovelamento. Além disso, ele descreve o sistema por meio de um parâmetro unidimensional, o perfil de energia, uma simplificação que não oferece nenhum aspecto geométrico da estrutura.

Uma segunda abordagem envolve uma compreensão mais detalhada da paisagem de energia livre e da entropia associadas à sua conformação espacial. Assim, nos denominados **modelos baseados em entropia**, são geradas um grande número de conformações através de simulações computacionais. As cadeias protéicas são então caracterizadas energeticamente por meio de um potencial de energia, obtido através de uma soma de termos de contato entre dois elementos da estrutura. Um exemplo clássico deste tipo de abordagem pode ser encontrado no modelo de Go [57], no qual cada aminoácido de uma proteína pode ser descrito através de uma representação onde cada átomo figura explicitamente na estrutura, ou por meio de uma representação simplificada em que cada aminoácido apresenta-se como uma conta esférica sem estrutura interna.

A idéia básica neste tipo de modelo, é que a geometria da proteína é o principal fator guiando o processo de enovelamento, e desta forma a energia li-

vre do sistema pode ser associada aos estados conformacionais por meio de um termo entrópico. Sua grande virtude é a capacidade de descrever a existência de fenômenos cooperativos na formação da estrutura espacial da proteína. Simulações realizadas para pequenas proteínas [58] confirmam, por exemplo, a existência de contatos específicos a partir dos quais a estrutura desenovelada atinge o estado nativo. Tal evidência indica que a proteína não se enovela de forma aleatória, seguindo na realidade uma sequência bem definida de passos. Por outro lado estes modelos não abordam a interação entre os 20 diferentes tipos de aminoácidos existentes na natureza e a possibilidade de que estes contatos possam formar interações inexistentes na estrutura nativa final. Ou seja, estes modelos negligenciam tanto o papel da sequência específica de aminoácidos para a formação de cada proteína, como a existência de estados metaestáveis que armadilhem o processo de enovelamento.

Uma outra abordagem termodinâmica para o problema do enovelamento, pode ser dada considerando-se a informação contida na sequência de aminoácidos que constitui a proteína. Nesta visão, a heterogeneidade das interações entre os aminoácidos produz uma paisagem de energia essencialmente rugosa, com diversos estados metaestáveis sendo formados. O objetivo dos **modelos baseados em energia** é entender de que forma uma proteína caracterizada por uma sequência específica de aminoácidos difere de um sistema aleatório. O principal elemento neste tipo de modelagem é a função de energia potencial que descreve a interação entre os contatos de dois aminoácidos da estrutura. Uma matriz para estas interações foi determinada por Miyazawa and Jernigan [59] em 1985.

O ponto de partida dos modelos baseados em energia é o estudo de cadeias formadas por uma sequência aleatória de aminoácidos, assim num modelo simplificado a energia de cada interação é escolhida aleatoriamente e a energia total para uma

configuração de $N$ contatos é dada pela soma das energias de interação entre pares de aminoácidos na estrutura. Como consequência destas considerações estabelece-se um primeiro modelo denominado modelo de energia aleatória (REM do inglês, *Random Energy Model*) [60], o qual prevê um estado fundamental $E_c$ e, a partir deste, um espectro contínuo. Como mostrado por Shakhnovich [61], a paisagem de energia para este modelo é composta por diversos mínimos locais, todos separados por pequenas barreiras de energia, que o distingue do comportamento de uma proteína real, a qual possui um mínimo global bem mais profundo do que os dos estados meta-estáveis. Nesse quadro, as características cinéticas do problema também tornam-se muito distintas daquela exibida por proteínas reais, uma vez que segundo esta prescrição uma proteína facilmente poderia ser armadilhada num desses estados meta-estáveis [62].

O desenvolvimento de computadores cada vez mais velozes tem permitido uma descrição mais detalhada da estrutura atômica das proteínas e dos potenciais de energia envolvidas na interação entre estes átomos. A construção da denominada **modelagem molecular**[27] baseia-se na simulação dos movimentos atômicos por meio de um campo de forças, para valores de temperatura e pressão conhecidos. É possível através desta abordagem analisar as interações provenientes entre a estrutura protéica e o solvente, onde a mesma se encontra imersa, tratando-o como um contínuo, ou até mesmo pela descrição explicita das moléculas que o compõem.

Essencialmente este tipo de modelo pretende resolver as equações de movimento de Newton através de um processo de integração numérica, o que permite investigar a evolução temporal entre várias conformações. Tecnicamente, a resolução das equações de movimento envolvem a utilização de potenciais efetivos clássicos parametrizados, os quais descrevem as interações intra e inter-moleculares.

O "coração" da dinâmica molecular reside em última instância, na função potencial escolhida, a qual deve ser realista o suficiente para fornecer descrições precisas da estrutura do sistema e ao mesmo tempo de fácil e rápida implementação numérica. Embora exista na literatura uma grande diversidade de potenciais e de programas, tanto de uso comercial quanto acadêmico (GROMACS, AMBER, CHARMM e THOR) [63, 64, 65, 66, 67], todos possuem dois fatores comuns: a existência de potenciais "ligados", usualmente representados por termos harmônicos, e de potenciais de interação à distância. Os primeiros são utilizados para descrever as ligações covalentes entre pares de átomos ângulos entre ligações químicas vizinhas e ângulos de torção em torno de ligações; enquanto que o segundo descreve interações entre átomos não ligados covalentemente, levando em conta as atrações e repulsões eletrostáticas, bem como as interações dipolares.

Como já discutido nas seções anteriores, devido ao elevado número de elementos constituintes, o número de graus de liberdade de um sistema molecular pode ser astronômico e desta forma torna-se impossível cobrir toda a superfície de energia. Como solução alternativa, os algoritmos tentam vasculhar não toda a superfície, mas sim o caminho que leva o sistema para uma configuração de menor energia. Para tanto entra em cena a segunda parte do problema de dinâmica molecular, a otimização de geometria. A otimização consiste essencialmente de um procedimento para obtenção, a cada passo de tempo fixado pelo algoritmo, do conjunto de cooordenadas que minimiza a energia do sistema como um todo.

Nesse espírito podem ser encontrados na literatura diversos tipos de algoritmos para efetuar tal minimização, tais como: o "steepest-descent" [68], o método dos gradientes conjugados [69] e os métodos de busca aleatória como o *"Generalized Simulated Annealing* [70]. É importante observar que assim como o potencial

de energia precisa ser acurado o suficiente para descrever as grandezas de interesse, a otimização precisa ser eficiente para que o sistema possa ser avaliado em tempos razoáveis. Dinâmicas que pretendem chegar a tempos da ordem de $1\mu s$ podem necessitar de tempos computacionais efetivos da ordem de dias, dependendo do tamanho do sistema, uma vez que cada passo da dinâmica gira em torno de 1fs. Embora se constitua de um poderoso método para a análise tanto de estruturas protéicas, como da interação destas com membranas e solventes, a grande limitação dos métodos de dinâmica molecular reside no fato de que atualmente a maior parte dos cálculos cobrem apenas uma pequena fração do tempo real envolvido no processo de enovelamento.

Um quarta abordagem são os **modelos de rede** indicados para estudar características gerais do enovelamento protéico. Tradicionalmente, se baseiam na aproximação de que a estrutura interna dos aminoácidos (seus átomos) pode ser negligenciada como consequência o caráter entrópico associado aos graus de liberdade internos podem ser também negligenciados. E assim, qualquer interação deverá ser considerada isotropicamente desta forma, os aminoácidos são representados por contas, conectadas em redes. Cada sítio pode estar vazio ou contendo uma conta, o que simularia o efeito do volume excluído. No caso de uma rede quadrada os ângulos de ligação entre cada aminoácido estão restritos aos valores $\pi/2$, $\pi$ ou $3\pi/2$ radianos, enquanto que o comprimento das ligações tem um valor fixo caracterizado pelo parâmetro de rede. Estes modelos, também denominados de modelos minimalistas, estão preocupados em elucidar perguntas básicas acerca do enovelamento a saber: como uma cadeia polimérica pode enovelar-se num estado nativo único, como se dão os mecanismos de cooperatividade ao longo do processo, por quê algumas sequências primárias enovelam-se enquanto que outras não ? E o caráter da

frustração na evolução do processo [71].

O mais conhecido desta classe é o modelo HP [72], onde apenas dois tipos de aminoácidos são considerados: um denominado **H**, de caráter hidrofóbico e outro **P** de caráter polar estes aminoácidos são distribuídos aleatoriamente numa rede cúbica ou quadrada. É assumido nesse caso um potencial de interação entre sítios vizinhos dependendo dos aminoácidos que os ocupam (**H** ou **P**), dessa forma o modelo contempla a possibilidade de aminoácidos, que não são consecutivos sequência primária, poderem formar "contatos". Como observado por Tang [73] em sua revisão a respeito de modelos dessa natureza, a maior parte das estruturas geradas nesse tipo de modelo não se enovelam num único estado, o que não descreve a situação real protéica.

Em suas versões mais recentes [73], os modelos em rede consideram uma matriz de interação para um alfabeto composto por 20 aminoácidos, no mesmo contexto dos modelos baseados em energia. É da heterogeneidade destas iterações que se revela o caráter frustrado do sistema, que passa a exibir uma superfície de energia rugosa. Um dos grandes resultados obtidos com esse tipo de metologia é a a formação de estruturas elementares locais durante o processo de enovelamento. Essa formação ocorre numa sequência de eventos hierárquicos que se mostram fundamentais para o mecanismo de estabilização da estrutura como um todo [74].

# Capítulo 3

# Modelo de caminhantes angulares Gaussianos

*"... você marcha, José! José, para onde? "*

Carlos Drummond de Andrade - José

## 3.1 Caminhantes aleatórios e caminhantes auto-excludentes

Em 1828, o botânico escocês Robert Brown (1773-1858) realizou experiências nas quais observou, com a ajuda de um microscópio, que numa suspensão de grãos de pólen em água cada grão se movia irregularmente, numa trajetória errática que seria posteriormente denominada de Browniana [75]. Constatando que esse fenômeno era reproduzido com o uso de diversas substâncias orgânicas, Brown acreditou haver encontrado a molécula primitiva da matéria viva. Esse suposição foi posteriormente

refutada pelo próprio Brown, ao perceber que substâncias inorgânicas pulverizadas também possuiam comportamento similar. Após essa primeira incursão "equivocada" pela biologia, os *caminhantes aleatórios* (ou *"random walks" ; RW*), resurgiriam com fundamentação matemática, no contexto econômico em 1900, graças ao francês Louis Bachelier (1870-1946), em sua tese Teoria da Especulação[1], sobre opções de preço em mercados especulativos [76].

Uma descrição corpuscular da matéria, fundamentada num modelo de caminhante aleatório, foi introduzida por Albert Einstein (1879-1955) em 1905 [77]. Com a proposição do número de Avogadro, Einstein relacionou as grandezas estatísticas do movimento Browniano com o comportamento dos átomos e deu aos experimentalistas um método de contagem dos átomos, o que foi comprovado experimentalmente por Jean-Baptiste Perrin (1870-1942)[2] em 1908. Atualmente estudos de processos difusivos estão interessados em propriedades estatísticas associadas à caminhantes deslocando-se de maneira errante sobre um substrato. Essa dinâmica é recorrentemente utilizada como um conceito importante para se determinar as propriedades de escala de muitos fenômenos físicos [78, 79, 80, 81].

No modelo de caminhante aleatório, cada passo é completamente independente de todos os passos anteriores, tais processos estotásticos são também chamados de Markovianos. Contudo, para diversos processos físicos essa suposição não é apropriada. Ao idealizarmos um polímero, por exemplo, como uma longa cadeia flexível destituída de qualquer estrutura interna, onde cada ligação corresponde a um passo do caminhante, temos de levar em conta a exclusão espacial. Uma caminhada aleatória submetida a tal condição é denominada *caminhada aleatória*

[1]Orientada pelo também matemático Henri Poincaré(1854-1912), a tese de Bachelier foi preterida por seus contemporâneos e permaneceria na obscuridade até meados da década de 1960.

[2]Por suas medições Perrin receberia o prêmio Nobel em 1926.

*auto-excludente* (ou *self-avoiding walk; SAW*). Na Figura 3.1, apresentamos padrões típicos gerados por uma caminhada aleatória e uma caminhada auto-excludente, ambas sobre uma rede quadrada. Em geral, as quantidades geométricas analisadas nesse tipo de modelo são: o deslocamento quadrático médio em relação à origem $< r^2 >$ e o raio de giração da estrutura final $R_g$, definido como a distância quadrática média de cada posição ao centro de massa da estrutura,

$$R_g \equiv \sqrt{\sum_{i=1}^{N} \frac{\rho_i^2}{N+1}}, \tag{3.1}$$

onde $\rho_i$ é a distância de cada um dos elementos da estrutura ao centro de massa e $N$ é o número de passos do caminhante.

Tais grandezas exibem uma relação de escala, ou seja, o raio de giração, por exemplo, exibe um comportamento tipo lei de potência com o número de passos $N$ do caminhante,

$$R_g \sim N^{\nu}, \tag{3.2}$$

onde $\nu$ é o expoente que caracteriza o tipo de difusão subjacente ao processo, com a seguinte classificação:

$$\begin{cases} \nu < 1/2 & \text{(subdifusivo)}, \\ \nu = 1/2 & \text{(difusivo)}, \\ \nu > 1/2 & \text{(superdifusivo)}. \end{cases} \tag{3.3}$$

Um modelo tipo campo médio para formação de estruturs poliméricas embebidas em um bom solvente, desenvolvido por Flory[82, 83], prevê um expoente

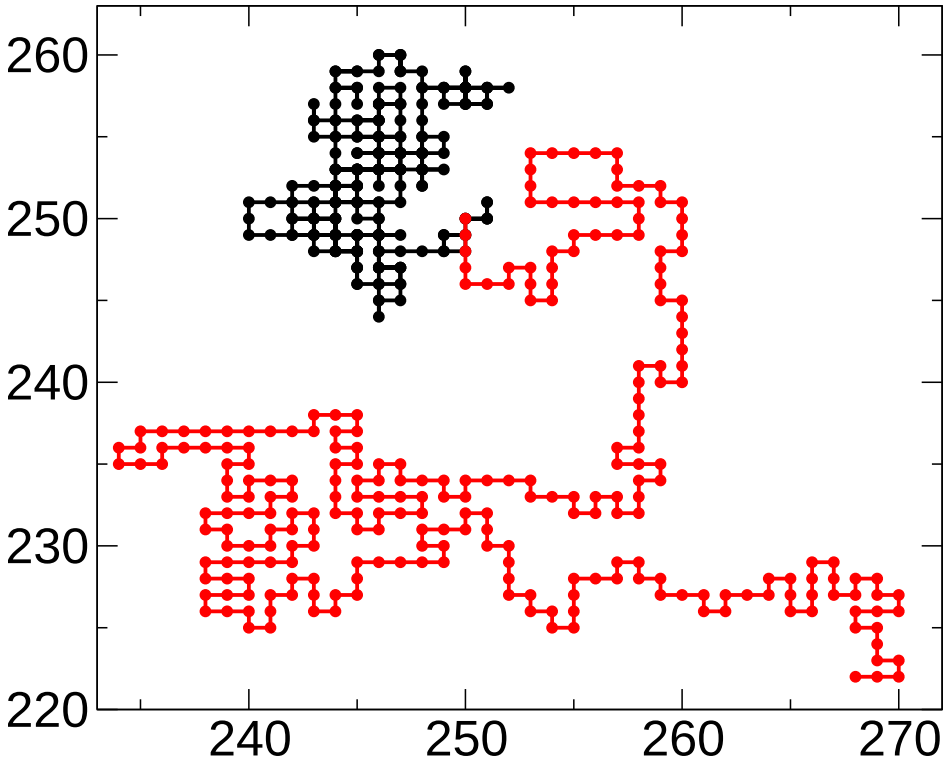


Figura 3.1: Padrões típicos, obtidos através de simulação numa rede quadrada, para caminhantes aleatórios (em preto) e caminhantes aleatórios auto-excludentes (em vermelho), ambos com 250 passos e partindo do ponto $(250, 250)$.

dependente da dimensão $D$, dado por:

$$\nu_{Flory} = \frac{3}{D+2}, \tag{3.4}$$

de modo que $\nu_{Flory} = 3/5$ para o caso em três dimensões pertencendo, também, a um regime superdifusivo, este resultado encontra-se reproduzido no Apêndice A desta Tese. Recentes simulações [84] de *RW* e *SAW* em três dimensões apresentam resultados diferentes para o expoente com $\nu_{RW} = 1/2$ e $\nu_{SAW} = 0.588$, indicando que no caso auto-excludente a partícula escapa mais rápido da origem que no regime

difusivo.

## 3.2 Modelo de caminhantes angulares Gaussianos

Ao tratarmos de cadeias protéicas, observamos que o raio das estruturas com o número de aminoácidos, também obedece um comportamento de escala, como pode ser observado no gráfico log-log experimental, baseado em 1826 cadeias protéicas, exibido na Figura 3.2. No contexto discutido na seção anterior, este comportamento pertence a um regime sud-difusivo, com $\nu_{exp} = 0.40 \pm 0.02$, diferente daquele previso por Flory, $\nu_{Flory} = 0.60$ e observado no contexto de superfícies amassadas como indicado por Gomes e colaboradores [85]. Constatamos dessa forma que mesmo as estimativas construídas em argumentos de campo médio, onde a entropia da superfície da estrutura é levada em conta, não é capaz de descrever esta característica geométrica básica.

Como discutido no capítulo 2, diversas abordagens para descrição das propriedades estruturais das proteínas têm sido propostas. Devido ao número astrônomico de possíveis configurações para proteínas globulares (compostas por 50 a 500 aminoácidos), metodologias convencionais fundamentadas em Monte Carlo ou dinâmica molecular tornam-se impraticáveis, devido ao seu alto custo computacional.

Nesta seção, apresentamos uma estratégia alternativa, baseada num modelo de caminhante aleatório em três dimensões, como forma de construirmos cadeias protéicas com diferentes comprimentos e porcentagens de estruturas secundárias. No modelo, cada passo possui um comprimento radial $l_0$ fixo, mas os ângulos diedrais, $\Phi$ e $\Psi$ que compõem a cadeia são escolhidos independentemente por meio de uma distribuição de probabilidades Gaussiana, seguindo a proposta de Shaw e colaboradores [79]. Os valores médios e o desvio de cada distribuição são definidos de acordo

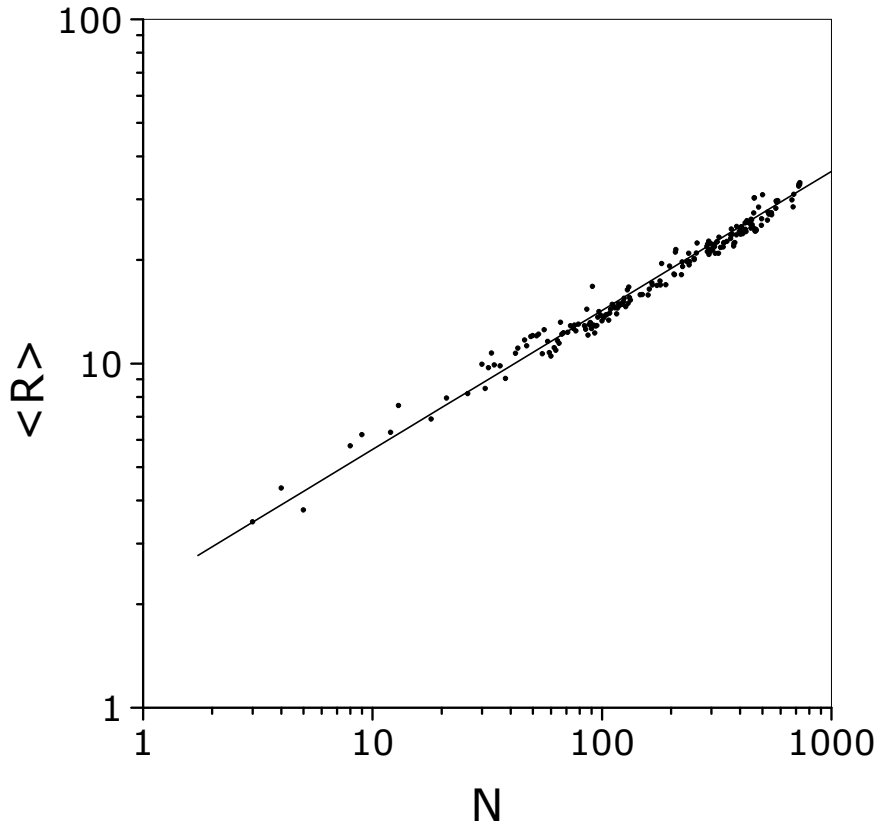


Figura 3.2: Comportamento do raio médio $< R >$ em função do número de aminoácidos $N$ para um conjunto de 1826 cadeias protéicas, com expoente $\nu \approx 0.40 \pm 0.02$. A linha contínua indica o ajuste linear dos dados

com as regiões permitidas para $\Phi$ e $\Psi$ presentes no mapa de Ramachandran. Aqui utilizamos os valores calculados por estatíticas de diversas estruturas secundárias, propostos pelo programa PRELUDE [86] a Tabela 3.1 indica sete possíveis pares de ângulos diedrais e suas conformações associadas.

Desta forma, as distribuições Gaussianas para os ângulos são explicitamente descritas como:

$$P(\Phi) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(\Phi - \Phi_o)^2}{2\delta^2}\right), \tag{3.5}$$

e

$$P(\Psi) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(\Psi - \Psi_o)^2}{2\delta^2}\right), \tag{3.6}$$

onde $\delta$ é o desvio padrão de cada distribuição, enquanto que $\Phi_o$ e $\Psi_o$ são respectivamente os valores médios dos ângulos diedrais $\Phi$ e $\Psi$.

| $\Phi_o$ | $\Psi_o$ | Conformação |
|---|---|---|
| -65° | -40° | $A$ |
| -89° | -1° | $C$ |
| -117° | 142° | $B$ |
| -69° | 140° | $P$ |
| 78° | 20° | $G$ |
| 103° | -176° | $E$ |
| -83° | 133° | $O$ |

Tabela 3.1: Sete possíveis pares para ângulos diedrais ($\Phi$,$\Psi$) e suas conformações associadas [86]. Configurações em hélice-$\alpha$ denotadas por $A$ e folha-$\beta$ por $B$.

Para simularmos proteínas com uma determinada porcentagem de estruturas secundárias $f$, fixamos o número de passos do caminhante $N$ e estabelecemos um processo de crescimento descrito pelas seguintes regras:

1. Nos primeiros $l_g \times f$ passos escolhe-se **um** dos pares de ângulos da Tabela 3.1 e por meio das distribuições descritas nas Equações 3.5 e 3.6, construímos sequencialmente a próxima posição como função da posição anterior.

2. Nos próximos $l_g \times (1 - f)$ passos utilizamos, a cada passo, ângulos aleatoriamente escolhidos entre os sete possíveis pares, utilizando mais uma vez as distribuições descritas nas Equações 3.5 e 3.6, construímos sequencialmente a próxima posição como função da posição anterior.

3. Nos próximos $l_g$ passos as regras 1 e 2 são repetidas, até construirmos uma cadeia de tamanho $N$.

A determinação da posição $(x_i, y_i, z_i)$, como função da posição anterior $(x_{i-1}, y_{i-1}, z_{i-1})$, envolve uma transformação entre coordenadas cartesianas e internas, descritas pelos ângulos $\Phi$ e $\Psi$, como observado por Park e colaboradores [87] [3]. Neste modelo minimalista, a construção do esqueleto peptídico envolve apenas os ângulos diedrais. Todos os potenciais efetivos descritos no Capítulo 3, sejam eles de contato ou de ação à distância, são levados em consideração indiretamente através da escolha dos ângulos da Tabela 3.1. Uma vez que a distribuição de ângulos diedrais encontra-se confinada às regiões permitidas pelo mapa de Ramachandran, espera-se que os fenômenos estéricos estejam inclusos nessa abordagem implicitamente. Devido à sua simplicidade computacional, o método permite a construção de uma elevada quantidade de amostras, ou seja, de diferentes conformações (da ordem de $10^4$).

É importante perceber, que nesse modelo, a **regra 1** induz um crescimento ordenado, durante um comprimento característico $(l_g \times f)$, enquanto que a **regra 2** descreve a quebra desse padrão estrutural, quebra esta que pode ser entendida como resultante da instabilidade das forças envolvidas na formação de conformações específicas, como no caso dos momentos de dipolo efetivos das estruturas hélice-$\alpha$.

Pode-se então simular diversas destas configurações, variando-se a fração $f$ e o tipo de estrutura secundária utilizada na regra 1 do modelo. Em nosso estudo utiliz-

[3] Para detalhes consultar o Apêndice B desta Tese.

amos: (a) estruturas hélice-$\alpha$; (b) folhas-$\beta$ ou (c) uma mistura de hélice-$\alpha$ e folha-$\beta$. A escolha desses motivos estruturais decorre de sua relevância nos estágios iniciais do enovelamento protéico e por estes serem os blocos fundamentais na formação de estruturas terciárias [88, 89]. Nas Figuras 3.3, 3.4 e 3.5 exibimos configurações típicas compostas por $N = 250$ resíduos, para os três casos acima especificados, assim como os respectivos os mapas de Ramachandran para um conjunto de 100 amostras, em cada um desses casos. Em todas as simulações a largura da distribuição vale $\delta/\pi = 0.1$, o comprimento $l_g = 100$ e o comprimento característico entre cada resíduo (distância radial) vale $l_o = 3.8$ Å.

## 3.3 Análise das grandezas relevantes

Uma vez criadas diferentes configurações, através do algorítmo descrito na seção anterior, utilizamos algumas grandezas a fim de caracterizar estruturalmente as cadeias geradas artificialmente e compará-las às cadeias reais. Seguindo a sugestão de Tang e colaboradores [90], elegemos o raio de giração $R_g$; o comprimento de contorno $l_c$; o número de contatos $n_c$; o número de coordenação $z_c$ e a energia entre os contatos $E$. Além destes parâmetros introduzimos outra quantidade denominada "parâmetro de compactação", definido como:

$$\gamma \equiv \frac{R_g}{D_{max}}, \tag{3.7}$$

onde $R_g$ é o raio de giração da estrutura e $D_{max}$ é a maior distância entre dois resíduos da mesma. Discutiremos a seguir a relevância de cada uma dessas grandezas e de que forma os parâmetros de nosso modelo as descrevem.