

**ANÁLISE DA HIDROFOBICIDADE  
NA EVOLUÇÃO DE PROTEÍNAS**

*Ricardo H. Theodoro da Silva*

Tese de Doutorado  
Pós-Graduação em Biofísica Molecular

# **Análise da Hidrofobicidade na Evolução de Proteínas**

**Ricardo H. Theodoro da Silva**

Exame de doutoramento apresentado como parte das exigências para obtenção do título de Doutor em Biofísica Molecular, área de concentração Biofísica Molecular do Departamento de Física do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho” – UNESP, Câmpus de São José do Rio Preto, São Paulo.

Orientador: **Prof. Dr. Vitor Barbanti Pereira Leite**

Co-orientador: **Prof. Dr. Jorge Chahine**

São José do Rio Preto

Setembro de 2009

”O deserto que atravessei  
Ninguém me viu passar  
Estranho e só  
Nem pude ver que o céu é maior”

*Zélia Duncan*

Agradeço primeiramente a Deus.

Aos meus queridos pais,

ao apoio de meu orientador Prof. Dr. Vitor Barbanti Pereira Leite,

ao meu co-orientador Jorge Chahine.

Aos Professores Drs: Antonio Caliri, Michel E. Beleza Yamaghishi, Luis Paulo B. Scott e Sidiney Jurado de Carvalho,

que participam desta banca. A ajuda e compreensão do coordenador da pós graduação Elso Drigo Filho

Aos meus colegas de grupo Luciana, Leandro, Grilo e Ronaldo. ao meu camarada André.

aos meus colegas Gabriel, Ricardinho, Ézio e Joaquim.

A todos os professores e funcionários, que de uma forma ou de outra participaram da minha formação.

Agradeço ao apoio de meu irmão Renato e meus parentes, em especial, ao tio Evaldo.

Agradeço a ajuda de minha cunhadas Leila, Lígia e Edite.

Agradeço o apreço de minhas queridas sobrinhas Camila, Gabriela, Daniela e Letícia

Ao apreço de meus sobrinhos Guilherme, Pedro, Lúcio e Luciano. A ajudante da Sofia, Fabiana

Aos meus colegas de trabalho da UNIFEV, em especial ao Rogério, Angelo e Paulo

Enfim a todos aqueles que de alguma forma participara desta longa caminhada.

*Dedico este trabalho a quem  
sempre me apoiou nas horas  
mais difíceis. Minha amada  
esposa Lucimara e  
minha filha Sofia.*

# Sumário

<b>Resumo</b>	<b>1</b>
<b>Abstract</b>	<b>2</b>
<b>1 Bioquímica e o Problema do Enovelamento de Proteínas</b>	<b>3</b>
1.1 Importância e Bioquímica . . . . .	3
1.2 Estrutura das Proteínas . . . . .	4
1.2.1 O que é uma proteína? . . . . .	4
1.3 Níveis Estruturais . . . . .	4
1.3.1 Estrutura Primária . . . . .	5
1.3.2 Estrutura Secundária . . . . .	5
1.3.3 Estrutura Terciária e Quaternária . . . . .	5
1.4 Classificação e Características dos Aminoácidos . . . . .	5
1.4.1 Grupo R apolare/alifático . . . . .	6
1.4.2 Grupos R aromáticos . . . . .	7
1.4.3 Grupos R polares não carregados . . . . .	7
1.4.4 Grupos R carregados positivamente (básicos) . . . . .	8
1.4.5 Grupos R carregados negativamente (ácidos) . . . . .	8
1.5 Enovelamento de Proteínas . . . . .	8
1.5.1 Interações Energéticas no Processo de Enovelamento . . . . .	9
1.6 Importância dos Modelos Minimalistas . . . . .	9
1.7 Superfície de Energia e o Conceito Funil . . . . .	10
1.8 Projetabilidade . . . . .	14
1.9 Resultados Anteriores e Motivação . . . . .	19
<b>2 Evolução de Proteínas</b>	<b>28</b>
2.1 Explorando a Evolução . . . . .	29

2.1.1	Variações entre espécies homólogas . . . . .	29
2.1.2	Taxa de Evolução . . . . .	30
2.1.3	Regra de Seleção . . . . .	34
2.1.4	Evolução Neutra . . . . .	35
2.2	Bancos de Dados . . . . .	36
2.2.1	O que é Bancos de Dados e Base de Dados . . . . .	36
2.3	Tipos Bancos de Dados . . . . .	37
2.3.1	Banco de Dados Relational . . . . .	38
2.3.2	Banco de Dados Orientado a Objeto . . . . .	38
2.3.3	Banco de Dados de Sequências Biológicas . . . . .	39
2.3.4	Bancos de Dados Primários . . . . .	40
2.3.5	Bancos de Dados Secundários . . . . .	40
2.3.6	Bancos de Dados Especializados . . . . .	41
2.3.7	Falhas dos Bancos de Dados . . . . .	42
2.4	Alinhamento . . . . .	43
2.4.1	Bases Evolucionárias . . . . .	43
2.4.2	Homologia vs. Similaridade . . . . .	44
2.4.3	Similaridade vs. Identidade . . . . .	45
2.4.4	Alinhamento Global vs. Alinhamento Local . . . . .	45
2.4.5	Matriz Dot Plot . . . . .	46
2.4.6	Programação Dinâmica . . . . .	47
2.4.7	Matrizes de Substituição(Pontuação) . . . . .	48
2.5	Alinhamento Múltiplo . . . . .	53
2.5.1	Método Hierárquico . . . . .	54
2.5.2	CLUSTALW . . . . .	54
2.5.3	Cluster Hierárquico(Agrupamentos) . . . . .	55
2.6	Sorting points into neighborhoods (SPIN) . . . . .	57
2.6.1	Estabelecimento Formal do SPIN . . . . .	57
2.6.2	Algoritmo STS . . . . .	58
2.6.3	Algoritmo Neighborhood . . . . .	59
2.7	Hidrofobicidade . . . . .	61
2.8	Escalas Hidrofóbicas . . . . .	63
2.8.1	Método do Particionamento . . . . .	64
2.8.2	Métodos da Área Acessível ao Solvente (ASA) . . . . .	65
2.8.3	Métodos Cromatográficos . . . . .	65

2.8.4	Metagênese de Sítio Dirigido . . . . .	66
2.8.5	Métodos de Propriedades Físicas . . . . .	66
<b>3</b>	<b>Resultados</b>	<b>68</b>
<b>4</b>	<b>Discussão e Considerações finais</b>	<b>81</b>
	<b>Referências Bibliográficas</b>	<b>83</b>
<b>5</b>	<b>Apêndice A</b>	<b>91</b>

# Resumo

Efeito das mutações sobre a estabilidade das proteínas é uma questão crucial na evolução da proteína. Tais efeitos dependem fortemente do caráter hidrofóbico global da proteína. Em um trabalho recente (J. Chem. Phys. **125**,084904(2006)), nós sugerimos dois cenários de enovelamento com conseqüências distintas na evolução da proteína. O limite de baixa hidrofobicidade, corresponde ao regime em que ocorre concomitantemente o colapso e a formação da estrutura nativa. Sob estas condições as proteínas são pouco robustas a mutações, o que implica em uma alta homologia entre proteínas de diferentes espécies. O limite de alta hidrofobicidade, corresponde ao regime em que a proteína sofre um colapso antes do enovelamento, e neste caso as proteínas são mais robustas a mutações, sugerindo uma menor homologia entre proteínas de diferentes espécies. Neste trabalho, nós estudamos a homologia de quatro proteínas para 41 espécies diferentes, correlacionando as suas homologias com suas hidrofobicidades médias. As proteínas estudadas foram lisozima, citocromo-c, mioglobina e histona H3, utilizando seis escalas hidrofóbicas diferentes. Junto com o cálculo da homologia, foi realizada uma comparação da similaridade estrutural (rmsd). Os resultados confirmam a hipótese acima, indicando que proteínas, em condições de baixa hidrofobicidade, têm baixa variabilidade de seqüências e conformações, para alta hidrofobicidade, as proteínas exibem variabilidade de seqüências e conformações.

**Palavras-chave:** evolução, enovelamento de proteína, cenários de enovelamento, projetabilidade, hidrofobicidade, homologia de proteínas, mutação em proteínas

# Abstract

Effect of mutations on stability of proteins is a crucial issue in protein evolution. Such effects depend strongly on the overall hydrophobic protein character. In a recent work we suggested two scenarios for folding with distinct protein evolution consequences (J. Chem. Phys. **125** 084904, 2006) Under low hydrophobic conditions proteins collapse concomitantly with the formation of their native state, and are less robust to mutations, which implies higher homology among proteins of different species. On the other limit, at high hydrophobicity proteins collapse before folding, and in this case they are more susceptible to mutations, suggesting lower homology among proteins of different species. In this work we investigate this conjecture studying the homology of four proteins for 41 different species, correlating it with their average hydrophobicity. The proteins studied were lysozyme, cytochrome-c, myoglobin and histone H3, using six different hydrophobic scales. Along with the homology calculation, a comparison of structural similarity (rmsd) was also carried out. The results confirm the above hypothesis, indicating that proteins at low hydrophobicity display low variations on sequences and conformations. On other hand, at high hydrophobicity, proteins exhibit high variability on sequences and conformations. **Keywords:** evolution, protein folding, scenarios of folding, projetabilidade, hydrophobicity, homology of proteins, mutations in proteins.

## Capítulo 1

# Bioquímica e o Problema do Enovelamento de Proteínas

### 1.1 Importância e Bioquímica

. Elas funcionam como enzimas que catalisam um complexo conjunto de reações químicas. As proteínas também são os reguladores dessas reações, tanto diretamente, como componentes de enzimas, quanto indiretamente na forma de mensageiros químicos, conhecidos como hormônios e também como receptores para esses hormônios. Elas atuam no transporte e armazenamento de substâncias, tais como íons metálicos,  $O_2$ , glicose, lipídeos e muitas outras moléculas. Sob forma de fibras moleculares e outras organizações contráteis, as proteínas geram o movimento mecânico coordenado de inúmeros processos biológicos, como por exemplo, a separação dos cromossomos durante a divisão celular. Proteínas do sistema imune, como as imunoglobulinas, formam um sistema biológico de defesa essencial em animais superiores. Proteínas são elementos ativos fundamentais na expressão genética. Além disso as proteínas têm funções estruturais importantes, como é o caso do colágeno, que dá a força tênsil característica dos ossos, tendões e ligamentos. Fica claro então que as proteínas estão no centro da ação da maioria dos processos biológicos[1]

## 1.2 Estrutura das Proteínas

### 1.2.1 O que é uma proteína?

Proteínas são formadas por seqüências de grupamentos químicos, denominados **aminoácidos** ou **resíduos**, ligados covalentemente entre si por ligação peptídica[3]. Denomina-se de **peptídeo** uma seqüência de poucos aminoácidos, ao passo, que longas seqüências de aminoácidos são designadas por **polipetídeo** ou **proteínas**.

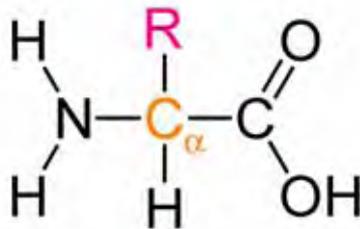


Figura 1.1: Estrutura básica dos aminoácidos

Os aminoácidos são as unidades estruturais básicas das proteínas. Na natureza são encontrados somente 20 diferentes tipos de aminoácidos, denominados de **aminoácidos padrões**. Cada um dos aminoácidos padrões, como podemos ver pela figura 1.1, tem uma estrutura básica composta por um átomo de carbono tetraédrico central ( $C_{\alpha}$ ), ligado a um grupamento amina ( $NH_2$ ), uma carboxila ( $-COOH$ ) um átomo de hidrogênio  $H$  e uma cadeia lateral diferenciada ( $R$ ). A cadeia lateral ( $-R$ ) de cada aminoácido é o que diferencia um aminoácido de outro. Os centros tetraédricos destes aminoácidos, com exceção de um aminoácido, são quirais, existindo apenas o isômero L, em proteínas naturais.

Os aminoácidos ligam-se entre si pela formação de uma **ligação peptídica**. A ligação peptídica é formada quando o grupo carboxílico de um aminoácido se uni ao grupo amina de outro aminoácido, eliminando uma molécula de água. Várias ligações peptídicas dão origem a cadeias poliméricas das proteínas. As proteínas são sintetizadas nos ribossomos, onde o RNA mensageiro é transcrito de acordo com seu código genético: cada tripleto de nucleotídeos especifica um aminoácido correspondente[2].

## 1.3 Níveis Estruturais

Toda proteína tem uma função específica a realizar. A função de uma proteína é melhor compreendida quando se considera as relações tridimensionais entre os átomos que

os compõem. As descrições estruturais das proteínas têm sido tradicionalmente descritas em termos de quatro níveis de estrutura.

### 1.3.1 Estrutura Primária

A **estrutura primária** é dada pela seqüência de aminoácidos ao longo da cadeia polipeptídica. É o nível estrutural mais simples e mais importante, pois dele como veremos mais adiante, deriva todo o arranjo espacial da molécula. São específicas para cada proteína, sendo determinadas geneticamente. A estrutura primária resulta em uma longa cadeia de aminoácidos semelhantes a um colar de contas. Sua estrutura é somente a seqüência dos aminoácidos[4]

### 1.3.2 Estrutura Secundária

O termo **estrutura secundária** é dada pelo arranjo espacial de aminoácidos próximos entre si na seqüência primária da proteína.. Alguns tipos de estrutura secundária são particularmente estáveis e freqüentemente encontrados nas proteínas. Os mais proeminentes tipos de estruturas secundárias são a  $\alpha$  hélice e as folhas  $\beta$  [5].

### 1.3.3 Estrutura Terciária e Quaternária

Enquanto o termo estrutura secundária refere-se ao arranjo espacial dos aminoácidos que estão adjacentes na estrutura primária, a **estrutura terciária** refere-se ao arranjo tridimensional global de todos os átomos em uma proteína. A estrutura inclui aspectos envolvendo distâncias mais longas dentro da seqüência de aminoácidos. No interior da estrutura tridimensional da proteína podemos encontrar aminoácidos que estejam distantes na estrutura primária e ou que pertencem a tipos diferentes estruturas secundárias. Certas proteínas podem conter duas ou mais cadeias polipeptídicas separadas **subunidades**, que podem ser idênticas ou não. O arranjo destas subunidades protéicas em complexos arranjos tridimensionais constitui a **estrutura quaternária**[5].

## 1.4 Classificação e Características dos Aminoácidos

Para entendermos quais forças são responsáveis pela formação da estrutura tridimensional da proteína, deve se compreender as propriedades físico-químicas dos constituintes das proteínas, os aminoácidos padrões. Todas as proteínas de todas as espécies são formadas pelos aminoácidos padrões. A notável gama de funções exercidas pelas proteínas resulta da diversidade e versatilidade desses vinte aminoácidos naturais[2].

Uma boa compreensão das propriedades físico-química dos aminoácidos nos possibilita entender a formação das estruturas complexas das proteínas. Os vinte aminoácidos padrões têm uma parte que é comum a todo aminoácido formada pelo grupo amina, grupo carboxílico e átomo de hidrogênio e outra que varia de aminoácido para aminoácido, que é composta pela **cadeia lateral** [2].

A cadeia lateral de cada um dos aminoácidos pode variar em tamanho, forma, carga, capacidade de formação de pontes, polaridade e reatividade química[5]. A forma mais utilizada na literatura para classificação dos aminoácidos é através da polaridade de cada aminoácido. A polaridade é definida como sendo a tendência do aminoácido em interagir com a água em pH fisiológico ( $pH \simeq 7$ ). Para os aminoácidos naturais a polaridade pode variar desde um comportamento apolar (**hidrofóbico**) até comportamento totalmente solúvel em água (**hidrofílico**). Neste trabalho foi utilizada a classificação de polaridade utilizada em [5] á apresentada na tabela 1.1.

Hidrofobicidade segundo a polaridade dos Aminoácidos

Apolares	Aromáticos	Polares não carregados	Polares carregados (+)	Polares carregados (-)
Alanina (Ala)	Fenilalanina (Phe)	Asparagina (Asn)	Arginina (Arg)	Aspartato (Asp)
Glicina (Gly)	Tirosina (Tyr)	Cisteína (Cys)	Histidina (His)	Glutamato (Glu)
Isoleucina (Ile)	Triptofano (Trp)	Glutamina (Gln)	Lisina (Lis)	
Leucina(Leu)		Serina (Gln)		
Metionina (Met)		Treonina (Thr)		
Prolina (Pro)				
Valina (Val)				

Tabela 1.1: Classificação dos aminoácidos pela polaridade da cadeia lateral segundo [5]

#### 1.4.1 Grupo R apolare/alifático

Os grupos R desta classe de aminoácidos são denominados de hidrofóbicos ou apolares. Estes aminoácidos tendem a se aglomerar para minimizar o contato com as moléculas de água (efeito hidrofóbico)[2].

As volumosas cadeias laterais da **alanina**, **valina**, **leucina** e **isoleucina** são importantes na estabilização das estruturas das proteínas pela promoção das interações hidrofóbicas em seu interior. A **glicina** é o aminoácido de estrutura mais simples, sua cadeia lateral é composta por somente um átomo de hidrogênio. Embora a glicina seja apolar, a sua pequena cadeia lateral não contribui efetivamente para a existência de interações hidrofóbicas. A **metionina**

é um dos dois aminoácidos que contém enxofre, ela possui um grupo tio éster em sua cadeia lateral.

A **prolina** também tem uma cadeia lateral apolar. Mas este aminoácido tem uma característica única. Ele possui uma estrutura em anel, que se liga tanto ao ao átomo de carbono alfa  $C_\alpha$  quanto ao átomo de nitrogênio. Por conta dessa restrição, a prolina influencia pronunciadamente a arquitetura das proteínas, pois impõe maiores restrições conformacionais do que outros aminoácidos[2].

### 1.4.2 Grupos R aromáticos

Os aminoácidos **fenilalanina**, **tirosina** e **triptofano** possuem cadeias laterais relativamente simples. Estes três aminoácidos podem participar de interações hidrofóbicas. A tirosina possui um grupo hidroxila, de modo que este aminoácido pode formar pontes de hidrogênio. Esta característica faz com que a tirosina possa atuar como um grupo funcional importante na atividade das enzimas. As propriedades da tirosina e e do triptofano os tornam mais polares do que a fenilalanina.

### 1.4.3 Grupos R polares não carregados

Os grupos R para esta classificação de aminoácidos, são mais solúveis em água (hidrofilicos), que os aminoácidos apolares por conterem grupos funcionais que formam pontes de hidrogênio com a água. Nesta classe de aminoácidos são incluídos **serina**, **treonina**, **cisteína**, **asparagina** e **glutamina**. As polaridades da serina e da treonina são devidas aos seus respectivos grupos hidroxila; a cisteína tem o seu grupo sulfidril, e a asparagina e glutamina aos seus grupos amina.

Os aminoácido asparagina e glutamina são amidas de dois outros aminoácidos, também encontrados nos aminoácidos naturais, aspartato e glutamato respectivamente. Assim a asparagina e a glutamina podem ser hidrolisados por ácido e base.

A cisteína é facilmente oxidada para formar um aminoácido dimérico unido covalentemente chamado de **cistina**, no qual duas moléculas de cisteína estão unidas por uma ligação dissulfeto. Os aminoácidos unidos por ligações dissulfeto são fortemente apolares hidrofóbicos. As ligações dissulfeto têm um importante papel na estabilização de estruturas de muitas proteínas, em virtude da formação de ligações covalentes entre diferentes partes de uma molécula ou entre duas cadeias protéicas distintas.

#### 1.4.4 Grupos R carregados positivamente (básicos)

Os grupos R mais hidrofílicos são aqueles que são positiva ou negativamente carregados. Os aminoácidos que tem os grupos R com carga líquida positiva em pH 7 são a **lisina**, a **arginina** e a **histidina**. A lisina tem um segundo grupo na posição  $\epsilon$  da sua cadeia alifática. A **arginina** tem o grupo guanidino, que é carregado positivamente. A **histidina** é o único aminoácido padrão que tem um  $pK_a$  próximo da neutralidade. Em muitas reações catalisadas por enzimas, um resíduo de histidina facilita a reação ao servir de doador ou aceitador de prótons.

#### 1.4.5 Grupos R carregados negativamente (ácidos)

Os dois aminoácidos, tendo R como uma carga líquida negativa em pH 7 são o **aspartato** e o **glutamato**, sendo que cada um deles possui um segundo grupo carboxila.

A discussão feita nesta seção anterior se destina a compreender algumas propriedades e características dos aminoácidos padrões que possam contribuir para a estrutura final e estas propriedades serão discutidas mais adiante.

### 1.5 Enovelamento de Proteínas

Inspirado nos artigos de Pauling[6] e Watson e Crick[7], Anfinsen avaliou a atividade da ribonuclease bovina em diversas condições físico-químicas diferentes. Ele observou que para condições desfavoráveis a proteína quebrava cinco pontes dissulfeto e para condições favoráveis a proteína formava novamente cinco ligações dissulfeto e se tornava novamente ativa (renaturava)[8]. Anfinsen encontrou que todas as propriedades físico-químicas da conformação renaturada da ribonuclease eram virtualmente idênticas a sua estrutura nativa.

Com os resultados obtidos mencionados do parágrafo anterior, Anfinsen postulou que a estrutura global da proteína poderia ser atingida sem nenhum maquinário biológico ao acreditar que toda informação para a proteína atingir sua estrutura tridimensional estava contida somente em sua seqüência de aminoácidos. Além disso, Anfinsen postulou, que o estado nativo de uma proteína ( conformação biologicamente ativa), é um mínimo global dentre as energias livres acessíveis. A partir daí, entender os mecanismos pelos quais a seqüência de aminoácido atinge sua estrutura nativa é denominado de **Problema do enovelamento de Proteína**[8]

Posteriormente, Levinthal [10] notou que o espaço das conformações cresce exponencialmente com o número N de aminoácidos da seqüência da proteína. Daí, Levinthal concluiu

que seria impossível, para uma proteína, encontrar o mínimo global (estado nativo) visitando todas as conformações possíveis. Para ilustrar seu raciocínio Levinthal supôs o seguinte: se admitirmos que cada um dos 100 aminoácidos de uma proteína possam assumir 2 orientações possíveis então a proteína poderá assumir  $M = 2^{100} \approx 10^{30}$  conformações diferentes. Continuando, se a proteína gastasse  $10^{-12}$  segundos para visitar uma cada uma das conformações possíveis, então a proteína levaria  $10^{18}$  segundos para se enovelar. Assim a proteína levaria um tempo maior do que a idade do universo para se enovelar[8].

Das observações de Levinthal surgiu um paradoxo. Para resolver este paradoxo, Levinthal admitiu que para a proteína se enovelasse rapidamente, ela não deveria realizar uma pesquisa aleatória entre todas conformações possíveis, mas deveria se enovelar através de uma série de estados intermediários metaestáveis, ou seja, a proteína se enovelaria através de um caminho cinético[12].

### 1.5.1 Interações Energéticas no Processo de Enovelamento

Walter Kauzmann inspirado na visão molecular dos processos biológicos de Pauling[6] e Watson e Crick [7], realizou uma série de experimentos para identificar quais eram os principais tipos de interações que determinavam a estrutura terciária das proteínas. Em seus experimentos Kauzmann [11] notou que as pontes de hidrogênio das cadeias laterais das proteínas, realizadas com as moléculas de água vizinhas, tinham a mesma magnitude do que das pontes de hidrogênio intra-cadeias encontradas no estado nativo da proteína.

Além disso, Kauzmann observou em seus experimentos que as interações mais importantes no, processo de enovelamento da proteína, se referiam a tendência dos aminoácidos hidrofóbicos em associarem-se entre si para diminuir o seu contato com a água, bem como dos aminoácidos polares se associarem com as moléculas de água[24].

Depois do trabalho publicado por Kauzmann, passou-se a analisar a influência da hidrofobicidade no enovelamento de proteína e ainda hoje existem muitas pesquisas querendo compreender os detalhes deste tipo de interação. A visão que emerge é que as interações hidrofóbicas são as mais importantes no processo de enovelamento, deixando para as pontes de hidrogênio e as interações locais (interação eletrostática, interação de van der Waals) responsabilidade pelos detalhes internos da estrutura enovelada[24].

## 1.6 Importância dos Modelos Minimalistas

A estrutura das proteínas envolve milhares de átomos, que interagem entre si e com o meio biológico, tornando difícil o estudo teórico deste sistema. Ainda hoje, é muito custoso

computacionalmente para se modelar uma proteína usando as coordenadas de todos os átomos e os diversos tipos de interações presentes nas proteínas. Por causa disso até hoje são utilizados modelos minimalistas que tentam representar as principais características das proteínas. Estes modelos, que são representações rudimentares das proteínas, por envolverem poucos parâmetros tem a vantagem de se poder calcular as grandezas de interesse de forma exata, sem considerações adicionais ou aproximações.

A simulação computacional de modelos minimalistas tem ajudado no entendimento do processo de enovelamento de proteínas. Isto porque a simulação de proteínas reais envolve uma certa quantidade de parâmetros e interações atômicas de modo que o tempo computacional para simular o enovelamento seria muito grande. Além disso, diversos resultados importantes foram obtidos através de modelos minimalistas[24].

Os trabalhos teóricos em proteínas se iniciaram com simulações em rede tridimensionais de Go [25, 26]. Neste modelo foi utilizado um potencial que continha toda a informação para a proteína se enovelar e por isso é dito que este potencial é não físico[24]. Logo após em 1989, Chan e Dill construíram um novo modelo de rede criando o modelo **HP** (Hidrofóbico-Polar). O modelo HP assume que a proteína pode ser considerada como um heteropolímeros com dois tipos de monômeros, o polar **P** e o hidrofóbico **H**. Tal modelo é capaz de simular o efeito hidrofóbico e prever quais seqüências são capazes de se enovelar[35].

Posteriormente, Chan e Dill desenvolveram um modelo de rede em 3 dimensões onde os 27 monômeros de um polímero são distribuídos em uma rede 3x3x3. Eles tinham o intuito de explorar as conseqüências de um potencial físico sem aproximações, para explorar as propriedades do espaço conformacional e suas seqüências geradas. Shakhnovich e Gutin enumeraram exaustivamente todas as conformações maximamente de um polímero de 27 monômeros em uma rede 3x3x3, encontrando um total de 103346 conformações diferentes. Mesmo o modelo de rede sendo uma grande aproximação para o que ocorre nas proteínas naturais, foi possível obter resultados importantes sobre o enovelamento de proteínas porque o modelo de rede detém as idéias essenciais das proteínas.

## 1.7 Superfície de Energia e o Conceito Funil

O conceito de funil e a superfície de energia, junto com a nova geração de experimentos contribuíram para o entendimento teórico e experimental do complicado processo de enovelamento de proteína. Uma nova abordagem surgiu então a partir de uma descrição estatística do relevo de energia para processo de enovelamento[8, 12, 13, 14].

A teoria do relevo de energia para o conceito de funil figura (1.2), nos fornece um

referencial teórico para a compreensão global do processo de enovelamento. Para o conceito do funil o relevo de energia representa a energia livre de todos estados conformacionais, com base energética dominante. A maior parte da topografia da superfície de energia pode ser dita afunilada e as suas sequências são consideradas bem projetadas.

Leopold [13] define o funil de enovelamento, como uma coleção de estruturas colapsadas geometricamente similares. Para o modelo do funil uma proteína é dita enovelável se ela tem um estado nativo termodinamicamente estável e acessível em seu funil de enovelamento. Por outro lado, uma proteína é não enovelável, quando o seu relevo de energia possui vários mínimos globais e múltiplos funis de enovelamento[12]

O potencial termodinâmico que descreve um processo a pressão e temperatura constantes, como é o enovelamento de proteína em meio biológico, é a energia livre de Gibbs ( $\Delta G = E - T\Delta S$ ). Por isso o sistema proteína-solvente pode ser descrito como função da energia livre e do espaço das conformações da proteína. Desta forma, Leopold demonstrou que se pode obter uma descrição detalhada da superfície de energia de uma proteína especificando sua energia livre média sobre a coordenada de reação em função das coordenadas de cada átomo na proteína. Assim o relevo de energia da proteína apresentará mínimos locais, que são pequenas excitações de energia correspondentes a mudanças conformacionais locais e individuais. Essas energias são da ordem de  $K_B T$ , ou seja, flutuações de energia dos átomos da proteína[14].

Observando a figura 1.2 podemos notar muitas estruturas com altas energias e poucas com baixas energias. Quanto mais próximo a proteína se encontra do seu estado nativo, mais baixa é a energia dessas estruturas. Uma proteína em seu estado desenovelado encontra seu estado nativo dinamicamente através de caminhos presentes em seu relevo de energia. Tais caminhos são determinados pela forma e rugosidade encontradas no relevo de energia de uma proteína. À medida que a proteína se aproxima do seu estado nativo, mais similar torna se a estrutura m relação a estrutura nativa e menor é o número de estados acessíveis. O conceito venceu o paradoxo de Levinthal, que é a improbabilidade estatística da proteína encontrar seu estado nativo em tempo funcional por tentativas aleatórias.

Dessa forma a dinâmica de enovelamento, para o conceito do funil, pode ser considerada como um processo, em que um conjunto de estruturas colapsadas similares de baixa energia, caminham para o estado nativo com sua energia livre associada. O relevo de energia de uma proteína no processo de enovelamento, assemelha-se a um funil parcialmente rugoso com armadilhas locais, onde a proteína pode residir transientemente como na figura 1.3. Não existe uma um único caminho de enovelamento, mas uma multiplicidade de rotas convergentes que seguem para a estrutura nativa[17].

Os trabalhos de Socci e Onuchic demonstraram que para modelos de rede de het-

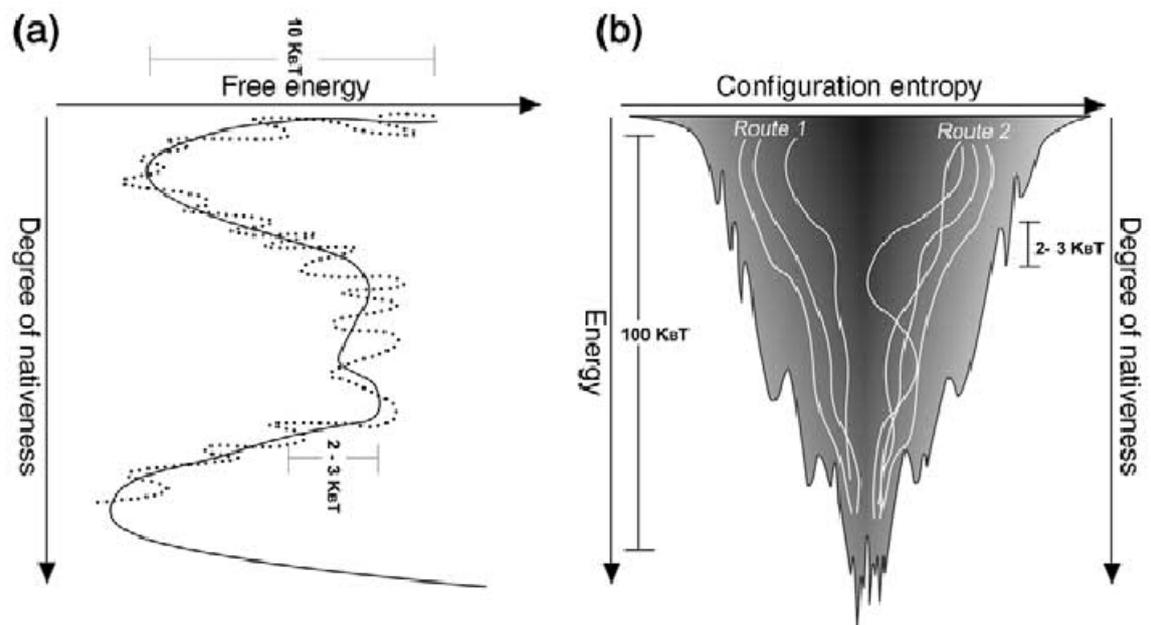


Figura 1.2: A figura (a) no eixo x a energia livre e o eixo y o grau de similaridade com o estado nativo. No gráfico (b) estão representados a entropia de configuração e a coordenada de reação em relação ao estado nativo e da energia, respectivamente nos eixos x e y. Como podemos notar na figura (b) não existe um caminho, mas uma rotas, que permitem vários caminhos que a proteína pode tomar para se enovelar. Na parte superior da figura (b) encontramos estruturas com altas energias. Na medida que a proteína vai se enovelando, menor torna-se o valor de energia, bem como a variação de entropia. Ao atingir a estrutura de menor energia e menor variação de entropia a proteína se encontra enovelada. Comparando a figura (a) com a (b) notamos que as flutuações em energia vão diminuindo até atingir seu estado nativo

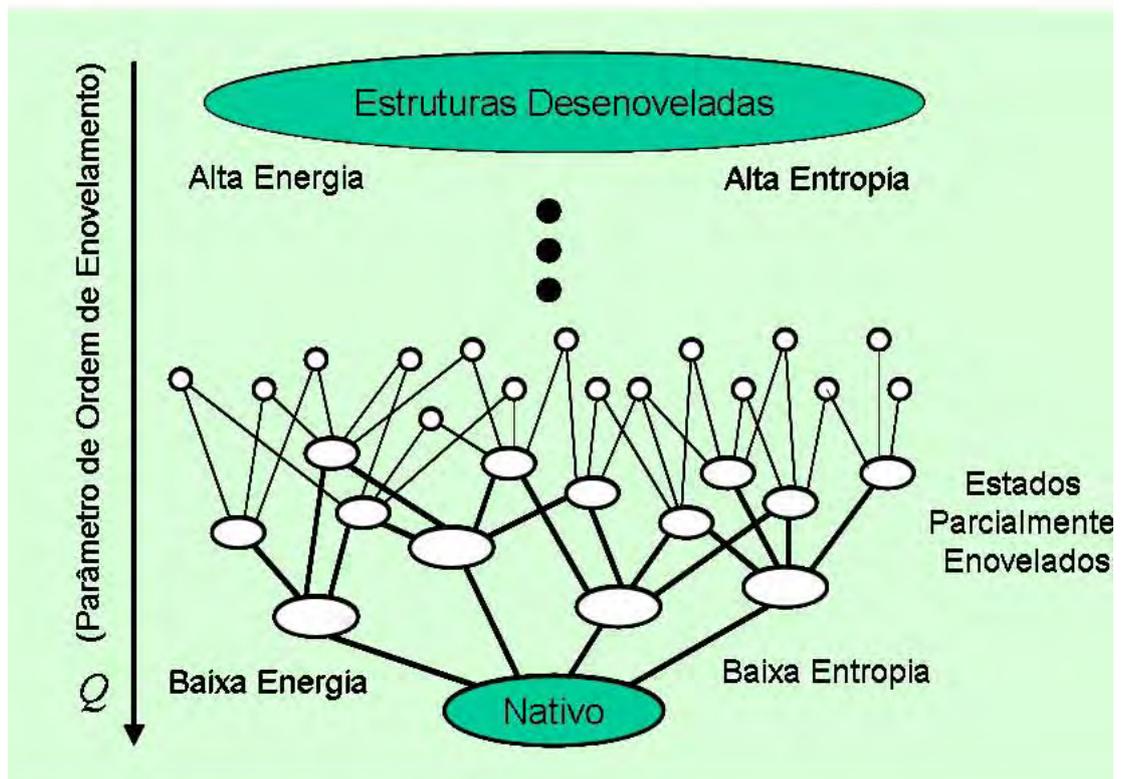


Figura 1.3: Esquema de conformações de uma proteína com relação ao grau de similaridade com o estado nativo. No eixo horizontal temos a entropia e no vertical, a coordenada de reação, ou seja, grau de similaridade da estrutura com o estado nativo. No topo da figura existe alta entropia, pois o estado desenovelado possui um ensemble de estruturas muito grande. Quando a proteína se enovela, move-se para baixo na coordenada de reação, dessa forma o espaço conformacional e a energia é reduzida drasticamente e a proteína assim encontra seu estado nativo que é único, estável termodinamicamente e cineticamente acessível.

eropolímeros as proteínas que se enovelam mais rapidamente ou seja, que tem menor tempo mínimo de enovelamento maximizam a razão  $T_f/T_g$ . A temperatura de vidro é definida como sendo aquela abaixo da qual a proteína fica armadilhada em mínimos locais, impossibilitando assim o enovelamento completo. Já a temperatura de enovelamento  $T_f$  é definida como sendo, aquela em que temos uma probabilidade de 50% de encontrar a proteína em seu estado nativo. Podemos pensar a temperatura  $T_f$  como caracterizando as interações no estado nativo e  $T_g$  como caracterizando o conjunto das interações não nativas presentes em outras configurações[17]. Assim quando a inclinação do funil é dominante sobre a rugosidade do relevo, o enovelamento é rápido e a competição entre a inclinação do relevo e a rugosidade pode ser avaliada pela razão  $T_f/T_g$ . [17]

Seqüências de proteínas, com rápidos enovelamentos, além de possuem grandes valores para a razão  $T_f/T_g$ [18] eliminam ao máximo a frustração no estado nativo (**frustração mínima**). Estas estruturas nativas são robustas a mutações nas sequências, de modo que seja fracamente dependente de variações que possam ocorrer no meio ambiente do enovelamento ou de mutações. De outra forma a menor variação em pH, temperatura ou alguma mutação afetasse a configuração da estrutura nativa, isso poderia favorecer uma outra estrutura nativa[17]. Ou ainda se a proteína já tivesse atingido uma forma estrutural útil para um determinado organismo ancestral uma mutação poderia destruir sua forma útil, o que poderia em última análise, provocar a morte do organismo mutado[36].

A robustez de uma proteína é obtida através da seleção natural, escolhendo seqüências em que as interações presentes na estrutura funcionalmente útil não estejam em conflito, ou seja, de forma que, as interações na estrutura nativa devam ser minimamente frustradas[8].

## 1.8 Projetabilidade

Se por um lado estudos teóricos enfatizam que seqüências que se enovelam mais rápidas são aquelas que tem seu estado minimamente frustrado[8]. Por outro lado temos resultados que sugerem que as estruturas minimamente frustradas comportam um grande número de seqüências que se enovelam em uma mesma estrutura. Estes dois resultados sugerem que certas estruturas possuem uma certa tolerância a mutações.

Uma das perguntas sempre pertinente ao processo de enovelamento é: Como uma proteína sempre se enovela em seu estado nativo único. A resposta é :a **evolução**. Sequências aleatórias de aminoácidos usualmente podem estar armadilhadas (glassy) e usualmente não possuem um estado nativo único. Mas proteínas naturais não são seqüências aleatórias. Elas formam uma classe de moléculas biológicas que são selecionadas na natureza pela evolução.

Encontramos em cada proteína um estado nativo de menor energia, que está bem separado dos outros estados[19].

Existem entre 50 a 100 mil proteínas diferentes no corpo humano e um número muito maior de proteínas no mundo biológico[19]. As estruturas das proteínas podem ser classificadas para diferentes estruturas enoveladas utilizando, por exemplo, as classificações SCOP[20] e CATH[21]. Proteínas que possuem a mesma estrutura enovelada (**motif**) tem, em sua maioria as mesmas estruturas secundárias com o mesmo arranjo e as mesmas conexões topológicas, a menos de pequenas variações[19].

Proteínas com relações evolucionárias próximas, freqüentemente tem alta similaridade de seqüência e estrutura. Contudo é intrigante que uma mesma estrutura enovelada, possa ser compartilhada por proteínas de diferentes origens evolucionárias e funções biológicas. Chothia [22], foi o primeiro que observou a desigualdade existente entre o número de estruturas terciárias e seqüências de proteínas. A partir desta observação, ele estimou que para as proteínas naturais existia aproximadamente 1000 estruturas terciárias **fold**s diferentes. Entre as características aparentes destes folds estão a hélice- $\alpha$  e a folha- $\beta$ , regularidades e simetrias. Entretanto como no caso de seqüências, estruturas de proteínas ou folds formam uma classe especial [19].

Na figura 1.4 comparamos o crescimento no número de seqüências depositadas no **PDB** com o número de estruturas enoveladas. Os pontos pretos representam o número de seqüências diferentes acumuladas ao longo dos anos (eixo x). Os pontos vermelhos referem se ao número estruturas terciárias (mesmos motifs) distintas acumuladas, segundo o critério SCOP, que foram observadas ao longos dos anos no PDB.

Por causa da grande desigualdade existente entre os números de estruturas e seqüências, foi utilizada uma escala logarítmica na figura 1.4. Analisando o gráfico da figura, é fácil notar que ao longo dos anos a desigualdade entre o número de estruturas de seqüências e o número de estruturas terciárias (folds) distintas (mesmos motifs), segundo o critério SCOP só tem se tornado mais acentuado, confirmando as hipóteses de Chothia [22], de que as proteínas constituem uma classe relativamente pequena de estruturas tridimensionais relevantemente distintas[19].

Existe algo especial a cerca dos folds de proteínas naturais, estas estruturas são meramente resultados da evolução ou existe alguma razão fundamental atrás de sua evolução. Um dos primeiros a atentar sobre este assunto foi Finkelstein e colaboradores. Eles argumentaram que certos motifs mais fáceis de estabilizar e desta forma mais comum, ou porque eles possuem os menores valores de energia ou porque tem um espectro não usual sobre seqüências aleatórias

Depois da descoberta da alta desigualdade observada entre as seqüências e estruturas das proteínas, alguns pesquisadores propuseram a hipótese da **projetabilidade**. O "princípio da projetabilidade" foi proposto como um mecanismo da seleção natural para estruturas das

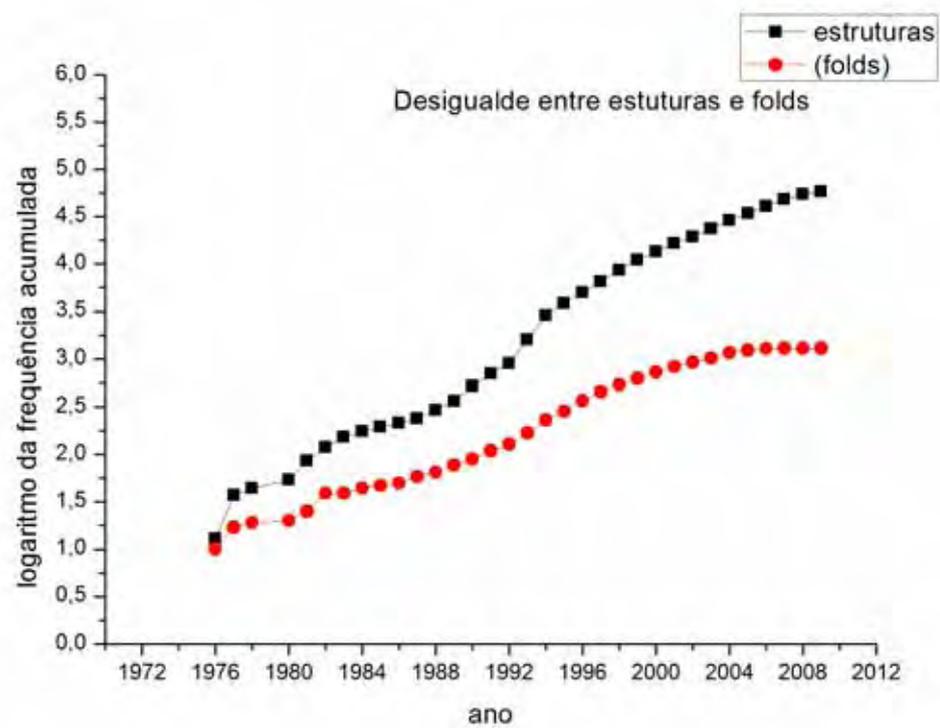


Figura 1.4: **Projetabilidade:Relação entre o logaritmo das freqüências acumuladas para as estruturas terciárias e as estruturas enoveladas segundo o critério SCOP. Os dados foram retirados das estatísticas apresentadas no banco de dados PDB**

proteínas [32, 33, 34] A projetabilidade de uma estrutura pode ser definida grosseiramente como o número de sequências de proteínas diferentes que tem a mesma estrutura terciária. De acordo com esta visão as diferenças nas propriedades geométricas podem tornar, ou não uma estrutura com maior ou menor projetabilidade. O fato é, que podemos verificar a projetabilidade de uma estrutura contando o número sequências diferentes que se enovelam em uma mesma estrutura. Uma estrutura nativa que tem um grande número de sequências diferentes é dita bem projetável, ao passo que uma estrutura com um pequeno número de sequências diferentes podem ser denominadas de estrutura mal projetada. Por causa disso, podemos afirmar que a entropia do número de sequências sobre uma estrutura pode variar de estrutura para estrutura [39].

A primeira consideração sobre esta questão foi realizada por Finkelstein em [37]. Ele apresentou uma gaussiana fenomenológica da distribuição do número de seqüências adotadas pelas estruturas de proteínas.[39]. Posteriormente Shakhovich apresentou em [38] uma justificativa analítica para a hipótese da Distribuição Gaussiana de Finkelstein e apontou suas falhas. Depois Wolynes [31], usando uma analogia entre estatística de seqüências e mecânica estatística de modelos simples de spin, analisou a projetabilidade para a uma aproximação mais simples que a de [38] e encontrou uma gaussiana equivalente a gaussiana encontrada por Finkelstein. Ele concluiu então que as estruturas com maior densidade(maior número de contados) são mais projetáveis, concordando com a hipótese de Finkelstein[39].

Posteriormente Li et al.[32] apresentaram um estudo da enumeração completa de um polímero de 27 monômeros, para um modelo HP, de duas letras, em uma rede 3x3x3. Eles pesquisaram, o estado fundamental para todas as  $2^{27}$  seqüências possíveis usando um conjunto completo de 51704 estruturas distintas não relacionadas por simetria. Eles encontraram, que uma das estruturas pode ser projetada (serve de estrutura nativa) para nada menos do que 3794 seqüências. Por outro lado, eles também encontraram, que 4256 estruturas não representam estado nativo para nenhuma das seqüências possíveis[39].

Recentemente England e Shakhovich[40] explorando a analogia entre estatística de seqüências e a mecânica estatística de vidro calcularam o número de seqüências que podem se enovelar em uma dada conformação com uma energia desejada. Para este fim, estes autores obtiveram uma expressão para a energia livre (no espaço de seqüências) de todas as seqüências que se enovelam em uma estrutura:

$$F_{seq} = -\frac{\beta}{4}(Trv^2)(TrC^2) - \frac{\beta}{8}(Trv^4)(TrC^4) \quad (1.1)$$

onde  $\beta$  é o inverso da temperatura de seleção de seqüência,  $Tr$  é o traço da matriz,  $v$  está relacionada à energia de interação dos aminoácidos e  $C$  é a matriz de contato da estrutura da

proteína

Usando uma analogia simples da termodinâmica podemos encontrar a entropia do espaço de sequências para uma cada estrutura, tomando o logaritmo do número de sequências que se enovelam em uma estrutura dada pela energia 1.1. Deste modo fica claro, que o conjunto de traços  $Tr^*$  (ou o maior autovalor  $\lambda_{max}$  da matriz de contatos  $C$ ) determinam a entropia no espaço de sequências, e desta forma, a projetabilidade da estrutura terciária de uma proteína. O primeiro termo da equação (??  $TrC^2$  é proporcional ao número de contatos entre os resíduos de uma dada estrutura (densidade de contatos DC). Utilizando a aproximação do primeiro termo da equação 1.1, encontra-se a aproximação original de Finkelstein. Se for considerados mais termos, além do primeiro termo da equação 1.1, pode se obter diferenças geométricas de estruturas para um mesmo valor de DC[40].

Recentemente, B. Shakhnovich e outros sugeriram, que a densidade de contatos (DC) das estruturas está significativamente correlacionada com o tamanho da família da proteína, de tal forma que estruturas com alta densidade de contatos possuem um grande número de sequências e estruturas com baixa densidade possuem um número pequeno de sequências diferentes. Eles encontraram que a diversidade funcional está correlacionada com DC. Entretanto deve ser considerado o fato, de que baixos valores de DC tipicamente correspondem a pequenas famílias de proteínas, e altos valores de densidade de contatos podem ocorrer tanto em pequenas famílias quanto grandes famílias de proteínas. Portanto, deve-se salientar que a projetabilidade não é um conceito que sozinho, pode determinar o tamanho de uma família de proteínas. De fato, processos evolucionários devem atuar juntos com as interações físicas para se determinar o tamanho de uma família de proteínas

Bloom e seus colaboradores, olhando diretamente para manifestações evolucionárias da projetabilidade, analisando a taxa de evolução no fermento. O modo escolhido para determinar a taxa de evolução, foi através da taxa de divergência da sequência de genes ortológicos (de espécies diferentes) de duas espécies relacionadas, avaliando o número de substituições não sinônimas por sítio DN. Usando o genoma completo de duas espécies de fermento, *S. cerevisiae* e *S. bayanus*, os pesquisadores calcularam a quantidade de divergência DN, para aproximadamente 200 genes capacitando assim à predição de estrutura, que foi realizada. O ganho destes resultados é que a quantidade DN está relacionada com muitas propriedades das proteínas[49].

Bloom e seus colaboradores encontraram que a densidade de contato (DC) (ou o máximo autovalor da matriz de contato) está relacionada positivamente com a taxa evolucionária. De fato, as estruturas com alta projetabilidade podem acomodar mais sequências por serem mais tolerantes e isto pode ser seguido uma maior divergência de sequências além de um aumento na taxa evolucionária. Não obstante, somente a projetabilidade ou somente

a termodinâmica não explicam ou podem explicar a robustez das proteínas encontradas na natureza[39].

## 1.9 Resultados Anteriores e Motivação

Os efeitos de mutação sobre a estabilidade de proteínas é uma questão crucial na evolução das proteínas. Vários trabalhos enfatizam o princípio de frustração mínima, afirmando que estas seqüências boas devem ser as mais otimizadas, de modo que elas são as que satisfazem as melhores condições de estabilidade termodinâmica, unicidade do estado nativo e acessibilidade cinética[18]. Por outro lado, se as proteínas devem ser selecionadas através da seleção natural deve se aceitar alguma tolerância na otimização, permitindo que proteínas com uma certa estabilidade termodinâmica se enovelam em uma mesma estrutura.

Deste modo, pode se aceitar mutações na seqüência de uma proteína, desde que esta mutação mantenha uma certa estabilidade termodinâmica e eficiência cinética no processo de enovelamento[50]. A tolerância na substituição de aminoácidos, foi observada experimentalmente por [51, 52, 53]. A motivação inicial dos nossos trabalhos, iniciado durante o mestrado, foi compreender:

1. como as mutações afetam seqüências otimamente desenhadas
2. a regra da hidrofobicidade no processo de enovelamento diante de mutações.

Muitos trabalhos[8, 41, 42, 43, 44, 45] têm sido desenvolvidos para responder se a formação de conformações compactas devidas ao efeito hidrofóbico ajudam a proteína enovelar-se ou não? Ou de outra forma, em qual cenário a proteína preferiria se enovelar? Na literatura são classificados dois cenários diferentes para o enovelamento de proteína. No primeiro cenário, a proteína sofre um rápido colapso não específico de estruturas seguido de um lento rearranjo até alcançar a estrutura nativa. No outro cenário, a proteína se enovela diretamente em seu estado nativo, sem a presença de um colapso de estruturas.

Existem trabalhos que mostram que algumas proteínas sofrem um rápido colapso seguido de seu posterior enovelamento completo em sua estrutura nativa [54, 55, 56]. Existe a evidência experimental [43] de que algumas proteínas se colapsam concomitantemente com a formação da estrutura nativa. Alguns estudos desenvolvidos por [8, 41, 42] correlacionam a cinética de enovelamento com quatro parâmetros definidos a seguir:

1. A temperatura de enovelamento  $T_f$ , definida como sendo a temperatura na qual metade da cadeia está enovelada ou como se diz em modelos de proteínas usados em simulações,

é a temperatura na qual a probabilidade de se encontrar a proteína com seus contatos nativos formados é de 50% ( $P(N) = 0.5$ );

2. A temperatura de vidro  $T_g$  é definida como aquela em que a cinética é dominada por armadilhas, por causa da presença de muitos mínimos no relevo de energia, o que por sua vez provoca o desvio do comportamento exponencial, tornando o enovelamento em uma estrutura nativa impossível.
3. A temperatura  $T_\theta$ , associada com o colapso não específico, que representa o rápido colapso hidrofóbica de algumas proteínas;
4. O gap de estabilidade  $gap = E_1 - E_0$ , que nada mais é do que a diferença entre o primeiro estado excitado e o estado nativo.

Estas grandezas têm sido usadas para definir quantidades adimensionais, que possam ser correlacionadas com as taxas de enovelamento[50].

Um desses parâmetros adimensionais, definido como  $\sigma = (T_\theta - T_f)/T_\theta$ , foi desenvolvido por Klimov e seus colaboradores [44, 45]. Através de simulação computacional, eles encontraram, que proteínas com rápidos enovelamentos, têm pequenos valores para  $\sigma$ . Este resultado implica que a proteína está se enovelando diretamente em seu estado nativo. Se por um lado os dados experimentais de espalhamento de raios X e dicroísmo circular produzidos por [43] corroboram os resultados teóricos de Klimov e seus colaboradores. Por outro lado, Chiu e Goldstein, utilizando a equação de difusão, demonstraram que proteínas marginalmente estáveis se enovelam rapidamente na presença de um colapso não específico, favorecendo a formação de estados compactos[50].

Gutin et al.[46], analisando o colapso de proteínas de 27 monômeros, para um modelo de rede 3x3x3, concluíram que se a atração global entre os aminoácidos é dominante, então um rápido colapso de estruturas compactas precede um lento rearranjo, até a proteína se enovelar completamente em seu estado nativo [50]. Os requisitos para um rápido enovelamento, proposto por [8, 43], é que a razão  $T_f/T_g$  seja maximizada. Socci e Onuchic [15, 16] demonstraram que ao maximizar a razão  $T_f/T_g$ , estamos minimizando o tempo de enovelamento da proteína[50]. Um outro critério, desenvolvido por Sali[41] correlaciona rápidos enovelamentos com grandes gaps de energia. Chiu e Goldstein[42] correlacionaram também rápidos enovelamentos a outra grandeza, que é o Zscore, definido por 1.2

$$Zscore = \frac{\langle E \rangle - E_0}{\sigma} \quad (1.2)$$

onde  $\langle E \rangle$  é o valor médio das energias,  $E_0$  é a energia do estado nativo e  $\sigma$  é o desvio padrão. A proposta de nossos trabalhos no mestrado, que tem seus resultados apresentados em

[50], foi analisar o modo pelo qual variam as taxas de enovelamento com relação à temperatura e também com seu grau de frustração diante a mutações realizadas sobre um polímero de 27 monômeros, através de modelo de rede 3x3x3, para diferentes cenários hidrofóbicos[50]. A energia configuracional de um estado foi dada por 1.3:

$$E = N_l E_l + N_u E_u \quad (1.3)$$

onde  $N_l$  é o número de contatos não covalentes entre monômeros de mesmo tipo *like contacts* representados por  $E_l$ .  $N_u$  é o número de contatos não covalentes entre monômeros de tipos diferentes, com energia  $E_u$ .

A seqüência escolhida, para se realizar as mutações foi a seqüência 012. A cadeia desta seqüência é dada por ABABBBCBACBABABACACBACAACAB. Esta seqüência foi escolhida por possuir uma razão  $T_f/T_g = 1.6$ . O estado nativo da seqüência 012 está entre as 103346 conformações maximamente compactas para um polímero de 27 monômeros em uma rede 3x3x3. As conformações maximamente compactas possuem 28 contatos não covalentes e a seqüência 012 não possui nenhum dos seus contatos não covalente entre monômeros de diferentes monômeros (frustrados).

Com o intuito manter a proporção de 11 monômeros do tipo A, 10 monômeros do tipo B e 6 monômeros C, as mutações foram realizadas segundo a seguinte regra: Toda vez que se encontrava dois monômeros diferentes ao longo da cadeia, permutava-se, trocando de posição estes monômeros de lugar. Cada permutação era considerada uma mutação simples e a cada mutação simples se considerava uma nova seqüência mutada diferente (mutação simples). Nesta perspectiva, mutação dupla é quando ocorre duas mutações simples concomitantemente.

Para uma mutação simples, foram encontradas 236 seqüências mutadas. Das quais, 206 seqüências possuíam estrutura nativa única, sendo que deste total, 198 (96%) tinham a mesma estrutura nativa. No caso de mutações duplas foram encontradas 19815 seqüências mutadas. Destas 8933 tinham estado nativo único, sendo 3991(45%) seqüências com a mesma estrutura nativa.

Para todas as seqüências, com estado nativo único, foram avaliadas as grandezas de estabilidade termodinâmica, gap de energia, Zscore bem número de contatos frustrados (entre monômeros diferentes no estado nativo, para as 103346 conformações maximamente compactas. Os resultados estão apresentados na figura 1.5

Os dois gráficos do lado esquerdo referem se ao caso de 1 mutação simples e os da direita para o caso de 2 mutações. Os gráficos superiores referem se a relação entre gap de energia e Zscore, para as seqüências com 1 mutação(lado esquerdo) e 2 mutações(lado direito). Os dois gráficos inferiores representam a relação entre o número de contatos frustrados  $f$  no

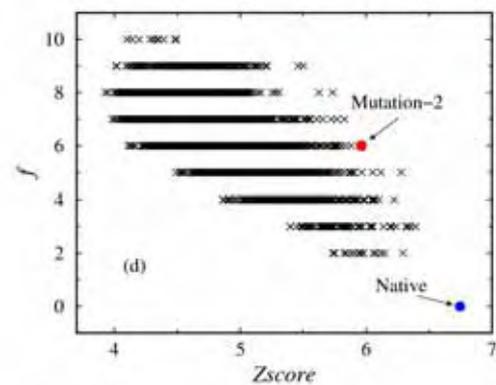
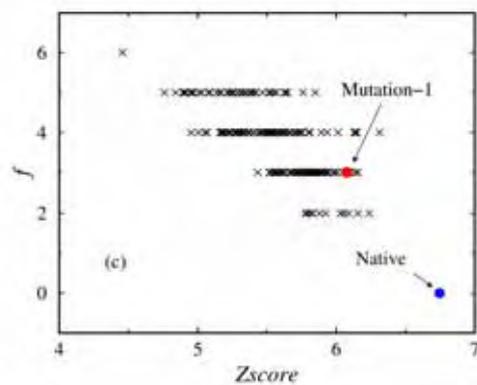
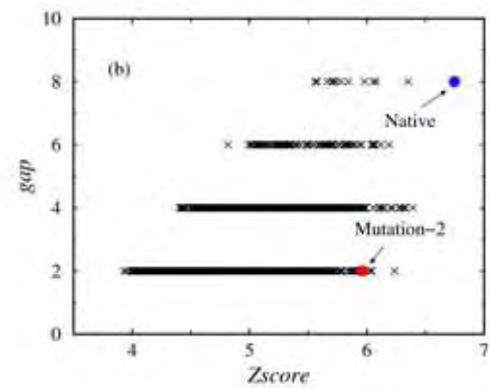
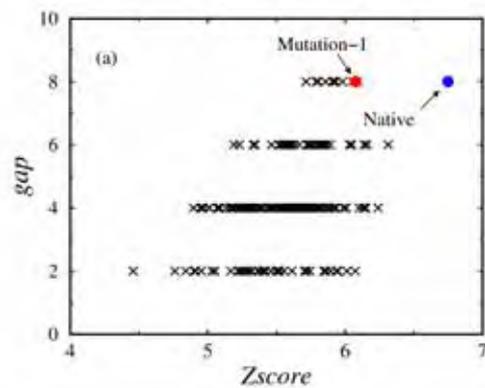


Figura 1.5: O gráfico na parte superior esquerda temos a relação entre o gap de energia vs.  $Zscore$  para as seqüências com mutações simples com estado nativo único. O gráfico da esquerda se refere à relação do número de contatos frustrados  $f$  (entre monômeros diferentes) no estado nativo vs.  $Zscore$  para as seqüências com mutações simples. Do lado direito superior temos a relação entre gap de energia vs.  $Zscore$  para as seqüências com mutações duplas com estado nativo único. No gráfico da parte inferior direita temos a relação entre número de contatos frustrados  $f$  (entre monômeros diferentes) no estado nativo vs  $Zscore$ .

estado nativo versus Zscore , para 1 mutação (lado esquerdo) e 2 mutações lado direito. O ponto azul refere-se aos valores da sequência 012.

Ao analisarmos a relação do gap com o Zscore, é fácil notar, que tanto para sequências com 1 mutação, quanto para 2 mutações, existem certas sequências que possuem gap e Zscore consideravelmente elevados, se comparados aos valores da sequência nativa. Estes resultados nos sugeriram avaliar os efeitos sobre uma sequência mutada. Escolhemos duas sequências mutadas para avaliarmos os efeitos, que a mutação provoca na cinética e termodinâmica para dois cenários hidrofóbicos diferentes. Foi escolhida uma sequência com uma mutação simples com 3 contatos frustrados no estado nativo e grandes valores de gap de energia e Zscore. A outra sequência mutada escolhida, foi uma sequência com mutação dupla e seis contatos frustrados no estado nativo, que possuía um grande valor de Zscore mas um baixo valor de gap de energia

Podemos analisar o comportamento do sistema através dos parâmetros de hidrofobicidade:

$$\bar{E} = \frac{1}{2}(E_l + E_u) \quad (1.4)$$

$$E_{het} = (E_u - E_l) \quad (1.5)$$

A grandeza da equação (1.4) indica a formação de contatos. De modo, que se esta grandeza for menor do que zero a formação de contatos é favorecida e a cadeia sofre um colapso não específico. A grandeza da equação (1.5), nos fornece uma idéia da rugosidade do relevo, que é determinada pela heterogeneidade dos diferentes tipos de aminoácidos, determinando os níveis de energia e assegurando que a cadeia seja um heteropolímero. Para avaliarmos o grau de compactação, ou melhor, a hidrofobicidade através da equação (1.6),

$$k = -\frac{\bar{E}}{E_{het}} \quad (1.6)$$

a equação (1.6) é ajustada para o valor de k varie entre zero e um. De modo que para k=1, a proteína esteja submetida a um potencial para um regime de alta hidrofobicidade e para k=0 um potencial de baixa hidrofobicidade. Ao escolher um potencial de alta hidrofobicidade é favorecida a formação de estruturas colapsadas. O enovelamento, em alta hidrofobicidade, ocorre em um cenário em que a cadeia sofre um rápido colapso (colapso não específico), seguido de um lento rearranjo até atingir o estado nativo. Por outro lado ao escolher um potencial de baixa hidrofobicidade as interações entre monômeros diferentes tem a mesma intensidade, módulo do que as interações entre monômeros diferentes, deste modo este potencial não favorece o aparecimento do colapso da proteína. No cenário de enovelamento, para baixa hidrofobicidade, a proteína a compactação e o enovelamento ocorrem simultaneamente diretamente na estrutura nativa da proteína ( colapso específico).

Para avaliarmos os efeitos das mutações na termodinâmica e cinética, selecionamos duas sequências mutadas. Uma das sequências escolhidas tinha 3 contatos nativos frustrados ( $f=3$ ) na sua estrutura nativa e tinha sofrido uma mutação simples. A outra sequência mutada, que foi escolhida entre as mutações duplas, possuía 6 contatos frustrados ( $f=6$ ) na sua estrutura nativa. Ambas sequências escolhidas tinham grandes de Zscore, mas a sequência de uma mutação simples tinha um alto valor de gap de energia e a sequência com mutação dupla tinha baixo valor de gap de energia.

Realizamos simulações, que avaliaram a termodinâmica e a cinéticas das duas sequências mutadas bem como da sequência nativa (sequência 012) para dois regimes hidrofóbicos. Para um regime de alta hidrofobicidade (AH), escolhemos um potencial com  $E_l = -3$  e  $E_u = -1$ . Para o regime de baixa hidrofobicidade foi escolhido um potencial com  $E_l = -3$  e  $E_u = 3$ .

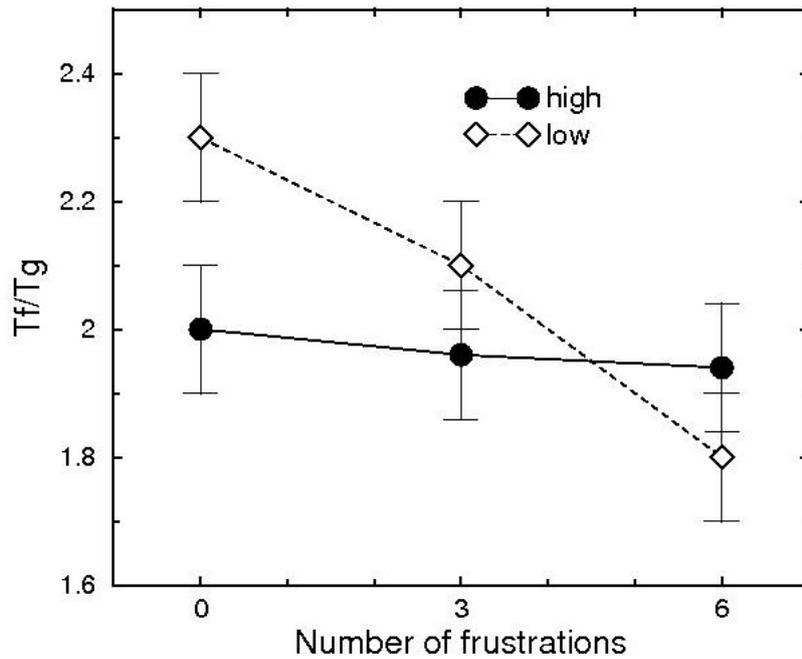


Figura 1.6: Parâmetro de envelhecimento  $T_f/T_g$  como função do número de contatos frustrados  $f$  na conformação nativa e hidrofobicidade.  $f=0$ ,  $f=3$  e  $f=6$  correspondem respectivamente a sequência nativa, uma mutação e duas mutações

As grandezas, de tempo de envelhecimento  $\tau$ , as temperaturas  $T_f$  e  $T_g$  bem como a razão  $T_f/T_g$ , foram avaliadas nos regimes de AH e BH para as 3 sequências. Em ambos regimes

hidrofóbicos, pode se observar que a grandeza  $T_g$  não dependia do grau de frustração, o que era de se esperar, já que a grandeza  $T_g$  está associado com a rugosidade do relevo de energia e este não depende dos detalhes da seqüência.

Entretanto foi observado, que a grandeza  $T_f$  diminui significativamente com o aumento da frustração  $f$  para o regime BH e se manteve praticamente constante para o regime de AH. Os resultados destas observações podem ser resumidas pela figura 1.6

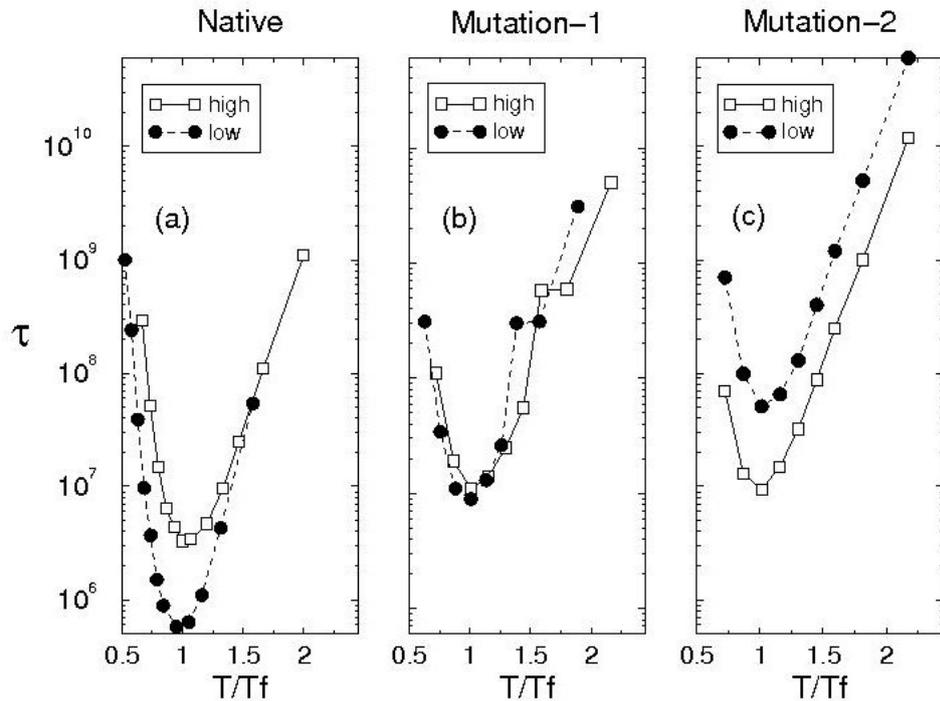


Figura 1.7: Tempos de enovelamento para seqüências: nativa(native), 1 mutação(Mutation-1) e 2 mutações(Mutation-2) respectivamente da esquerda para direita, para dois regimes hidrofóbicos diferentes. A linha contínua refere se ao regime de alta hidrofobicidade e a linha tracejada se refere ao regime de baixa hidrofobicidade. Os gráficos, da esquerda para a direita, se referem a relação de  $\tau$ (medido em passo de Monte Carlo) Vs.  $T/T_f$  para as seqüências nativa, com uma mutação e com duas mutações respectivamente

A cinética das 3 seqüências foi avaliada medindo se o tempo de enovelamento  $\tau$  para 100 corridas independentes. Doze valores de  $\tau$  foram calculados para 100 simulações para um intervalo de temperatura  $1.2 < T < 3.0$ . Os resultados destes cálculos estão apresentados

em 1.7

Observa-se na figura 1.7, a forma em U da curva de  $\tau$  em função de  $T/T_f$ , para todas as seqüências nos dois regimes hidrofóbicos. Ao olharmos para a seqüência nativa, observamos que ela se enovela mais rapidamente no regime de baixa hidrofobicidade (BH) do que no regime de alta hidrofobicidade. Para a seqüência com uma mutação ( $f=3$ ), a cinética para os regimes de AH e BH tem comportamentos similares e os tempos de enovelamento sofrem um aumento relativo, dentro da ordem de grandeza. Contudo, se analisado o comportamento a seqüência com duas mutações, observa-se que o regime de BH torna-se bem mais lento do que no regime de AH, de modo que nota-se uma inversão nos comportamentos na medida que ocorre o aumento da frustração na seqüência.

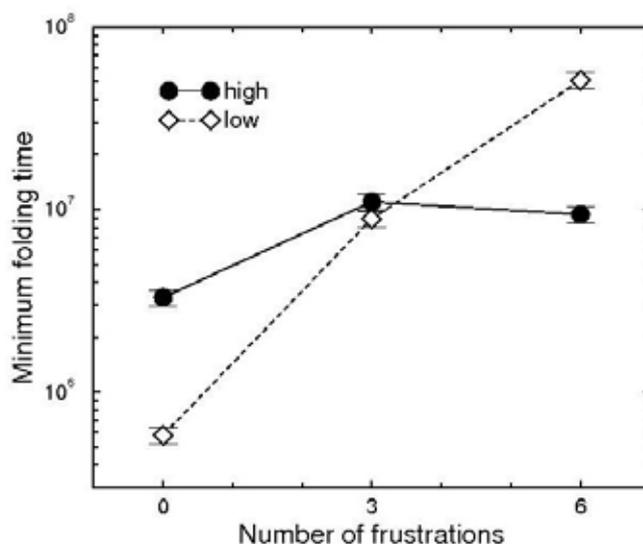


Figura 1.8: Parâmetro de enovelamento  $T_f/T_g$  como função do número de contatos frustrados  $f$  na conformação nativa e hidrofobicidade.  $f=0$ ,  $f=3$  e  $f=6$  correspondem respectivamente à seqüência nativa, 1 mutação e 2 mutações

Os resultados são resumidos na figura 1.8, onde os tempos mínimo de enovelamento  $\tau_{min}$  são mostradas como função da frustração e do regime de hidrofobicidade da seqüência. Analisando a figura 1.8, observa-se que, quando se aumenta a frustração no regime de AH, entretanto em BH observa-se que o aumento de contatos frustrados provoca um aumento considerável no tempo de enovelamento  $\tau_{min}$ . Desta forma, este resultado nos indica seqüências que se enovelam em AH

suportam mais frustração do que seqüências que se enovelam em BH.

Os resultados deste trabalho, sugerem, que seqüências otimamente bem projetadas, caracterizadas por alta estabilidade, se enovelarão mais rapidamente em baixa hidrofobicidade(caracterizado por um colapso específico) do que no regime de alta hidrofobicidade (caracterizado por um colapso não específico). No entanto para seqüências menos otimizadas (com frustração) e menor estabilidade, a situação se inverte e a proteína se enovelará mais rapidamente no regime de AH do que no regime de BH.

Deste modo, nos faz imaginar que o colapso não específico possa ser um modo de se vencer a frustração existente no estado nativo e nos faz especular sobre esta questão. Em um processo evolucionário, seqüências otimizadas bem projetadas removeriam o colapso não específico que precede o enovelamento das proteínas. Por outro lado, se o processo de enovelamento não resultar em seqüências otimizadas, esperamos que estas proteína sofram um colapso não específico, ou seja, o colapso ajudaria no enovelamento quando existir alguma frustração na cadeia.

Se pensarmos sobre alguns resultados até aqui apresentados poderíamos imaginar que para uma certa classe de proteína, que ocorram em diversas espécies e enovelam em um regime de AH, ou seja, regime em que as proteínas sofrem um rápido colapso seguido de um lento rearranjo, teria uma homologia entre as seqüências menores do que para uma classe de proteínas que se enovelam em um regime de BH, regime em que a proteína se sofre um colapso específico. Desta forma se compararmos duas classes de proteínas veremos que a diminuição da hidrofobicidade haverá um aumento do homologia da classe. Nossa proposta é pesquisar classes de proteínas que ocorrem em diversas espécies e analisar a relação entre hidrofobicidade média e homologia de algumas classes para diversas espécies diferentes.

## Capítulo 2

# Evolução de Proteínas

Nosso trabalho no doutorado foi avaliar a relação entre hidrofobicidade média de algumas classes de proteínas e correlacionar esta com a homologia média da classe. Nossa hipótese parte do fato de existir dois mecanismos básicos de enovelamento. No primeiro, a proteína se encontra em um regime de alta hidrofobicidade (AH). Neste regime a proteína rapidamente sofre um colapso não específico e depois se rearranja até atingir seu estado nativo. No outro mecanismo, a proteína se encontra em regime de baixa hidrofobicidade (BH). Em BH, a proteína vai se colapsando e enovelando diretamente na estrutura nativa da proteína. Em nossos trabalho anterior [50] , foi observado que as mutações afetam mais a estabilidade termodinâmica e a cinética de proteínas que se enovelam em um regime de baixa hidrofobicidade do que de no regime de alta hidrofobicidade sugerindo assim que proteínas que se enovelam com alta hidrofobicidade tivessem uma menor homologia (diferença entre seqüência) do que proteínas que se enovelam em um regime de baixa hidrofobicidade.

Nosso trabalho foi então pesquisar algumas classes de proteínas distintas. Tomamos o cuidado de escolhermos quatro classes de proteínas que ocorriam para as mesmas diferentes espécies. Selecionamos quatro classes de proteínas: **Lisozima**, **Citocromo c**, **Mioglobina** e **Histona H3** para 41 espécies diferentes que estão apresentadas na nos diversos bancos de dados de proteínas. A tabela 5 apresentada no apêndice A nos fornece o ID de cada sequência de proteína para cada uma das quatro classes de proteína escolhida (Citocromo c, Lisozima, Mioglobina, Histona H3) para 41 espécies diferentes. Será apresentado a partir de agora os procedimentos metodológicos utilizados neste trabalho.

## 2.1 Explorando a Evolução

Como os membros de uma família humana, os das famílias moleculares tem características em comum. Tal semelhança familiar é mais facilmente detectada comparando-se estruturas tridimensionais, o aspecto de uma molécula mais proximamente ligado à função da proteína. Comparando, por exemplo, para as ribonuclease bovina e humana, nota-se similaridade entre as estruturas tridimensional destas, apesar das seqüências de proteínas não possuírem tão alta similaridade. Comparando as ribonuclease do homem e do porco nota-se que há uma similaridade mais baixa entre as seqüências bem menor do que a similaridade destas estruturas [2].

Infelizmente, as estruturas tridimensionais determinadas até hoje correspondem a um pequeno número de proteínas diferentes. Além disso, existe o fato, observado por Chothia, que na natureza o número de folds distintas ser bem menor do que o número seqüências. Por outro lado, o desenvolvimento das técnicas de clonagem e sequenciamento de DNA, hoje possibilita obter as seqüências de proteínas através das seqüência gênica. A similaridade entre seqüências manifesta a as relações evolutivas destas seqüências. Por isso, ainda hoje é pertinente a análise das estruturas primárias.

Ao analisar as similaridades das seqüências de proteínas observa-se que as relações evolutivas também podem se manifestar nas seqüências de proteínas [2]. Por isso ainda hoje é pertinente a análise das propriedades das estruturas primárias.

### 2.1.1 Variações entre espécies homólogas

Proteínas de espécies relacionadas com proximidade filogenética tem alta identidade entre as seqüência. De acordo com a teoria evolucionária, espécies relacionadas evoluíram a partir de um ancestral comum. Da mesma maneira, segue que todas as seqüências de uma família de proteínas devem ter evoluído a partir da proteína ancestral comum.

A comparação de estruturas primárias (seqüência) relacionadas evolutivamente nos fornece quais aminoácidos são essenciais à sua função; este tipo de aminoácido é denominados de aminoácidos **invariante**. Em algumas posições podemos encontrar aminoácidos com somente propriedades semelhantes, de tal modo que nesta posição podemos encontrar aminoácidos que podem ser classificados como sendo do mesmo tipo. A substituição de um aminoácido por outro de mesma classificação de uma determinada posição é denominada de **substituição conservativa** [2]. Por outro lado, em certas posições muitos resíduos de aminoácidos podem ser tolerados em uma dada posição indicando que as exigências daquela posição não são tão específicas tal aminoácido pode ser classificado como **variável** [1].

Para ilustrar melhor estes pontos, vamos considerar a estrutura primária de uma proteína praticamente universal, o **citocromo c**. O citocromo c está presente na mitocôndria como parte da cadeia de transporte de elétrons, num sistema metabólico complexo que atua na oxidação de nutrientes para produzir ATP. Em vertebrados o citocromo c tem somente uma cadeia polimérica constituída por 103 ou 104 resíduos de aminoácidos. Em outros phyla pode se encontrar seqüência de citocromo c com até 8 aminoácidos a mais do que ocorrem nos vertebrados[1].

Emanuel Margoliash, Emil Smith, e seus colaboradores, elucidaram mais de 100 seqüências de citocromo c para várias espécies eucarióticas diferentes, variando da espécie humana a levedura de cerveja. As seqüências de 22 espécies são apresentadas na figura 2.1.

Nesta matriz triangular inferior estão apresentadas as diferenças entre as seqüências de citocromo c de 22 diferentes espécies. Nestes dados um total de 23 dos 104 aminoácidos do citocromo c são invariantes e a maioria dos demais resíduos são substituídos conservativamente. São encontradas oito posições que podem ser ocupadas por seis ou mais aminoácidos diferentes, de modo que podemos considerar estes como hiper-variáveis[1].

A figura 2.1 enfatiza que a diferença entre os grupos de espécies semelhantes é comparável a taxonomia clássica. De fato, a diferença na seqüência do citocromo C dos primatas é menor quando comparado com o citocromos c dos outros mamíferos, do que quando se compara com o citocromo dos primatas, por exemplo, com o citocromo dos insetos. Da mesma forma, os citocromos c dos fungos distinguem se tanto dos mamíferos quanto dos insetos. A partir de dados apresentados na figura 2.1 fica evidente que podemos utilizar as diferenças entre as seqüências de proteínas para se realizar uma pesquisa acerca da evolução das proteínas.

### 2.1.2 Taxa de Evolução

Proteínas diferentes evoluíram a taxas diferentes. Para justificar tal afirmação podemos mencionar, que para o ser humano e o macaco rhesus encontra se uma diferença de apenas 1% entre as sequências de seus citocromos c, uma diferença de 3 – 5% entre as suas sequências de hemoglobinas e cerca 30% de diferença entre seus fibrinopeptídeos. Um ordenamento similar dos graus relativos de variação destas proteínas é observado quando outras espécies são comparadas[4].

Da mesma forma que os tempos de divergência de várias espécies podem ser de alguma forma avaliados a partir de outros dados, as taxas com que as proteínas variam também podem ser estimadas. Uma forma de se calcular os tempos de divergências é através do cálculo da taxa



de alteração característica, conhecida como unidade de período evolutivo, que é definida como o tempo necessário para que a seqüência de aminoácidos de uma proteína sofra alteração em 1% depois de duas espécies terem divergido.

Na figura 2.2 temos no eixo y a porcentagem de mutações aceitas, ou (**unidades PAM**), que nos fornece uma medida da quantidade de mutação sofrida por uma proteína entre as diversas espécies. No eixo x temos a unidade de tempo (medida em milhões de anos) desde a divergência. Neste gráfico temos 4 classes de proteínas: **Histona H4**, **Citocromo C**, **Hemoglobina** e **Fibrinopeptídeos**. Este gráfico permite comparar proteínas como o citocromo c, que possui uma unidade de período evolutivo de 20 bilhões de anos com uma proteína que tem muito menos, a histona h4 (600 milhões de anos), com outras que sofrem variações maiores como a hemoglobina (5,8 milhões de anos) e os fibrinopeptídeos (1,1 milhões de anos)[1].

Essas informações prévias não implicam que as taxas de mutações dos DNAs que especificam estas proteínas sejam diferentes, mas sim que a taxa em que as mutações são aceitas na proteína depende de quanto as alterações nos aminoácidos afetam sua função.

O citocromo c é uma proteína pequena que, ao exercer sua função biológica, tem grande parte de sua área de superfície interagindo com grandes complexos protéicos. Assim, qualquer mutação na molécula do citocromo c poderá afetar as interações do citocromo c com o complexo, a menos é claro, que os complexos sofram mutações simultâneas de modo a acomodar a alteração, acontecimento que é bastante improvável. Por isso, que o citocromo c é uma proteína um tanto intolerante a mutação, de modo que sua cadeia polimérica varia pouca durante a evolução[1].

A histona H4 é uma proteína que se liga ao DNA nos cromossomos eucarióticos. Sua função central no empacotamento genético evidentemente a torna extremamente intolerante a quaisquer alterações mutacionais. De fato, a histona H4 está tão bem adaptada a sua função que se comparadas as seqüências das histonas H4 da vaca e da ervilha, que divergiram a 1,2 bilhões de anos, são observadas somente duas alterações conservativas nos 102 resíduos da seqüência[1].

A hemoglobina, assim como o citocromo c, é uma máquina molecular intrincada. No entanto, diferentemente do citocromo c a hemoglobina é encontrada na forma livre, tornando os grupos da sua superfície mais tolerantes a alterações do que aqueles do citocromo c. Isto explica a maior taxa de mutação encontrada para a hemoglobina frente ao citocromo c[1].

Os fibrinopeptídeos são polímeros de vinte aminoácidos, que são hidrolisados proteoliticamente nos vertebrados **fibrinogênio**, quando esta é convertida em fibrina no processo de coagulação. Assim que são cortados, os fibrinopeptídeos são descartados, de modo que existe pouca pressão seletiva para que eles mantenham suas seqüências de aminoácidos conservadas,

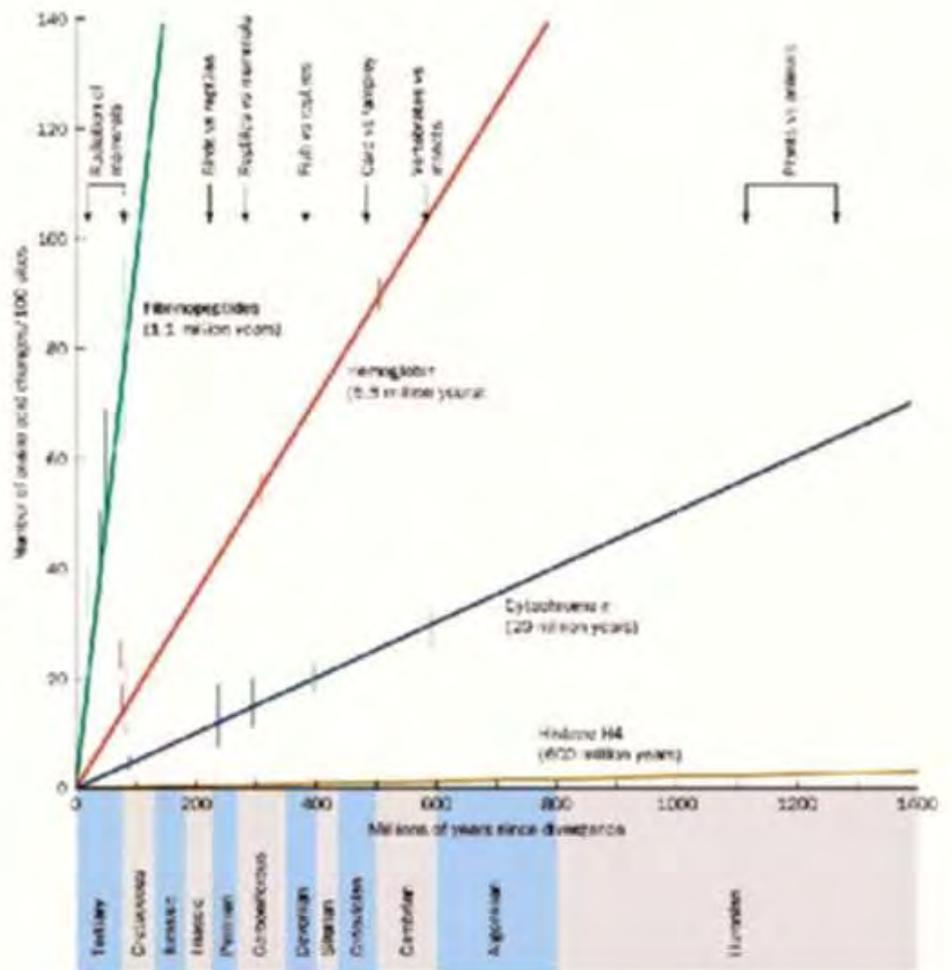


Figura 2.2: Taxas de evolução para quatro proteínas não relacionadas. O gráfico foi construído colocando-se as diferenças médias por unidade PAM, das seqüências de aminoácidos dos dois lados de um ponto de ramificação de uma árvore filogenética, contra o tempo quando, de acordo com os registros fósseis, espécies correspondentes divergiam a partir de um ancestral comum. As barras de erro indicam a distribuição experimental dos dados das seqüências. A taxa de evolução de cada proteína, que é proporcional à inclinação, está indicada em unidades de período evolutivo

assim as taxas de variação para os fibrinopeptídeos são altas. Se presumirmos que os fibrinopeptídeos estão evoluindo aleatoriamente, então as unidades de período evolutivo, indicam que apenas  $1,1/5,8 = 1/5$  das alterações de aminoácidos são aceitáveis, isto é, são inócuas, enquanto que essa quantidade é  $1/18$  para o citocromo c e  $1/550$  para a histona H4[1].

### 2.1.3 Regra de Seleção

Substituições de aminoácidos em uma proteína resultam, na maior parte das vezes, de alterações em uma única base no gene que codifica a proteína. Se tais mutações pontuais ocorrem sobretudo em conseqüências de erros no processo de replicação do DNA, então a taxa na qual uma dada proteína acumula mutações seria constante em relação ao número de gerações celulares. Se os processos de mutações resultarem da degradação química aleatória do DNA, então a taxa de mutação seria constante em relação ao tempo absoluto. Para escolher entre essas hipóteses, comparemos a taxa de divergência do citocromo c em insetos e mamíferos[4]. ,

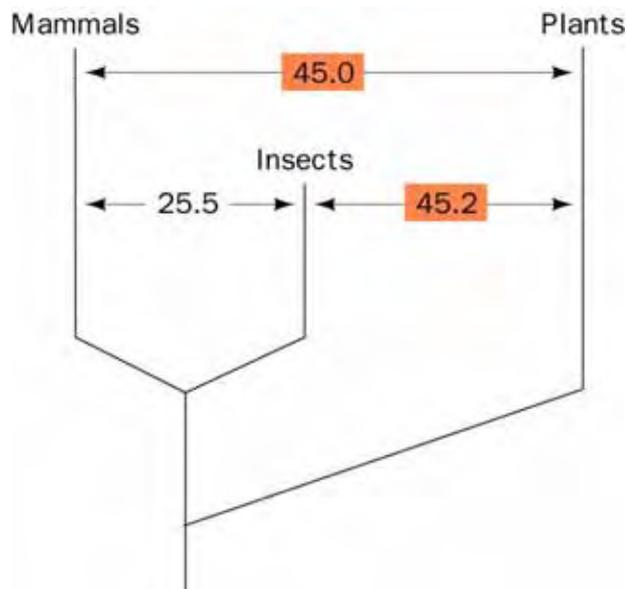


Figura 2.3: **Árvore filogenética para o citocromo c:** A árvore mostra o número médio de diferenças de aminoácidos entre citocromos c de mamíferos, insetos e plantas. Desde o ponto de ramificação, mamíferos e insetos divergiram igualmente das plantas(Adaptada de Dickerson, R.E. e Timkovich R., in BOYER, P.D.[Ed.], *The Enzymes*[3rd ed], Vol11, p.447, Academic Press[1975]

Insetos apresentam tempos de gerações mais curto do que os mamíferos. Conseqüentemente, se a replicação do DNA for a principal fonte de erros mutacionais, então, desde a

época em que as linhagens de insetos e mamíferos divergiram, poderíamos concluir, que devido ao maior número de gerações celulares, que os insetos teriam evoluído a partir de plantas mais do que os mamíferos[4]. As distâncias médias entre as espécies ou grupos de espécies são definidas como o número médio das diferenças entre as sequências do citocromo c. Analisando o dendograma 2.3 observa-se que a distância média entre insetos e plantas (45,2) é essencialmente a mesma distância entre mamíferos e plantas (45,0). Assim é possível concluir que o citocromo c acumula mutações em uma taxa uniforme em relação ao tempo, e não em relação ao número de gerações celulares. Isso, por sua vez implica que mutações pontuais no DNA acumulam a uma taxa constante com o tempo, ou seja, por meio de alteração química aleatória, ao invés de resultarem principalmente de erros no processo de replicação [4].

#### 2.1.4 Evolução Neutra

Se as mutações ocorrem a uma taxa constante em todos os genes, como podemos explicar o fato de encontrarmos um largo intervalo de taxas evolucionárias para as proteínas e porque cada "classe" de proteína tem uma taxa evolucionária constante. A explicação mais plausível é que a maioria das mutações que ocorrem entre os aminoácidos são devido a mutações neutras, que não afetam significativamente a função da proteína [1].

De acordo com a hipótese de mutação neutra, a taxa constante de divergência é a mesma que a taxa de mutação neutra por gene, que é dada pelo produto da taxa total de mutações pela fração de mutações que efetivamente neutra. Daí, se a taxa de mutação é a mesma para todos os genes, a taxa de mutação neutra seria diferente para cada gene por causa das diferentes frações de mutações que ocorrem em diferentes proteínas que são efetivamente neutras. Todo gene ou proteína diferiria de outras proteínas, pela quantidade de aminoácidos que podem variar sem afetar sua função. Se o aminoácido exato não é crucial para a função da proteína, uma grande fração de suas mutações seria neutra e a sequência da proteína evoluiria rapidamente[4].

Geralmente o grau de variação na estrutura primária de uma proteína está inversamente correlacionada com a importância biológica de cada resíduo. Os aminoácidos que mais variam são aqueles que se encontram na superfície de uma proteína e que não envolvem interações funcionais com outras moléculas. Os aminoácidos mais conservados são aqueles que estão diretamente envolvidos na função biológica, por exemplo, os aminoácidos do citocromo c que interagem com o grupo heme e os sítios ativos das enzimas[4].

Todas as observações realizadas aqui indicam que os tipos de mudanças em nível molecular tem no mínimo conseqüências funcionais. Deste modo, a ocorrência de mutações

que não afetam a funcionalidade da proteína pode ser explicada como sendo resultado da acumulação de mutações neutras. A seleção natural parece possuir uma correlação negativa com as mutações, deletando mutações que afetam a função da proteína[4].

É claro que mudanças funcionais devem ocorrer durante a evolução, como é evidenciado pela diversidade de organismos. Esta diversidade não é frequentemente evidente em nível molecular. Proteínas com a mesma função que ocorrem em diferentes espécies usualmente tem propriedades similares[1]

Examinando as "pegadas" presentes nas seqüências de proteínas modernas, o pesquisador em biologia molecular pode se tornar um arqueólogo molecular capaz de aprender sobre os eventos no passado evolutivo. As comparações de seqüências podem em geral revelar tanto as vias de descendência evolutiva quanto às estimativas de marcos evolutivos específicos. Além disso, o estudo das comparações de proteínas pode revelar mecanismos e propriedades de famílias de proteínas. As comparações podem ser realizadas na WEB utilizando se de dados de seqüência de proteínas que estão disponíveis em bancos de dados disponibilizados gratuitamente, ou não, na WEB[2].

O primeiro passo na escolha e análise de um conjunto de seqüência é, realizar uma pesquisa das seqüências de interesse para a pesquisa, que estão disponíveis nos bancos de dados de seqüência biológicas na WEB.

## **2.2 Bancos de Dados**

### **2.2.1 O que é Bancos de Dados e Base de Dados**

Um banco de dados efetivamente é uma planilha eletrônica, que possui um método eficiente de guardar uma vasta quantidade de informação. O armazenamento dos dados dependem de dois fatores: da natureza da informação a ser armazenada(seqüência,estruturas, imagens 2D e 3D) e do gerenciamento dos dados armazenados[59].

Uma característica marcante da moderna pesquisa em genômica é da geração de uma enorme quantidade de dados gerados. De forma que o volume crescente dos dados de sequências exigiu que os dados sobre as sequências de DNA fossem armazenados e disponibilizados para a comunidade científica. Assim pode se afirmar que a pesquisa em genômica exigiu o desenvolvimento de metodologias computacionais sofisticadas para armazenar e disponibilizar as informações a respeito da sequências de DNA e proteínas. A forma mais adequada para realizar

esta tarefa foi a criação de bancos de dados eletrônicos que disponibilizam as informações das sequências na WEB.

Um banco de dados disponibiliza informações sobre sua base de dados. A base de dados é uma coleção de dados logicamente relacionados com algum significado, de modo que associações livres de dados não constituem uma base de dados. Um sistema de banco de dados é basicamente um sistema de manutenção de registro para um propósito específico que atenda um determinado grupo de usuários. Os bancos de dados de sequências biológicas é destinado a pesquisadores da área de **Biologia Molecular**.

Bancos de dados são compostos pelo hardware do computador e pelo software de gerenciamento dos dados fornecidos pelo banco de dados. O objetivo central do desenvolvimento de uma base de dados é organizar os dados através de registros estruturados que permitam a recuperação fácil da informação. Cada registro ou anotação deve conter um **ID** um número de campos que assegurem atualizados os dados dos ítem. Por exemplo, campos para nomes, número de telefones, endereço e datas. Para recuperar um dado gravado em uma base de dados, o usuário pode especificar uma parte da informação, chamada de **valor** será pesquisado em um campo em particular e espera se que a informação completa seja restaurada. Este processo é chamado de fazer pergunta [58].

Ainda que a principal proposta de todos os bancos de dados seja de recuperar dados, os bancos de dados biológicas tem um alto nível de solicitação, conhecida como *conhecimento descoberto*, que se refere a identificação das conexões entre as partes que não se conhecia quando a sequência foi primeiro anotada. Por exemplo, base de dados contendo informação da sequência pode realizar uma tarefa computacional extra e identificar homologia de sequência e motifs conservados [58].

## 2.3 Tipos Bancos de Dados

Originalmente todas os bancos de dados usavam um formato de arquivo flat. Um arquivo flat é tão somente um longo arquivo texto que contém muitas entradas separadas por delimitadores, um caracter especial como a barra vertical (|). No interior de cada de entrada encontram se um número de campos separados por tab ou dois pontos. Exceto para valores nas colunas em cada campo, o arquivo texto completo não contém nenhuma instrução oculta para computadores realizarem uma pesquisa de informação específica. O arquivo texto pode ser considerado uma tabela. De modo que, para um computador pesquisar uma parte da informação, o computador precisa ler todo o arquivo texto para encontrar a informação desejada. Isto é manejável, ou melhor, tratável para pequenos bancos de dados, mas quando o

banco de dados cresce ou os tipos de dados tornam se mais complexos este estilo de banco de dados torna difícil a recuperação da informação. De fato, pesquisas em arquivos textos podem provocar o travamento do computador por ser uma operação que consome muito a memória do computador [58].

Para facilitar o acesso e a recuperação de dados programas de computador para organização, pesquisa e acesso aos dados tem sido desenvolvidos. Eles são denominados de **Sistemas Gerenciadores de Bancos de Dados SGBD**. Tais sistemas não possuem somente colunas de dados gravados, mas instruções operacionais para ajudar a identificar alguma conexão entre os dados gravados. A proposta de estabelecer uma estrutura de dados e facilitar a execução de pesquisas e combinar diferentes dados para formar um relatório de pesquisa final. Dependendo dos tipos de estruturas de dados, o sistema de SGBD podem ser classificados em dois tipos: **Sistema de Gerenciamento de Banco de Dados Relacional** e **Sistema de Gerenciamento de Banco de Dados Orientado a Objeto**. Conseqüentemente os bancos de dados que aplicam estas metodologias são chamados de **Banco de Dados Relacional (BDR)** e **Banco de Dados Orientado a Objeto (BDOO)** [58].

### 2.3.1 Banco de Dados Relational

O banco de dados relacional utiliza um conjunto de tabelas para organizar os dados. Cada tabela, chamada de **relação** é composta de colunas e linhas. Colunas representam campos individuais. Linhas representam os valores nos campos dos dados gravados. A coluna em uma tabela é indexada de acordo com uma característica comum denominada de **atributo**, tal que esses atributos possam ter referências cruzadas em outras tabelas. Para se realizar uma pesquisa em um BDR, o sistema seleciona itens linkados de diferentes tabelas e combina a informação em um relatório. De forma que, uma informação específica pode ser encontrada mais rapidamente em banco de dados relacional do que banco de dados de arquivo flat [58].

Banco de dados relacional podem ser criados por uma linguagem de programação específica chamada **Structured Query Language (SQL)**. A criação deste tipo de banco de dados pode ser um grande desafio na fase de projeto. Depois da criação do banco de dados original, uma nova categoria de dados pode ser adicionada sem a exigência de que todas as tabelas seja modificadas [58].

### 2.3.2 Banco de Dados Orientado a Objeto

Um dos problemas dos bancos de dados relacionais é que as tabelas não descrevem relações hierárquicas complexas entre os itens dos dados. Para contornar esse problema, bancos

de dados orientados a objetos foram desenvolvidos para guardar o dado como um objeto. Em uma linguagem de programação orientada a objeto, um objeto pode ser considerado como uma unidade que combina dados e rotinas matemáticas que atua sobre os dados. O banco de dados é estruturado tal como um objeto conectado por um conjunto de ponteiros definindo relações pré determinadas entre os objetos. A pesquisa no banco de dados orientado a objeto envolve navegar entre objetos com ajuda ponteiros conectando diferentes objetos. A linguagem *C++* é frequentemente utilizada para banco de dados orientados a objeto [58].

O sistema de banco de dados orientado a objeto é mais flexível. Os dados podem ser estruturados baseados em relações de hierarquia. Fazendo isto, as tarefas computacionais a dados que tem um relação complexa, tal como um dado multimídia. Entretanto este tipo de sistema de banco de dados perde em fundamentação matemática rigorosa. Existe também o risco que algumas relações entre objetos possam ser de difícil representação. Alguns bancos de dados incorporam características dos dois tipos de programação de banco de dados, criando um **Sistema de Gerenciamento de Banco de Dados Objeto-Relacional (SGBDOR)**

### 2.3.3 Banco de Dados de Sequências Biológicas

Os bancos de dados de proteínas atualmente utilizam os três tipos de estruturas de banco de dados: arquivos flat, relacionais e orientado a objeto. Apesar da fraqueza do uso de arquivos flat no gerenciamento do banco de dados, muitos bancos de dados de sequências biológicas ainda utilizam este formato.

Os bancos de dados eletrônicos podem ser construídos por arquivos flat, relacionais ou orientados a objetos. Arquivos flat são simplesmente arquivos texto. Este tipo de arquivo dificulta qualquer forma de organização que facilite a pesquisa documental dos computadores. Bancos de dados relacionais organizam os dados através de tabelas de forma que a pesquisa da informação compartilha características. Bancos de dados orientados a objeto organizam os dados como objetos associados a objetos de acordo com relações de hierarquia. Os bancos de dados de sequência biológica são compostos por esses três tipos. Baseado neste contexto podemos dividir os bancos de dados de sequência biológicas em três categorias[59]:

- Bancos de Dados Primários
- Bancos de Dados Secundários
- Bancos de Dados Especializados

### 2.3.4 Bancos de Dados Primários

Os **bancos de dados primários** contém os dados biológicos originais. Eles fornecem e disponibilizam seqüências ou a estrutura tridimensional de proteínas ou nucleotídeos através de um arquivo de colunas. Existem três grandes bancos de dados públicos que armazenam dados de seqüência de ácidos nucléicos produzidos e submetidos por pesquisadores na **WEB:GenBank**, **European Molecular Biology Laboratory (EMBL)** e **DNA Data Bank of Japan (DDBJ)**. Esses bancos de dados podem ser acessados gratuitamente na Internet. A maioria dos dados destes bancos são contribuições diretas dos autores com um mínimo de nível de anotação. Um pequeno número de seqüências, especialmente as publicadas nos anos 80, foram anotadas manualmente através de artigos publicados em jornais e revistas científicas pelo staff que gerencia o banco de dados[58].

Atualmente, a submissão de seqüência nos bancos GenBank, EMBL, ou DDBJ é condição necessária para a maioria das publicações científicas para assegurar que os dados moleculares possam ser utilizados livremente. Estes três bancos de dados tem uma colaboração bem próxima e dados novos são acrescentados diariamente. Eles se uniram e formaram o **International Nucleotide Sequence Database Collaboration (INSDB)**. Isto significa, que se conectando a um destes bancos de dados tem se acesso aos dados de seqüência dos três bancos de dados. Entretanto cada um destes 3 bancos de dados tem maneiras ligeiramente diferentes de apresentar os dados[58].

Felizmente, para a estrutura tridimensional de macromoléculas biológicas, existe um banco de dados centralizado, o **Protein Data Bank (PDB)**. Este banco de dados disponibiliza as coordenadas atômicas(de proteínas e ácidos nucléicos) das macromoléculas determinadas por difração de raios X ou por ressonância magnética nuclear (NMR). O formato do arquivo do PDB apresenta o nome da proteína, os autores, detalhes experimentais, estruturas secundárias, cofatores e as coordenadas atômicas. A interface WEB tem ferramentas de visualização das estruturas apresentadas[58].

### 2.3.5 Bancos de Dados Secundários

A anotação da informação em um banco de dados primário freqüentemente é mínima. Para melhorar o nível de informações sobre a seqüência existem os bancos de dados secundários. Este tipo de banco de dados de seqüência biológica realiza um pré-processamento a respeito da seqüência derivada dos bancos de dados primários. A quantidade de processamento realizado varia muito entre os bancos de dados secundários. Alguns fornecem simplesmente a tradução da seqüência para o DNA, no entanto outros bancos de dados secundários fornecem anotação e

informação mais complexas, como a estrutura tridimensional e função biológica[58].

Um exemplo de banco de dados secundário é o **SWISS-PROT**. Este banco de dados fornece anotações detalhadas da seqüência que incluem informações sobre estrutura, função e família de proteína. Os dados de seqüência são derivados principalmente de **TrEMBL**, um banco de dados de seqüência traduzidas de seqüência de ácidos nucléicos armazenadas no banco de dados EMBL[58].

A anotação de cada entrada do SWISS-PROT é cuidadosamente apurada por especialistas para se ter uma boa qualidade de informação. A anotação para cada seqüência inclui função, domínio estrutural, sítios catalíticos, ligantes, informações sobre vias metabólicas, doenças associadas e similaridade com outras seqüências. A anotação adiciona informações pertinentes à seqüência gravada. Muitas informações são obtidas da literatura científica e adicionadas pelos curadores do banco de dados. Além disso, o SWISS-PROT disponibiliza o recurso de links de referências cruzadas de interesse. Outra característica é a baixa redundância e um alto nível de integração com outros bancos de dados primários e secundários, tornam o SWISS-PROT um banco de dado popular na comunidade da biologia molecular[58].

Em 1992, esforços combinados dos bancos de dados SWISS-PRPOT, TrEMBL, e PIR terminaram na criação do banco de dados **UniProt**. O banco de dados UniProT tem uma maior cobertura do que somente um dos três bancos de dados sozinhos. O UniProt mantém a característica original do SWISS-PROT de baixa redundância, referências cruzadas e alta qualidade de anotação[58].

Existem também bancos de dados secundários que relacionam classificação de famílias de proteínas de acordo com suas funções e estruturas. Os bancos Pfam e Blocks contém informações sobre seqüência alinhadas como os motifs e padrões destas seqüências, que podem ser utilizados para classificação das famílias e a inferência das funções da proteína. O banco de dados **DALI** é um banco de dados de estrutura secundária de proteínas vital na análise da classificação da estrutura da proteína por identificar as relações evolucionárias entre proteínas[58].

### 2.3.6 Bancos de Dados Especializados

Bancos de Dados Especializados normalmente servem a uma comunidade de pesquisadores ou a um organismo em particular. O conteúdo destes bancos de dados podem ser seqüências ou outro tipo de informação. As seqüências nestes bancos de dados podem possuir sobreposição de seqüências em sua base de dados, entretanto eles tem seus dados atualizados diretamente por seus autores. Por serem mantidos por especialistas de uma certa área, estes banco de dados tem uma anotação que pode ser diferenciada, além de possuírem anotações específicas para cada

sequência. Muitos bancos de dados de genoma, que possuem uma taxonomia específica falha sem esta categoria de banco. Exemplos incluem os bancos **FlyBase**, **WormBase**, **ACEDB** e **TAIR**. Além disso, existem bancos de dados que contém dados originais derivados de análise funcional. Por exemplo, os bancos de dados **GenBank EST Microarray Gene Expression** e o **European Bioinformatics Institute** são alguns dos bancos de dados de expressão genética.

### 2.3.7 Falhas dos Bancos de Dados

Um dos problemas associados aos bancos de dados biológicos é o excesso de confiança na informação da seqüência e nas anotações sem se preocupar com a qualidade da informação. Frequentemente se ignora o fato de que pode existir muitos erros nos bancos de dados de seqüência. Existem bancos de seqüências em que as seqüências de proteínas são derivadas das seqüências de nucleotídeos de genes. A maioria dos erros cometidos na sequenciamento de nucleotídeos é causada por erros de técnicas experimentais. Pois algumas vezes as técnicas experimentais podem provocar um deslocamento de quadro, tornando a identificação difícil a identificação completa do gene e sua tradução para os aminoácidos bem mais difícil ainda. Outras vezes a seqüência de gene pode estar contaminados pelo vetor clone. Estes tipos de erros são mais comuns nos dados obtidos antes de 1990. A partir de então, a das técnicas experimentais aumentou consideravelmente[58].

A Redundância é outro grande problema dos bancos de dados de seqüência. Por diversos motivos, existe uma tremenda duplicação de informações nos bancos de dados. As causas da redundância, pode ser várias e incluem:

- submissão de seqüência idêntica pelos mesmos ou diferentes autores.
- revisão da anotação
- mau gerenciamento dos curadores do banco de dados

Passos tem sido tomados para se reduzir a redundância no banco de dados National Center for Biotechnology Information (**NCBI**). Os seus mantenedores criaram um banco de dados de seqüência não redundantes chamado de **RefSeq**, no qual seqüência de um mesmo organismo e fragmentos de seqüência são unidos em uma mesma entrada. Seqüências de proteínas derivadas das mesmas seqüências de DNA são explicitamente lincadas com entradas relacionadas. Seqüências variantes de um mesmo organismo com muito poucas diferenças, que podem aparecer por causa de erros de sequenciamento, são tratadas distintamente[58].

Como mencionado anteriormente, o SWISS-PROT também tem a preocupação de manter uma baixa redundância para as seqüências se comparado com outros bancos de dados. Outra maneira de tentar resolver o problema da redundância é criar aglomerados de seqüência, tal como o banco de dados UniGene realizou. Outro problema é o de anotações errôneas. Frequentemente o mesmo gene é encontrado como sendo anotado com diferentes nomes, resultando deste modo em múltiplas entradas de um mesmo gene, causando assim confusão sobre os dados[58].

Por outro lado genes não relacionados entre si são encontrados possuindo o mesmo nome. Para aliviar o problema de nomenclatura dos genes a re-anotação de genes e proteínas usando um conjunto comum de um vocabulário controlado para descrever gene e proteína é necessário. É necessário criar uma nomenclatura não ambígua para genes e proteínas[58].

Algumas das inconsistências na anotação são causadas simplesmente pelo genuíno desacordo entre pesquisadores. Outros erros podem surgir da imprudência dos pesquisadores em definir as funções das proteínas das seqüências submetidas. Existem também alguns erros que são simplesmente causados por omissão ou erros de digitação. Erros na anotação podem ser particularmente danosos, visto que a grande maioria das novas seqüências tem suas funções definidas bem como suas famílias baseadas na similaridade com seqüência que já foram anotadas nos bancos de dados. Um erro na anotação pode ser facilmente transferido a todos os genes similares em todo o banco de dados. Alguns erros podem ser corrigidos estudando as funções e famílias de proteínas. Outros são corrigidos através de trabalhos experimentais.

## 2.4 Alinhamento

### 2.4.1 Bases Evolucionárias

Há uma longa tradição em biologia de análises comparativas. Somente de forma ilustrativa podemos dizer que a comparação de características morfológicas dos passarinhos das ilhas Galapagos com outras espécies é que foi possível a Darwin desenvolver a teoria da seleção natural. Em essência estamos realizando o mesmo tipo de análise quando estamos comparando seqüências de genes e proteínas. Nesta atividade as similaridades e diferenças são analisadas para se poder depois realizar inferências relacionadas à estrutura, função e relações evolucionárias destas seqüências. O método comparativo mais comum é o de seqüências biológicas é **alinhamento de seqüência**, que fornece um mapeamento explícito entre os resíduos de duas ou mais seqüência[60].

O DNA e as proteínas foram e são produtos da evolução. A construção dos blocos

destas macromoléculas biológicas, as bases dos nucleotídeos e os aminoácidos das proteínas, formam uma seqüência linear que determina a estrutura primária destas moléculas. Estas moléculas podem ser consideradas fósseis que codificam a história de milhões de anos de evolução[58].

Durante a evolução, as seqüências moleculares sofreram mudanças aleatórias, algumas das quais foram selecionadas durante o processo de evolução. Durante a evolução as seqüências gradualmente acumularam mutações e se divergiram no tempo. Traços deste processo podem ainda ser observados em certas porções da seqüência[58].

### 2.4.2 Homologia vs. Similaridade

Um importante conceito na análise da seqüência é o conceito de **homologia**. Quando duas seqüências possuem uma mesma origem evolucionária detectável elas são ditas ser homólogas. Um termo relacionado à homologia, mas que tem um significado diferente é a **similaridade**. A similaridade se refere a porcentagem de resíduos que tem propriedades físico-químicas similares tais como tamanho, carga e hidrofobicidade. Por exemplo, se utilizando o critério de classificação hidrofóbica utilizada por Leningher, apresentada anteriormente neste trabalho, for encontrado que um aminoácido isoleucina de uma seqüência 1 coincide com o aminoácido leucina de outra seqüência 2 pode se afirmar que estes aminoácidos são similares, pois pertencem a mesma classificação hidrofóbica[59].

O alinhamento de seqüência possibilita a inferência através da relação entre duas ou mais seqüências. Se duas seqüências compartilham alguma similaridade significativa, isto indica que estas seqüências derivaram de uma mesma origem evolucionária comum. Quando um alinhamento de seqüência é gerado corretamente, ele reflete as relações evolucionárias de duas seqüências: regiões que estão alinhadas mas não idênticas correspondem a substituições de resíduos. É possível encontrar regiões onde resíduos de uma seqüência não concordem a nenhuma deleção ou inserção em relação à outra seqüência indicando assim que houve uma mudança em uma das seqüências durante a evolução. Mesmo assim pode ser possível que estas duas seqüências sejam derivadas de um ancestral comum, mas em algum ponto da evolução divergiram de tal modo que, as relações ancestrais comuns se tornam mais difíceis de serem observadas.

Um ponto importante é distinguir homologia de similaridade, porque os dois termos são freqüentemente confundidos na literatura científica. Para se esclarecer, homologia de seqüência é uma inferência ou conclusão a cerca de uma relação ancestral comum feita quando duas seqüências tem um alto grau de similaridade. Por outro lado, a similaridade pode ser

quantificada usando porcentagens. Por exemplo, pode se dizer que duas seqüências tem uma similaridade de 40%, mas não podemos dizer que estas seqüências compartilham 40% de homologia. Elas são homólogas ou não[60]

### 2.4.3 Similaridade vs. Identidade

Outro conjunto de dados relacionados que merecem ser distinguidos são a **similaridade** e a **identidade**. A similaridade e a identidade são sinônimos para as seqüências de nucleotídeos. Para proteínas estes dois conceitos são diferentes. No alinhamento de seqüência de proteínas, a identidade da seqüência se refere à porcentagem de aminoácidos de uma seqüência em relação à outra. A similaridade se refere à porcentagem de aminoácidos alinhados que tem propriedades físico-químicas similares. Dois aminoácidos diferentes que possuem as mesmas propriedades são ditos similares.

Existem dois modos de se calcular a similaridade/identidade de seqüência. Um modo envolve o uso do comprimento total de ambas as seqüências. O outro normaliza o tamanho da menor seqüência. O primeiro método utiliza a seguinte fórmula dada por 2.1, onde  $S$  é dado em porcentagem da similaridade,  $L_s$  é a quantidade de aminoácidos alinhados com características similares  $L_a$  e  $L_b$  são os comprimentos individuais de cada seqüência.

$$S(\%) = \frac{L_s \times 2}{L_a + L_b} \times 100 \quad (2.1)$$

A identidade  $I(\%)$  pode ser calculada por 2.2, onde  $L_i$  é o número de resíduos alinhados idênticos.

$$I(\%) = \frac{L_i \times 2}{L_a + L_b} \times 100 \quad (2.2)$$

O segundo modo de se calcular a identidade de seqüência é apresentado por 2.3,

$$I(\%) = \frac{L_s}{L_a} \times 100 \quad (2.3)$$

que é dada pela razão entre o número de aminoácidos com características similares  $L_s$  dividido pelo maior comprimento das duas seqüências alinhadas

### 2.4.4 Alinhamento Global vs. Alinhamento Local

Os métodos utilizados para se calcular a similaridade das seqüências podem ser agrupados em duas categorias: **alinhamento global** e **alinhamento local**. O alinhamento global toma duas seqüências e as compara ao longo de todo seu comprimento e tenta encontrar o melhor alinhamento entre as duas seqüências. O método de alinhamento global, em geral é aplicável a proteínas com alta similaridade, ou seja, para proteínas que possuem o mesmo

comprimento. Para seqüências divergentes ou de tamanhos diferentes este método não pode ser bem aplicado porque falha no reconhecimento de regiões altamente similares entre duas seqüência[60].

A maioria dos pesquisadores em biologia molecular depende da segunda categoria de algoritmo de alinhamento, que utiliza seqüência locais para em seus alinhamentos. Os métodos de alinhamentos locais, de modo geral comparam os trechos das seqüências, sem se preocupar com o alinhamento do resto da seqüência. Frequentemente, tais métodos fornecem mais que um resultado de alinhamento para duas seqüências pesquisadas porque pode ocorrer mais que um domínio para as seqüências analisadas[60]. Estes métodos são mais úteis para avaliar padrões conservados de seqüência mais divergentes em proteínas e DNA[58].

#### 2.4.5 Matriz Dot Plot

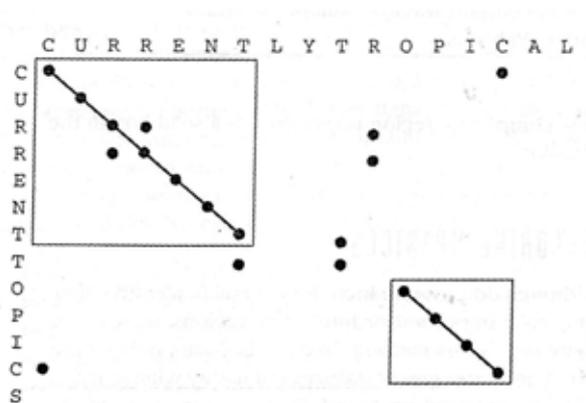


Figura 2.4: Construção do dotplot das frases: Current Topics e Currently Tropical. Indicando a região de alinhamento

Um dos modos fundamentais na comparação entre duas seqüências é pelo método visual chamado **dot-plot**. Os Dotplots fornecem uma rápida identificação de regiões de alinhamentos locais, repetidas direta ou inversamente, inserções, deleções ou regiões de baixa complexidade. Este método gráfico compara duas seqüências através de uma matriz. Para

demonstrar como o dot-plot é construído vamos ver como fica o alinhamento da frase1 *current topics* com a frase2 *currently tropical*[60].

Para começar as frases são escritas de cima para baixo de da esquerda para a direita ao lado da matriz como está mostrado na figura 2.4. Agora posição a posição marque o ponto em que existe coincidência entre as letras das duas frases que. Na coluna 1 existe um C da frase *current topics* que coincide com o C da frase *currently tropical*. Existe outro C na coluna 12 da frase *current topics*, de tal modo que este ponto também deve ser marcado. Usando a mesma metodologia move se através marcando todos os pontos em que existam coincidências entre as letras das duas frases que, no caso, estão sendo alinhadas. Conectando os pontos por setas se encontra o alinhamento ótimo das seqüências. Quanto mais próximo for o caminho ótimo da diagonal principal maior será similaridade entre as seqüências. De forma que, quando o movimento entre dois pontos sucessivos ocorrer na diagonal dois pares de letras estão alinhadas. Se o movimento é na direção horizontal um gap é inserido indexando as linhas da seqüência. Se o movimento é na direção vertical da mesma forma um gap também é inserido indexando as colunas da seqüência. Se o movimento acontecer para a esquerda ou para cima nota se que várias (aminoácidos ou nucleotídeos) de uma seqüência tem somente um caracter em comum. O caminho indicado pelas setas entre os pontos em comum nos fornece o alinhamento ótimo. Outra forma de pensar o caminho através do dotplot é como um script, isto é, a prescrição de uma série de operações para transformar a seqüência que está disposta horizontalmente na seqüência que está disposta verticalmente.[60].

#### 2.4.6 Programação Dinâmica

Programação dinâmica é um método que determina o alinhamento ótimo entre duas seqüências para os pares possíveis de caracteres entre duas seqüências. Isto é fundamentalmente uma forma de matematizar o conceito de caminho ótimo da matriz dot-plot. Este método pesquisa o conjunto de caminhos de maior pontuação para depois encontrar o alinhamento ótimo das seqüência[58].

A programação dinâmica primeiro constrói uma matriz bidimensional. Nesta matriz se dispões da esquerda para a direita uma das seqüências (sequência 1) a serem alinhadas. A outra seqüência fica disposta de cima para baixo. Desta forma cada linha se referirá a um caracter da seqüência 1 e cada coluna se referirá a cada caracter da seqüência 2. A comparação entre os caracteres é feita através de uma pontuação dada por uma matriz de substituição. A pontuação é feita uma linha de cada vez[58].

Inicia-se a comparação com o primeiro caracter da linha comparando com os todos

os caracteres das colunas da outra seqüência. A contagem da segunda coluna se inicia a partir da classificação das pontuações da primeira linha. A melhor pontuação é colocada no canto direito inferior de uma matriz intermediária. O processo realiza iterações até que todas as células sejam preenchidas. Deste modo a pontuação é acumulada ao longo da diagonal indo da parte superior esquerda à parte inferior direita[60]. Depois de calcular as pontuações de toda matriz, deve se encontrar o caminho que representa o alinhamento ótimo[60]. Isso é feito traçando de maneira reversa, ou seja, do canto direito inferior em direção ao canto esquerdo superior. O melhor caminho é aquele que obtiver o maior valor de pontuação total[58]. Se dois ou mais caminho tem o maior valor de pontuação total, então é escolhido um destes de forma arbitrária. O caminho pode se mover na diagonal, horizontal ou verticalmente. O movimento na horizontal ou vertical corresponde à introdução de um gap, inserção ou deleção de uma das duas seqüências[59].

#### 2.4.7 Matrizes de Substituição(Pontuação)

O algoritmo programação dinâmica apresentado na seção anterior, que possui um conjunto de valores para quantificar a possibilidade de um aminoácido ser substituído por outro no alinhamento. O sistema de pontuação é chamado de **matriz de substituição**. Estas matrizes são derivadas de análises estatísticas de substituições de aminoácidos de conjuntos de alinhamentos realizados entre seqüência altamente relacionadas[59].

As Matrizes de substituição para seqüência de nucleotídeos são relativamente simples. Um valor positivo ou alto é dado para uma pontuação quando se encontra coincidência entre caracteres e valores negativos ou de baixa pontuação para caracteres diferentes. Esta escolha é baseada na hipótese de que as freqüências de mutações são iguais para todas as bases. No entanto estas hipóteses podem não ser realísticas. Observações mostram substituições entre purinas e purinas ou entre pirimidinas e pirimidinas ocorrem com mais freqüência[58].

A determinação de matrizes de substituição para os aminoácidos são mais complicadas, porque as pontuações da matriz entre os diversos tipos de aminoácidos, devem refletir as propriedades físico-químicas tão bem quanto à probabilidade de um aminoácido ser substituído por outro. Certos aminoácidos, com propriedades físico-químicas similares, podem ser mais facilmente substituído do que de aminoácidos de características diferentes, pois substituições entre aminoácidos similares (mesma classe) preservam as características essenciais de função e estrutura. Entretanto, substituições entre aminoácidos de propriedades físico-químicas diferentes podem provocar a quebra do binômio função/estrutura. Este tipo de troca é menos provável de ocorrer porque tornar as proteínas não funcionais[60]. Por exemplo a fenilalanina, a tirosina

e o triptofano compartilham uma estrutura de anel aromático. Por causa de suas propriedades qualquer um dos três aminoácidos pode ser substituído pelos outros sem perturbar a função regular da proteína. Analogamente, arginina, lisina e histidina são aminoácidos básicos existindo uma alta probabilidade de um destes ser substituído pelos outros. O glutamato, o aspartato, a asparagina e a glutamina estão associados com alta frequências de substituições[58].

O grupo de aminoácidos hidrofóbicos (apolares) incluem metionina, isoleucina, leucina e valina. Aminoácidos pequenos e polares incluem serina, treonina e cisteína. Aminoácidos com estes grupos possuem uma alta probabilidade de um deles ser substituído pelos outros aminoácidos do mesmo tipo. Contudo, a cisteína contém um átomo de enxofre, que permite a formação da ligação dissulfeto. A substituição da cisteína com outros resíduos freqüentemente desestabiliza a estrutura da proteína, de modo à substituição deste aminoácido é muito pouco freqüente. Os aminoácidos pequenos e apolares tais como a glicina e a prolina também são únicos na quebra de estruturas secundárias regulares. Para esses aminoácidos as substituições são muito pouco freqüentes[58].

As matrizes de substituições de aminoácido são matrizes de  $20 \times 20$  que refletem a probabilidade de substituição dos aminoácidos. Essencialmente existem dois tipos de matrizes de substituições para aminoácidos. Um tipo é baseado na inrtecambealidade do código genético ou propriedades de aminoácidos. O segundo tipo é derivado de estudos empíricos de substituições de aminoácidos. O primeiro tipo de matriz de substituição tem se tornado menos utilizado do que o segundo tipo de matriz. De forma a aproximação empírica tem se tornado mais populares e por isso será o nosso próximo foco de discussão[58].

As matrizes empíricas, que incluem as matrizes **PAM** e **BLOSUM** são derivadas de alinhamentos de seqüência com alta similaridade. Através da análise das probabilidades de substituições é possível criar um sistema de pontuação desenvolvido dando altos valores de pontuação para as substituições mais prováveis e baixos valores de pontuação para substituições mais improváveis[58].

## **MATRIZ PAM**

A base para a formação das matrizes **PAM**[61] foi o exame do padrão de substituição em um grupo de proteínas que compartilhavam 85% de similaridade. As análises foram realizadas sobre 1572 trocas de aminoácidos que foram avaliadas para 71 grupos de proteína. A construção da matriz **PAM1** envolveu alinhamentos de seqüência e conseqüentemente a construção de árvores filogenética. A construção da árvore filogenética foi realizada usando se a unidade de distância **PAM**. A distância evolucionária de **1 PAM** corresponde a um aminoácido sofrer mutação entre 100 aminoácidos, ou grosseiramente 1% de divergência em uma seqüência

de proteína [60]. Avaliou-se o número de substituições entre as sequências que pertenciam a um nó, de modo que a pontuação das matrizes PAM foram derivadas das frequências de substituições de um aminoácido ser trocado por outro. Baseado nestas probabilidades, as pontuações foram geradas pela aplicação da formula 2.5:

$$Sub_{i,j} = \log\left[\frac{q_{i,j}}{p_i \cdot p_j}\right] \quad (2.4)$$

onde  $p_i$  é a probabilidades com que ocorre o resíduo  $i$  entre todas as proteínas, onde  $p_j$  é a probabilidades com que ocorre o resíduo  $j$  entre todas as proteínas. A quantidade  $q_{i,j}$  representa com que frequência os aminoácidos  $i$  e  $j$  são vistos alinhados um com outro nos alinhamentos múltiplos das famílias das proteínas. Desta forma, a razão  $Sub_{i,j}$  representa a taxa da frequência observada versus o produto da frequência do aminoácido  $i$  pela frequência do aminoácido  $j$ . Comumente observamos substituições  $Sub_{i,j}$  com valores maiores do que zero. Para substituições menos frequentes espera-se valores de substituições  $Sub_{i,j}$  menores que zero. Quando o número de frequências observadas e a frequência aleatória são as mesmas  $Sub_{i,j}$  é zero[60].

Várias hipóteses foram feitas na construção das matrizes PAM. Uma das mais importantes é que as substituições dos aminoácidos é independente da mutação prévia em uma mesma posição. Por causa desta hipótese, a matriz original foi extrapolado para se prederizer frequências de substituições à matriz PAM80 é produzida pelos valores da matriz PAM1 multiplica por ela mesma  $\times 80$ . Isto não quer dizer que 80 de 100 aminoácidos variaram, porque haverão intermediários entre as matriz PAM1 e PAM80. Por causa disso, a matriz PAM80 corresponde a %50 das taxas de mutações observadas[58].

O aumento do número da matriz PAM correlaciona aumento da unidade PAM e suas distâncias evolucionárias das seqüência das proteínas 2.5. Por exemplo , PAM250, que tem 20% de identidade dos aminoácidos, representa 250 mutações por 100 aminoácidos. Em teoria, o número de mudanças evolucionárias correspondem a um tempo evolucionário de 2,5 milhões de anos. Desta forma matrizes PAM de números baixos são mais apropriadas para alinhamentos de seqüência com alta similaridade e matrizes de número alto é mais apropriado para alinhar seqüência divergentes[58]

## **MATRIZ BLOSUM**

As matrizes PAM estão baseadas nas taxas evolucionárias das proteínas derivadas de alinhamentos de seqüência de identidade de 85%. Entretanto a maioria das seqüência que realizam o mesmo trabalho (função) tem identidades menores que 85% de similaridade, de modo

**TABLE 3.1.** Correspondence of PAM Numbers with Observed Amino Acid Mutational Rates

PAM Number	Observed Mutation Rate (%)	Sequence Identity (%)
0	0	100
1	1	99
30	25	75
80	50	50
110	40	60
200	75	25
250	80	20

Figura 2.5:

que as matrizes PAM não são adequadas para se avaliar alinhamentos de seqüência divergentes [59].

Reconhecendo as limitações da metodologia para o cálculo das matrizes PAM, Henikoff e Henikoff derivaram outras matrizes de substituição para aminoácidos usando uma metodologia diferente, denominada de **BLOSUM**. As matrizes BLOSUM formam um série de matrizes de blocos de substituições. Todas as matrizes BLOSUM derivadas foram baseadas nas observações diretas de todas as possíveis substituições de aminoácidos em alinhamentos de várias seqüências[59].

Os estudos dos padrões conservados das proteínas de Henikoff [62], possibilitaram a criação do critério de **BLOCO** de banco de seqüências. A idéia de **bloco** é derivada de uma noção mais familiar de um motif, que usualmente se refere aos aminoácidos conservados que conferem uma função ou estrutura específica. Quando estes padrões individuais de proteínas, de uma mesma família sem introdução de um gap, são observados, o resultado é um bloco de seqüência. Desta forma o termo **bloco** refere-se a um alinhamento de aminoácidos, não se referindo a seqüência toda de uma proteína[60].

Naturalmente, qualquer seqüência de proteína pode conter mais de um bloco, correspondendo assim cada bloco a um padrão de estrutura ou função. Com o conceito de bloco em mãos, Henikoff analisou os padrões conservados de substituições para mais 2000 blocos, que representavam mais de 500 grupos de proteínas.

Obviamente muito mais seqüências de proteínas foram avaliadas nos trabalhos realizados por Henikoff 1992[64] por Henikoff do que Dayhoff em 1978 [61], providenciando uma base de dados mais robusta para a análise dos padrões de substituição de aminoácidos. No entanto, a distinção mais importante entre as matrizes BLOSUM E PAM, é que as matrizes BLOSUM possuem dados de várias distâncias evolucionárias fornecendo uma visão mais apurada sobre as substituições de aminoácidos[60].

A pontuação das matrizes BLOSUM para um par de resíduos é derivado da seguinte equação:

$$Sub_{i,j} = \log_2 \left[ \frac{q_{i,j}}{p_i \cdot p_j} \right] \quad (2.5)$$

onde  $p_i$  é a probabilidades com que ocorre o resíduo i entre todas as proteínas, onde  $p_j$  é a probabilidades com que ocorre o resíduo j entre todas as proteínas. A quantidade  $q_{i,j}$  representa com que freqüência os aminoácidos i e j são vistos alinhados um com outro nos alinhamentos múltiplos das famílias das proteínas. Desta forma, a razão  $Sub_{i,j}$  representa a taxa o observado versus a freqüência aleatória para a substituição do aminoácido i pelo aminoácido j. A diferença desta forma de calcular a taxa de substituição é que para as matrizes BLOSUM utilizam ologaritmo na base 2 ao invés do logaritmo na base 10 usada para as matrizes PAM.

Cada matriz BLOSUM é assinalada por um número  $n$  (**BLOSSUM $n$** ), de modo que este número  $n$  representa o nível de conservação das seqüência que foram usadas para derivar aquela matriz de substituição em particular. Por exemplo a matriz BLOSUM62 é calculada para seqüência que compartilham com mais que 62% de identidade.

## GAPS

Os algoritmos que realizam as alinhamentos entre seqüência freqüentemente envolvem a aplicação de gaps para representar inserções e deleções. Por causa da seleção natural, os processos de inserção e deleção são relativamente raros em comparação a substituições. A introdução de gaps é custoso computacionalmente para refletir os eventos raros de inserção e deleção encontrados na evolução. No entanto, escolher os valores de penalidades pode ser uma tarefa mais ou menos arbitrária, por causa da teoria evolucionária que determina um custo para a deleção e inserção.

Os valores das penalidades para os gaps devem ser bem escolhidos de modo adequado. Se os valores das penalidades forem baixos demais pode ocorrer de seqüências não relacionadas possuírem altos valores de pontuação. Se os valores de penalidades forem altos demais, os gaps podem se tornar difíceis de mostrar identidade entre para alinhamentos entre seqüência com certa similaridade, o que não seria realístico. Através de estudos empíricos para proteínas globulares, um conjunto de penalidades apropriado tem sido desenvolvido. Estas penalidades são implementados como valores default na maioria dos programas desenvolvidos que estão disponibilizados na WEB [58].

## 2.5 Alinhamento Múltiplo

A extensão natural do alinhamento entre duas seqüências é o **alinhamento múltiplo**. Um alinhamento múltiplo deve envolver no mínimo três seqüências. Um alinhamento de duas seqüências é chamado de **alinhamento por pares**. A vantagem da realização de alinhamento múltiplo é que este tipo de alinhamento pode revelar mais as informações biológicas do que o alinhamento por pares. Um exemplo é, a identificação de padrões e motifs conservados de seqüências em uma família de seqüências, que não possui uma óbvia detecção pela comparação de duas seqüências. Muitos aminoácidos conservados, críticos para a funcionalidade da proteína, podem ser identificados mais facilmente no alinhamento múltiplo. O alinhamento múltiplo é um pré-requisito essencial para uma análise filogenética de famílias de proteínas e predição da estruturas secundárias e terciárias [58].

### 2.5.1 Método Hierárquico

Alguns dos métodos práticos mais apurados para automatizar os alinhamentos múltiplos são os métodos hierárquicos. Primeiro, todos os pares do conjunto de seqüências escolhidas são comparados pelo método de alinhamento por pares. Isto fornece então um conjunto de similaridades por pares, que podem ser realizados por programas de análise de aglomerados ou de cálculo de árvores hierárquicas. As árvores são calculadas de modo que os pares de seqüências mais similares estejam mais próximos do que os menos similares. Isto é feito avaliando a pontuação de todos os pares de alinhamentos possíveis entre as seqüências escolhidas[58].

### 2.5.2 CLUSTALW

O ClustalW é um dos programas de alinhamento múltiplo hierárquico mais populares na comunidade científica. Ele está disponível para utilização gratuita no endereço eletrônico <http://www.clustal.org>. Neste endereço você pode encontrar pacotes para downloads e servidores que realizam alinhamentos on line. Ele combina um método robusto para alinhamento múltiplo de seqüências com uma interface fácil de usar[60].

O programa usa uma série de matrizes de pontuação por pares para indicar a localização de gaps. Posteriormente segue-se um re-alinhamento das seqüências alinhadas para que seja refinado o alinhamento. O ClustalW pode interpretar a estrutura secundária, que pode ser usada para indicar a posição de quaisquer gaps inseridos; o ClustalW pode interpretar dois alinhamentos pré-existentes e posteriormente alinhar um com outro, ou pode alinhar um conjunto de seqüências de um alinhamento existente. O processo é repetido até que todas as seqüências estejam alinhadas[60].

O ClustalW também inclui opções de inferências filogenéticas através de construção de árvores. Contudo o ClustalW não disponibiliza ferramentas de visualização destas árvores. Entretanto saída pode ser compatível com o programa que cria árvores filogenéticas **PHYLIP**. O ClustalW pode ler uma variedade de formatos que podem produzir diferentes formatos de saída. Em nossos trabalhos utilizamos seqüências somente no formato FASTA[60].

O alinhamento hierárquico depende de um conjunto de etapas do alinhamento múltiplo. A divisão em etapas facilita e acelera a implementação do ClustalW. Primeiro o programa conduz um alinhamento por pares para cada par possível de seqüências. As pontuações para cada um dos pares de seqüências podem ser por pontuação(Score) ou porcentagem de idênticas das seqüências. Ambas as pontuações estão correlacionadas com as distâncias evolucionárias entre seqüências. As pontuações são convertidas então em distâncias evolucionárias para gerar uma matriz de distância para todas as seqüências envolvidas. Uma análise filogenética simples então

é executada baseada na matriz de distância às seqüências do grupo baseadas em pontuações da distância dos pares de seqüências[58].

Em consequência, uma árvore filogenética é gerada usando o método simples. A árvore com referência à proximidade evolucionária. Precisa ser enfatizado que a árvore resultante é uma árvore aproximada e não tem o rigor de uma árvore filogenética formalmente construída. No entanto, a árvore pode ser usada como um guia dirigindo o realinhamento das seqüências. Por essa razão, é que frequentemente utiliza-se o termo árvore guia. De acordo com a árvore guia, as seqüências estreitamente relacionadas duas a duas são realinhadas[60].

Primeiro se realinham as duas seqüências mais fortemente relacionadas. Para se alinhar as seqüências adicionais, as duas seqüências já alinhadas são convertidas em uma seqüência consenso. A seqüência consenso é então tratada como uma única seqüência na etapa subsequente. Neste etapa, posterior, a seqüência mais fortemente correlacionada com a seqüência consenso na árvore do guia é alinhada. Após o realinhamento de uma seqüências, uma nova seqüência é gerada até que toda seqüência escolhida tenha sido alinhada.[58].

### 2.5.3 Cluster Hierárquico(Agrupamentos)

Existem diversas abordagens de clustering (agrupamentos), tais como: probabilística, otimização, clumping e hierárquica [67, 68]. Cada abordagem difere, uma da outra, pela maneira como representa os elementos dos clusters. Os agrupamentos presentes neste trabalho são obtidos por meio de um algoritmo de clustering hierárquico. Este algoritmo faz o agrupamento dos indivíduos com características similares e os representa na forma de um dendograma como da figura 2.6, que consiste de um tipo especial de árvore, na qual os nós pais agrupam os exemplos representados pelos nós filhos. Assim, um agrupamento hierárquico reúne dados de modo que se dois exemplos são agrupados em algum nível, nos níveis mais acima deles continuam fazendo parte do mesmo grupo, construindo uma hierarquia de clusters. Com o uso desta técnica, pode-se analisar os clusters em diferentes níveis de granularidade, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos.

Duas abordagens podem ser derivadas do clustering hierárquico: **aglomerativo (Bottom-up)** e **divisivo (Top-down)**. Na primeira abordagem, os dados são inicialmente distribuídos de modo que cada exemplo represente um cluster e,então, esses clusters são recursivamente agrupados considerando alguma medida de similaridade, até que todos os exemplos pertençam a apenas um cluster. Na segunda abordagem, na decisiva, o processo inicia-se com apenas um agrupamento contendo todos os dados e segue dividindo-o recursivamente segundo alguma métrica até que alcance algum critério de parada, frequentemente o número de clusters desejados[66]

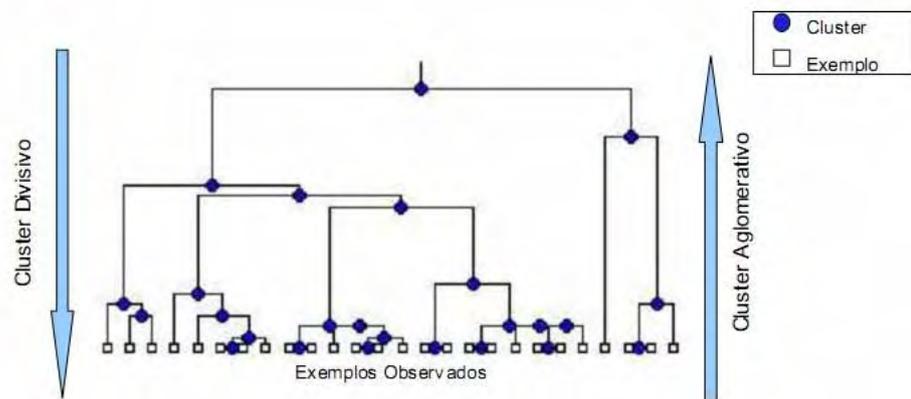


Figura 2.6: Representação de dendograma que demonstrando como os exemplos são agrupados em clusters. O círculo assinala um cluster (agrupamento) de dois exemplos. Do lado esquerdo da figura a seta representa a direção o cluster (aglomeramento divisivo). Neste algoritmo o processo inicia-se considerando um único agrupamento e em seguida divide se este grupo inicial recursivamente por uma determinada métrica até que alcance algum critério de parada, em que cada exemplo represente um cluster. A seta da direita indica a direção do algoritmo de cluster aglomerativo. Neste segundo algoritmo, inicialmente cada exemplo é considerado como sendo um cluster. Em seguida, os exemplos de maior similaridade são recursivamente agrupados até todos os exemplos pertençam a um só cluster

## 2.6 Sorting points into neighborhoods (SPIN)

**Sorting points into neighborhoods** significa Ordenação de pontos na vizinhança ou simplesmente SPIN [69]. Esta é a denominação de uma técnica utilizada para tratamento de dados de múltiplos objetos em aglomerados. No entanto, pela ênfase no particinamento dos dados ,em geral as aproximações de aglomerados negligenciam a questão a cerca de como os dados multidimensionais estão distribuídos. O SPIN, por outro lado, focaliza o significado do ordenamento e a apresentação, ajudando assim a elucidar a estrutura local e global dos dados [69].

### 2.6.1 Estabelecimento Formal do SPIN

A entrada do SPIN é uma matriz das distância  $D \in R^{n \times n}$  para os dados compostos de  $n$  pontos. A saída do SPIN é uma matriz de distância reordenada, obtida pela permutação de  $n$  objetos de acordo com uma permutação particular  $P \in S_n$ . Nós denotamos por  $P$  a matriz também como a matriz permutação associada com  $P$ . Na busca de um critério para apresentação visual dos dados o SPIN se preocupa com dois pontos, que por vezes podem ser conflitantes. Primeiro, as grandes distâncias devem ser empurradas para as bordas. Segundo as menores distâncias devem ser empurradas para a diagonal. Para cada uma destes pontos o programa SPIN utiliza um algoritmo diferente para a reordenação da matriz de entrada, que são denominados de **Side-to-Side** e **Neighborhood** respectivamente.

Estes atributos podem ser matematicamente formulados pela introdução da função custo  $F$  quantificando a qualidade da permutação. Desta maneira, o re-ordenamento da matriz pelo programa SPIN torna se em encontrar a permutação  $P$  que minimiza  $F \equiv F_D : S_n \rightarrow \Re$ . A família de funções encontradas é dada por 2.6:

$$F(P) = tr(PDP^T)W = \sum_{i,j}^n W_{ij}D_{P(i)P(j)} \quad (2.6)$$

, onde  $tr$  denota o traço da matriz,  $W \in R^{N \times N}$  são os pesos da matriz. Para esta família, o problema de otimização é conhecido como **Problema de Associação Quadrática (QAP)** , introduzido por [70]. O **QAP** geral é um problema de otimização extremamente difícil, tratado como *NP-Hard* para então ser aproximado e então ser utilizado na prática. A propriedade **STS** é capturada pela escolha da ponderação para pesos sendo definida por  $W = XX^T$  para um vetor  $X$  de score da forma  $X_i = i - \frac{n+1}{2}$ . A propriedade **Neighborhood** é refletida pela escolha de  $W$  simétrico e concentrado em uma região, determinada pelo parâmetro  $\sigma$ , em torno da diagonal principal.

## 2.6.2 Algoritmo STS

O algoritmo STS é um problema NP-Completo pela redução de um problema **k-click** em teoria de grafos. O algoritmo STS é pode ser explicado pelo fluxograma 2.6.2

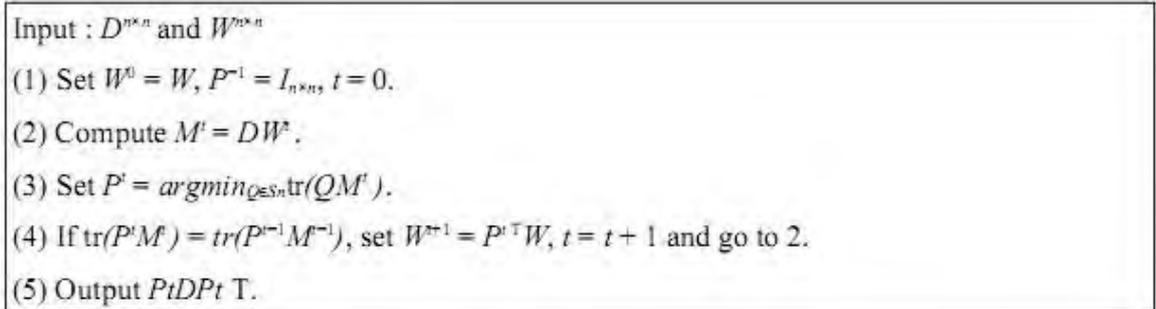


Figura 2.7: Inicia se o programa com uma matriz  $D(n,n)$  das dissimilaridades de  $n$  objetos. No passo inicial  $t=0$ , calcula se  $W_{t=0} = W$ , de modo a esta matriz seja simétrica em relação a diagonal principal. De 2 a 4 corresponde a 1 iteração  $t$ , através de um conjunto de permutações que podem ser realizadas, e assim o algoritmo realiza o re-ordenamento da matriz  $D(n,n)$  com uma saída  $PtDt^T$

Cada passo de 2-4 no fluxograma 2.6.2 corresponde a uma iteração. A complexidade deste algoritmo é  $O(n^2)$ . Cada iteração STS é vista como o mapeamento como o mapeamento do grupo das permutações nele mesmo,  $G_D : S_n \rightarrow S_n$ . Desta forma  $P$  é uma possível saída do STS se e somente é u ponto fixado de  $G_D$ . Após um número finito de passos, a matriz de distâncias converge. A prova é dada pelo fato de que toda iteração o algoritmo reduz a função de custo,  $F$ , garantindo assim convergência a um local de mínimo[69].

Note que o STS pode convergir a um  $P$  que não corresponde ao mínimo global de  $F$ ; para permutações iniciais diferentes o algoritmo pode terminar em diferentes pontos, com valores diferentes de  $F$ , visto que trata-se de um algoritmo que emprega heurística. Uma estratégia comumente utilizadas para combater tal problema é iniciar o algoritmo com diferentes gerações de permutação aleatórias, e escolher o melhor ponto obtido [69].

O algoritmo Side-to-Side gera uma matriz de distância que preferencialmente coloca os elementos mais escuros (que denotam grande similaridade) -figura 2.6.2 perto do canto superior esquerdo e inferior. Assim pontos que são colocados muito distantes na ordenação linear são

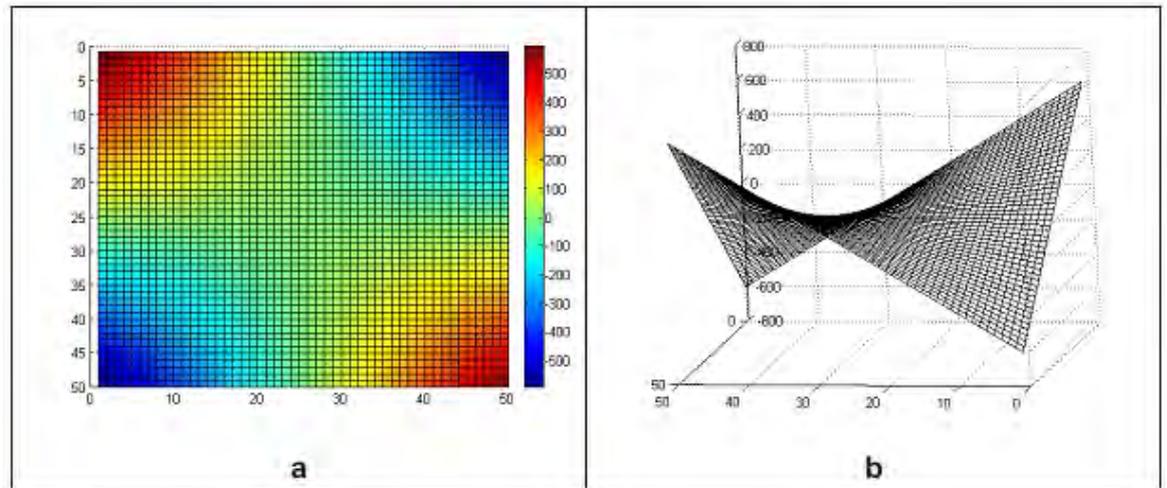


Figura 2.8: Em (a) Representação gráfica das dissimilaridades, segundo o algoritmo Side-to-Side os elementos com maior dissimilaridades (vermelho) e menor (azul) dissimilaridade estão dispostos nas extremidades (b) O gráfico disposto disposto em 3 dimensões

dispostos em extremidades opostas do gráfico, como pode ser notado na representação 3D do gráfico 2.6.2.

### 2.6.3 Algoritmo Neighborhood

O algoritmo Neighborhood se preocupa em recolocar um ponto A em colocar o melhor fig para vizinhança local, isto é, nenhum dos pontos vizinhos de A possuem as maiores distâncias. O algoritmo está sintetizado no fluxograma 2.6.3

Cada passagem dos passos 2-4 do fluxograma 2.6.3 constituem uma iteração Neighborhood. O tamanho da vizinhança é ditado pela escolha do W o que afeta a escala cujos objetos são distinguíveis. O passo 3 pode ser resolvido exatamente em tempo não-polinomial utilizando o algoritmo húngaro. Esta função reflete a suposição para localização mais adequada de todos os pontos de dados de determinada linha da matriz de entrada, ou seja em cada iteração, os pontos são enviados para uma nova localização, baseada na ordenação atual dos dados[69].

Sendo assim, o ponto A é enviado para um novo local  $i(A)$  e perto dele são movidos outros pontos que tenham distâncias próximas de A. No entanto, como vários pontos são permutados simultaneamente, não há garantia de que esta nova configuração continue sendo otimizada, sendo necessária uma nova iteração. Sabendo-se que o problema da Associação Lin-

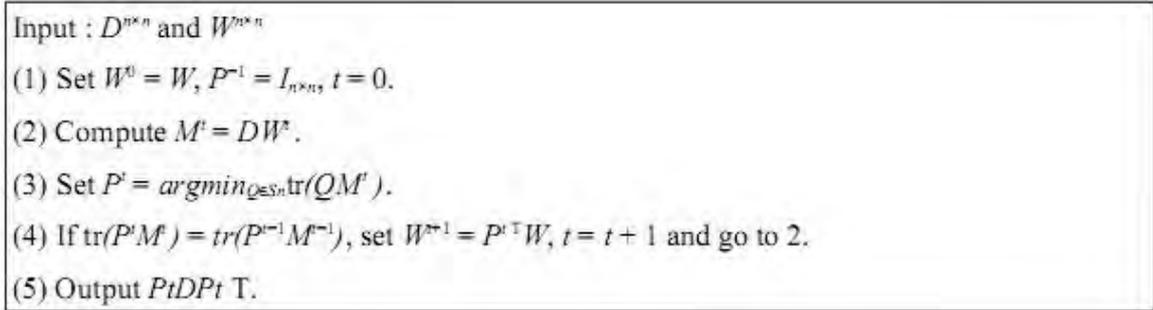


Figura 2.9: Inicia se o programa com uma matriz  $D(n,n)$  das dissimilaridades de  $n$  objetos. No passo inicial  $t=0$ , calcula se  $W$  de forma gaussiana, de modo a esta matriz seja simétrica em relação a diagonal principal. De 2 a 4 corresponde a 1 iteração  $t$ , através de um conjunto de permutações que podem ser realizadas, e assim o algoritmo realiza o re-ordenamento da matriz  $D(n,n)$  com uma saída  $P^t D P^t T$

de

ear é um problema conhecido e pode ser resolvido em  $O(n^3)$  [71]. A complexidade de cada iteração  $O(n^3)$ . O custo é melhorado em toda iteração e a convergência até certo ponto é garantida após um número finito de vezes.

De acordo com o passo 4, o algoritmo termina quando percebe que a iteração atual não obteve mudança em relação à anterior com relação ao cálculo da função de custo ( $W$ ). Isso previne que hajam ciclos de custo constantes. Mais uma vez, sabendo que o espaço de permutação é finito, tem-se a garantia de que o término da iteração se dará após um número finito de passos, convergindo a certo ponto[69].

A matriz dos pesos é dada por:

$$W_{ij} = e^{-\frac{(i-j)^2}{n\sigma}} \tag{2.7}$$

que é normalizada em uma matriz estocástica dupla, isto é, a soma de cada linha ou coluna é um. No caso a matriz  $M = D.W$  pode ser vista suavização da variaria  $\sigma^2$  de cada linha de  $D$ . Para um certo conjunto de dados existe um intervalo de escalas de tamanho considerável, onde grandes escalas refletem o tamanho excessivo no intervalo de dados, enquanto que valores menores dão uma noção de uma organização de possíveis fragmentos de grandes estruturas. Essa disparidade no layout dos dados é controlado pelo valor de  $\sigma$ . Além disso, a solução do problema de associação linear (passo 3 no algoritmo) pode ser eficientemente aproximado

pela busca do mínimo de cada linha de  $M$ , em seguida pela ordenação dos índices da matriz de distância, a fim de disponibilizar os valores de distância aproximados sempre ao redor da diagonal principal.

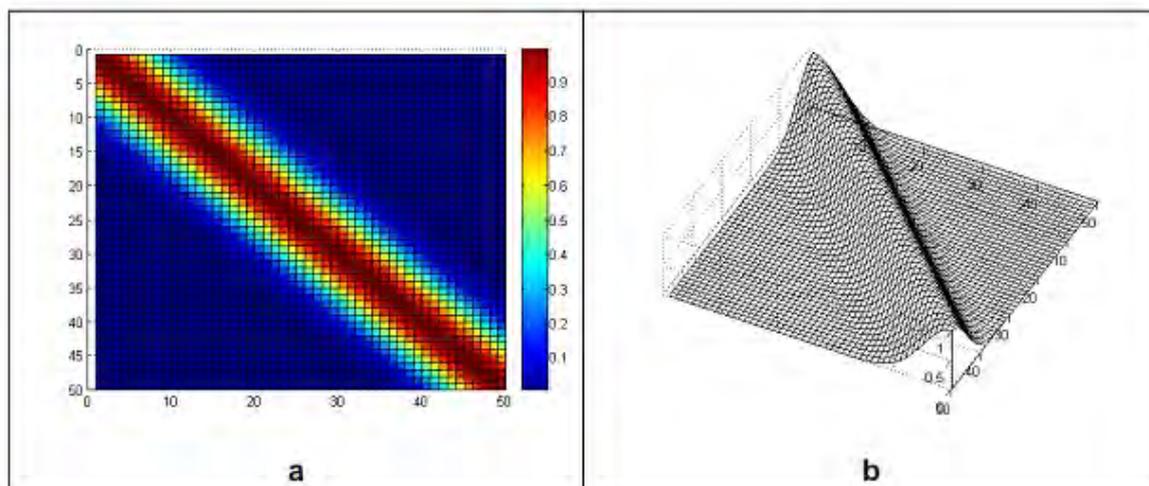


Figura 2.10: **Representação gráfica do algoritmo Neighborhood.** Em (a) vemos a cor vermelha representando elementos de maior dissimilaridade e o azul representando menor dissimilaridade como demonstra a escala ao lado da figura em (b) notamos como fica disposto o gráfico em 3D

O algoritmo Neighborhood (vizinhança), ao contrário do STS, tenta fazer com que elementos localizados perto da diagonal principal, identificados pela cor vermelha, sejam aqueles com menores distâncias como estão representados na figura 2.6.3. Pontos vizinhos na ordenação linear são agrupados em grupos que ficam concentrados em torno da região da diagonal principal, constituindo blocos de elementos semelhantes[69].

## 2.7 Hidrofobicidade

Em todos os sistemas em que as interações são mediadas por moléculas de água, as interações hidrofóbicas parecem causar o aglomeramento das unidades hidrofóbicas. O primeiro a observar a influência das interações sobre os sistemas biológicos foi Walter Kauzmann[11] e sua conclusão de que as interações hidrofóbicas são as mais importantes no processo de enovelamento é largamente aceita atualmente. Entretanto, a quantificação dos termos que influenciam a hidrofobicidade é difícil, pois o efeito hidrofóbico é um fenômeno multifacetado, que se manifesta

de diferentes maneiras para pequenas unidades moleculares ou grandes aglomerados, que estão envolvidos separadamente ou quando se encontra a combinação de ambos[73].

O efeito hidrofóbico é derivado das propriedades especiais da água como solvente. De fato, outros solventes polares, tais como o dietilsulfoxido e N,N dimetilformamide tendem a denaturar proteínas. O efeito hidrofóbico é a tendência da água em excluir moléculas e átomos apolares do contato com as moléculas de água.

Este efeito é devido à habilidade especial das moléculas de água em formarem pontes de hidrogênio com moléculas de água vizinhas. Por causa da diferença na eletronegatividade entre os átomos de oxigênio e hidrogênio, as ligações eletrônicas na molécula de água estão distribuídas assimetricamente, de forma que o oxigênio fica parcialmente carregado cargas negativas e o hidrogênio parcialmente carregado positivamente. Deste modo, o átomo de hidrogênio  $H$  da ligação  $O-H$  é atraído pelo átomo de oxigênio  $O$  da ligação  $O-H$  de outra molécula de água. Esta atração resulta em um próton de hidrogênio sendo compartilhado entre duas moléculas de oxigênio com duas moléculas de água. Isto é, a ligação de hidrogênio faz a interação água-água ser energeticamente mais favorável do que interação água-soluto hidrofóbico[5].

Moléculas polares eletricamente neutras, tal como o etanol, ou com cargas líquidas, como os aminoácidos, podem interagir com as moléculas. Por exemplo, o grupo funcional  $COO^-$  pode ser aceitador da ponte de hidrogênio e o grupo funcional  $NH_3^+$  pode funcionar como doador para a molécula de água, resultando assim em uma interação também energeticamente favorável [5].

No entanto, em água pura, moléculas apolares, como o hexano, não possuem nenhum grupo que possa ser doador ou aceitador para as pontes de hidrogênio das moléculas de água. Deste modo, ao introduzir uma molécula apolar na água as pontes de hidrogênio serão quebradas e posteriormente reconstruídas ao em torno da molécula apolar, criando assim uma cavidade. Desde que, as ligações de hidrogênio no espaço ocupado pela molécula apolar não possa ocorrer, uma rede de pontes de hidrogênio é criada entorno da molécula apolar, de tal modo que a rede criada em torno da molécula apolar não é aleatória, sendo sim mais ordenada. Este ordenamento provoca uma diminuição da variação de entropia  $\Delta S$ , significando um aumento da ordem do sistema. Mesmo a variação de entropia sendo negativa e  $\Delta G$  poder variar livremente, a variação da energia livre  $\Delta G = \Delta H - T\Delta S$  é positiva.. Aqui  $T$  é a temperatura absoluta e  $\Delta H$  é a entalpia. Este fenômeno torna-se um processo energeticamente desfavorável [5].

A energia livre de hidratação desfavorável de substâncias apolares provocado pelo ordenamento das moléculas de água vizinhas é o resultado final da expulsão da substância fase aquosa. Isto ocorre porque a área da superfície da cavidade contendo a agregação de substâncias apolares é menor do que a soma das áreas das superfícies de cada molécula individualmente[1].

A agregação de grupos apolares minimiza a área de superfície da cavidade e consequentemente ocorre a perda de entropia do sistema. Nesse sentido, os grupos apolares estão expulsos pela fase aquosa por causa das interações hidrofóbicas. A medida da energia livre para retirar um grupo  $-CH_2-$  de uma solução aquosa é cerca de  $-3KJ.mol^{-1}$ . Este valor é relativamente pequeno, no entanto quando este valor se junta a um conjunto de muitos grupos apolares, as interações hidrofóbicas tornam-se mais intensa[1].

## 2.8 Escalas Hidrofóbicas

Walter Kauzmann, em 1959 apontou que as interações hidrofóbicas eram as que interações de maior influência no processo de enovelamento de proteínas, indicando que as cadeias laterais dos aminoácidos deveriam ser classificadas por índices que mediam a hidrofobicidade e hidrofiliidade de cada aminoácido. De modo que, vários pesquisadores tem se dedicado a classificar os aminoácidos padrões por um índice que demonstre suas características hidrofóbicas e hidrofílicas e assim na literatura científica especializada foram geradas várias escalas hidrofóbicas [76].

As escalas hidrofóbicas são listas de classificação da hidrofobicidade relativa de cada aminoácido. Estas hidrofobicidades relativas são determinadas por técnicas experimentais diversas. Sabe-se que dentre as interações eletrostáticas, pontes de hidrogênio, hidrofóbicas, e de van der Waals, as interações hidrofóbicas são as mais importantes na determinação da estrutura nativa da proteína. Entretanto, o efeito hidrofóbico sozinho, não representa o comportamento completo dos aminoácidos. De qualquer modo certos pesquisadores acreditam que a melhor compreensão a respeito da hidrofobicidade relativa dos aminoácidos poderá melhorar as predições de como as proteínas se enovelam em suas estruturas terciárias [74].

Muito esforço tem sido investido em estudar o hidrofobicidade dos aminoácidos. Em consequência, muitas escalas diferentes de hidrofobicidade para aminoácidos padrões foram desenvolvidas[76]. O site endereço eletrônico <http://www.expasy.ch/tools/protscale.html> disponibiliza a maioria das escalas hidrofóbicas existente na literatura. Este endereço eletrônico fornece 57 escalas hidrofóbicas diferentes. Tais escalas hidrofóbicas são determinadas através de dados experimentais para técnicas e metodologias diferentes. Por isso valores dos índices de hidrofobicidade para um mesmo aminoácido podem variar entre as diferentes escalas. Podemos dividir grosseiramente em 5 categorias as metodologias para se obter as escalas hidrofóbicas. Agora apresentaremos resumidamente as características de cada uma dessas metodologias.

## 2.8.1 Método do Particionamento

O método mais comum de medida da hidrofobicidade é realizar o particionamento entre duas fases líquidas imiscíveis. Diferentes solventes orgânicos imiscíveis são utilizados para imitar o interior da proteína. Entretanto, estes solventes orgânicos são ligeiramente miscíveis com água e avaliação da mudança entre ambas as fases, de modo que fica difícil se obter uma escala pura para a hidrofobicidade de cada aminoácido[76].

Tanford e seus colaboradores propuseram a primeira escala hidrofóbica para nove aminoácidos. O álcool etílico e o dioxane foram usados como os solventes orgânicos e a energia livre de transferência de cada aminoácido foi calculada. As fases não líquidas podem ser igualmente usadas com divisão de métodos tais como fases micellar e fases de vapor. Fendler [78] e seus colaboradores mediram o particionamento de 14 aminoácidos usando micelas de sulfato dodecyl de sódio (SDS). Também, a afinidade da cadeia lateral do aminoácido para a água foi medida usando fases de vapor[80].

O potencial da hidratação e sua correlação à aparência dos aminoácidos na superfície das proteínas foram estudados por Wolfenden. As fases aquosas e do polímero foram usadas no desenvolvimento de uma nova, **escala Radzicka**[81]. Esta escala encontrou a distribuição dos coeficientes dentro-fora das cadeias laterais dos aminoácidos de proteínas globulares. A distribuição dos coeficientes das cadeias laterais dos aminoácidos para os solventes **ciclohexano-água** foi avaliada. Esta distribuição forneceu um índice experimental de suscetibilidade a atração através das forças de dispersão, que correspondiam as energias livres dos aminoácidos, que foi observado estarem linearmente correlacionados com as áreas das superfícies das cadeias laterais dos aminoácidos[81].

A **escala Kyte-Doolittle** avaliou a hidrofobicidade utilizando dados experimentais de proteínas. Eles avaliaram a hidrofobicidade média dos aminoácidos continuamente utilizando um comprimento pré-determinado. Para cada comprimento (janela) se avaliava, para o aminoácido central, a energia livre de transferência da fase água para vapor condensado, bem como a distribuição das cadeias laterais 100% e 95% enterrados. Kyte-Doolittle encontraram a hidrofobicidade fazendo a média destas três grandezas. Quando um dos valores encontrados tinha uma discrepância, fazia-se a média dos outros dois valores [82].

Roseman determinou a **escala  $\pi$ -r**, usando modelos de peptídeos para avaliar os valores de  $\pi - r$  para as cadeias laterais dos aminoácidos, onde  $P(H - R)$  corresponde ao coeficiente de partição entre o octanol/água do constituinte da cadeia lateral R. A grandeza  $\pi - r$  é dada pela equação (2.8)

$$\pi = \log\left[\frac{P(H - R)}{P(H - H)}\right] \quad (2.8)$$

$P(H - H)$  corresponde ao coeficiente de partição água/octanol para 2 átomos de hidrogênio [83].

Os métodos de particionamento possuem muitos inconvenientes. Primeiramente, é difícil imitar o interior da proteína, além do mais, a regra da auto-solvatação dos aminoácidos livres é muito difícil de ser aplicada. Além disso, ligações de hidrogênio que são perdidas em transferência aos solventes orgânicos não são reformadas mas freqüentemente no interior da proteína[84].

### 2.8.2 Métodos da Área Acessível ao Solvente (ASA)

Neste método para cada aminoácido é calculada a área de acessível ao solvente (**acessibilidade**), estática para moléculas de solvente através da utilização de programas de computador. A vantagem principal de usar métodos de cálculo de área de acessível é que a modelagem do interior do soluto ou da proteína não está envolvida. As coordenadas das proteínas são utilizadas em programas de computador para serem avaliados os graus de exposição e as correlações com hidrofobicidades para todos os aminoácidos[76].

A escala **Chothia** [84] calculou a área de superfície acessível de todos os aminoácidos de 12 proteínas com as coordenadas das cadeias distendidas[84]. Este estudo definiu como superfícies hidrofóbicas, as superfícies formadas por átomos apolares (carbono) ou átomos polares(nitrogênio, oxigênio, enxofre) como formando pontes de hidrogênio. As superfícies de hidrofílicas foram definidas como aquelas com os átomos polares sem ligações de hidrogênio. Entretanto, há algumas desvantagens de métodos acessíveis da área de superfície. Primeiro, a base de dados limitada da estrutura da proteína é bem limitada, assim como sua dependência na definição de átomos polares e não polar. Igualmente não mede a acessibilidade dinâmica das proteínas na solução. Além disso, a diferenciação entre o hidrofobicidade e o hidrofílicidade é feita sem considerara polaridade de superfície [76].

### 2.8.3 Métodos Cromatográficos

A cromatografia líquida da fase reversa(RPLC) é o método cromatográfico que avalia a hidrofobicidade do soluto[85]. A fase estacionária apolar imita as membranas biológicas. O uso do peptídeos tem muitas vantagens porque o particionamento não é extensa às cargas terminais em RPLC. A formação das estruturas secundárias também pode ser evitada pelo uso peptídeos curtos da seqüência.Outra escala usando cromatografia foi desenvolvida em 1971 usando a retenção do peptídeo no gel hidrofílico. Pliska e seus colaboradores[87] usaram a cromatografia de camada fina para relacionar valores da mobilidade de aminoácidos livres a

suas hidrofobicidades. A escala foi obtida através da cromatografia de fase líquida onde foram utilizados 121 peptídeos [88].

Os valores absolutos e as classificações relativas das hidrofobicidades relativas dos aminoácidos determinados por métodos cromatográficos podem ser afetados por um número de parâmetros. Estes parâmetros incluem a área de superfície do silicone e o diâmetro do pore, a escolha e PH do amortecedor aquoso, temperatura e a densidade da ligações das cadeias em fase estacionária[76].

#### 2.8.4 Metagênese de Sítio Dirigido

Este método utiliza a tecnologia de recombinação de DNA para se obter a medida atual da estabilidade da proteína. Em estudos detalhados de metagênese dirigida Yutani[89] e seus colaboradores substituindo 19 aminoácidos por Trp49 da síntese do triptofano e mediram a energia livre de desenovelamento. O interessante é que eles observaram que o aumento da estabilidade é diretamente proporcional ao aumento em hidrofobicidade até um determinado limite do tamanho. A desvantagem principal do método da metagênese sítio dirigida é que não todos os 20 aminoácidos naturais podem substituir um único resíduo em uma proteína. Além disso, estes métodos custaram problemas e foram úteis somente para a estabilidade de medição da proteína[89, 76].

#### 2.8.5 Métodos de Propriedades Físicas

As escalas hidrofóbicas desenvolvidas por métodos da propriedade física são baseadas na medida de propriedades físicas diferentes. Os exemplos incluem, capacidade de calor parcial molar, temperatura de transição e a tensão de superfície. Os métodos físicos são fáceis de serem aplicados. Uma das escalas hidrofóbicas mais populares[90] medindo a tensão superficial para uma solução de NaCl. Os inconvenientes principais de medidas da tensão de superfície são que as ligações de hidrogênio quebradas e os grupos carregados neutralizados permanecem na interface de ar-solução.

Uma escala hidrofóbica, que utiliza um método de propriedade física para se calcular a hidrofobicidade de cada resíduo é denominado de **escala consensus**,foi desenvolvido por Eisenberg [90]. A partir das coordenadas atômicas das estruturas de proteínas, Eisenberg e seus colaboradores, avaliaram a contribuição átomo da proteína para a energia livre de solvatação, que por sua vez, foi definida como produto da acessibilidade de um átomo ao solvente por um parâmetro atômico de solvatação [90].

Outra escala hidrofóbica, que se utiliza de propriedades físicas para se avaliar as hidro-

fobocidades, foi desenvolvida por Miyazawa e Jernigan [92], conhecida como **escala Miyazawa-Jernigan(MJ)**. A hipótese básica deste trabalho é de que a característica média dos contatos resíduo-resíduo, para uma grande quantidade de estruturas cristalográficas de proteínas, refletem as diferenças entre os resíduos de aminoácidos. A **aproximação quase-química** foi utilizada, porque a formação de pares de contatos assemelha-se a uma reação química. Esta aproximação, foi aplicada a um sistema, para se avaliar as fórmulas que relacionam médias estatísticas dos números de contatos com as energias de contato.

O número efetivo de moléculas de solvente para cada proteína foi escolhido para que o número de contatos resíduo-resíduo fosse igual ao valor esperado para o caso hipotético de interações de esfera dura e moléculas de solvente. Cada resíduo é representado pelo átomo central de sua cadeia lateral e os contatos são definidos como sendo pares com 6,5 Å de distância. Os números de coordenação, bem como os de moléculas de solvente foram estimados através do volume médio de cada aminoácido, que por sua vez, foi utilizado para se estimar o número de contatos resíduo-solvente e solvente-solvente do número de contatos resíduo-resíduo. Os resultados obtidos pela **escala Miyazawa-Jernigan(MJ)**, para um conjunto de 43 estruturas de proteínas demonstraram, que os valores estimados de energia de contato, tem uma razoável dependência, que é refletida pela distribuição de resíduos nos cristais de proteínas

## Capítulo 3

# Resultados

Conforme resultados de nosso trabalho anterior [50], proteínas pertencentes a uma certa classe, que se enovelsse em alta hidrofobicidade deveriam possuir uma baixa similaridade entre as seqüências para diferentes espécies. Por outro lado, classes de proteínas, que se enovelam em regime de baixa hidrofobicidade deveriam possuir uma alta similaridade de seqüência entre diversas espécies. Por isso, nosso objetivo de pesquisa no presente trabalho, consistiu em avaliar os efeitos da hidrofobicidade na evolução das proteínas. seqüências De início, pesquisamos nos bancos de proteínas (UniProt e PIR), qual era o maior número de seqüências de lisozima ocorriam para diferentes organismos. O problema de se encontrar seqüências para uma mesma proteína que ocorra em diferentes espécies é a redundância de seqüências encontradas nos bancos da dados de proteínas. Sem realizar nenhum corte, no início de nossas pesquisas encontramos mais de quarenta mil ocorrência para lisozima no banco de dados SWISS-PROT(UniProt), atualmente são encontrados mais de 250 mil ocorrências para a lisozima. Analisando os dados da primeira pesquisa para a lisozima realizada no SWISS-PROT, pode se notar que havia uma grande redundância para as seqüências, bem como os tamanhos das seqüências anotadas variavam largamente. Desta forma, viu se a necessidade da escolha das seqüências ser realizada segundo alguns critérios:

Para nossa análise ser consistente nosso conjunto foi escolhido de modo que as quatro seqüências de proteínas ocorrem para uma determinada espécie, de modo que para cada espécie encontrássemos 4 seqüências de classes de proteínas diferentes.

- seqüências de vertebrados
- o tamanho da seqüência pesquisada não deveria variar mais do que 35% o tamanho da lisozima encontrada na espécie humana

Utilizando os critérios acima, foram encontradas 212 seqüências de lisozima para espécies diferentes. Através dos dados das 212 diferentes espécies que possuíam seqüências de lisozimas, foi realizada uma nova busca, nos bancos de dados UNIPROT e PIR, para se encontrar seqüências de citocromo c, lisozima e histona H3 que ocorressem para as 212 espécies diferentes encontradas na primeira busca. O resultado dessa nova pesquisa é, que encontramos 41 seqüências de citocromo c, mioglobina e histona H3 para as mesmas espécies diferentes, dessa forma foi obtido um conjunto de 41 espécies diferentes que tinham seqüências de proteínas para **Lisozima, Citocromo c, Mioglobina e Histona H3**. A relação entre o ID das proteínas para as quatro classes está listada no quadro 5 do apêndice A.

Formado o banco de dados de quatro classes de proteínas, foi avaliada a identidade média para cada classe. Primeiramente submetemos cada uma das quatro classes de proteínas a um alinhamento múltiplo hierárquico, realizado pelo programa CLUSTALW no endereço eletrônico <http://:ebi.ac.uk/Tools/clustalw2/index.html>. As condições escolhidas para o alinhamento foram de utilizar a matriz de substituições BLOSSUM62 e uma análise sem gap. A saída do alinhamento foi escolhida como sendo por porcentagem de identidade entre as seqüências, ou seja as  $sequences(i, j) = I(i, j)$ , onde o valor  $I(i, j)$  se refere à identidade entre as seqüências i e j. O nosso interesse era em saber qual a distância das similaridades, ou melhor, como se utiliza na literatura as **dissimilaridades** entre as seqüências. As dissimilaridades entre as seqüências foram obtidas realizando a operação 3.1,

$$DIS(i, j) = 100 - I(i, j), \quad (3.1)$$

onde dessa forma  $DIST(i, j)$  se refere à distância entre as seqüências i e j.

Antes de se realizar a comparação, propriamente dita, das médias de hidrofobicidade e similaridade de cada classe de proteína foi analisada a distribuição das dissimilaridades das quatro proteínas durante o processo de evolução. Uma forma de se fazer isto é se comparar a dissimilaridade entre as seqüências i e j com a distância filogenética das espécies i e j. A priori, uma forma seria encontrar o genoma das 41 espécies, fazer o alinhamento destas seqüências e depois calcular as dissimilaridades destes genomas, de modo fosse possível considerar a dissimilaridade  $DISS(i, j)$  entre as seqüências i e j como sendo proporcional a distância filogenética das espécies das proteínas i e j, assim ao se comparar as dissimilaridades de cada uma das quatro classes de proteínas com a dissimilaridades dos genomas estaríamos comparando as dissimilaridades de cada classe de proteína com as distância filogenéticas das espécies envolvidas.

Entretanto, se fosse utilizado como medida da distância filogenética as dissimilaridade das seqüências dos genomas, seria cometido um grave erro, pois o tamanho entre os genomas para várias espécies varia largamente e, conseqüentemente a dissimilaridade entre as seqüências

das espécies  $i$  e  $j$  não corresponderia à distância filogenética entre as espécies  $i$  e  $j$ . Para contornarmos este problema, escolhemos comparar as dissimilaridades das proteínas com as dissimilaridades do DNA mitocôndrias das 41 espécies do nosso banco de dados. A mitocôndria, é um das organelas celulares mais importantes para a respiração celular. Ela está presente na maioria dos nos organismos de reprodução sexuada, de modo que o DNA mitocondrial é transmitido pelo cromossomo materno, e assim a seqüência de genes da mitocôndria é altamente preservado, sofrendo pouca variação em seu tamanho, entre as diversas espécies. Assim, a dissimilaridade entre duas seqüência de DNA mitocondrial poderá ser considerada proporcional a distância filogenética entre essas espécies

Pesquisamos o DNA mitocondrial para as 41 espécies nos endereços eletrônico <http://arsa.ddbj.nig.ac.jp/top.html> do banco de dados de nucleotideos DNA Data Bank of Japan (**DDBJ**), [www.ncbi.nlm.nih.gov/sites](http://www.ncbi.nlm.nih.gov/sites) do National Center Biotechnology Information (NCBI). Depois foram avaliadas as identidades entre as seqüência do DNA mitocondrial das 41 espécies, utilizando o programa CLUSTALW, de modo que posteriormente foram calculadas as dissimilaridades pela transformação utilizada na equação (3.1). O resultado da comparação entre a dissimilaridade das quatro classes de proteínas contra a distância filogenética fornecida pelas dissimilaridades das mitocôndrias está apresentada na figura (3.1)

Se tomarmos a dissimilaridade entre os DNAs mitocôndriais como sendo proporcional a distância filogenética e analisar o conjunto de gráficos representado na figura (3.1). comparando os dados da série Histona H3 → Citocromo C → Mioglobina → Lisozima, pode se observar um aumento da dissimilaridade em relação à distância filogenética, demonstrando que ao longo da evolução as sequências da lisozima divergiram mais do as sequências da histona H3, assim seguindo a série Histona H3 → Citocromo C → Mioglobina → Lisozima, observa-se um aumento da dissimilaridade média de uma classe para outra. Este aumento pode ser verificado através do cálculo das dissimilaridades das quatro proteínas. Os resultados do cálculo das dissimilaridades para as quatro classes estão apresentados na tabela 3.1 Conforme nota-se na

Proteína	Dissimilaridade
Histona	5.298
Citocromo c	11.046
Mioglobina	23.579
lisozima	35.166

Tabela 3.1: **Tabela que correlaciona as famílias de proteínas com suas dissimilaridades médias**

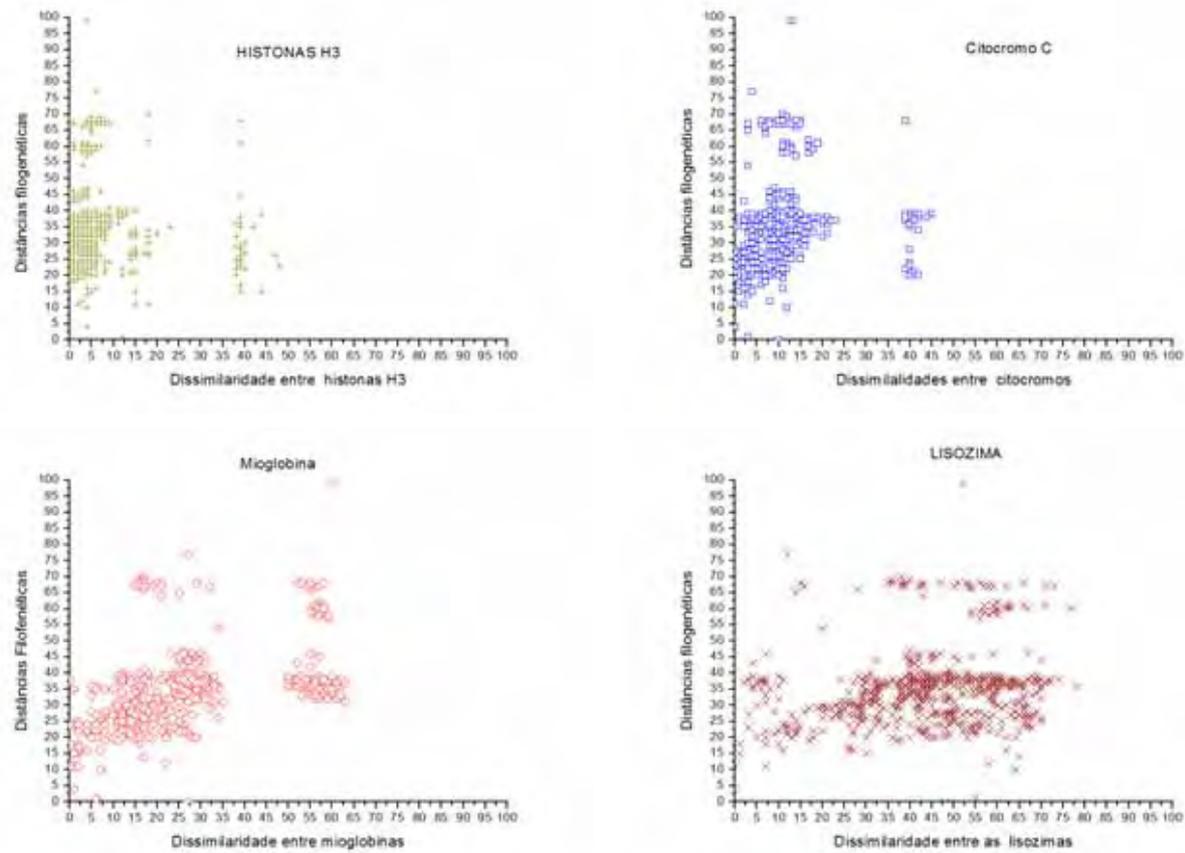


Figura 3.1: Cada ponto reflete a dissimilaridade (distância em homologia) entre um par de espécies ( $i,j$ ). Indicando no eixo y a dissimilaridade entre as mitocôndrias, que corresponde à distância filogenética das espécies  $i$  e  $j$ . No eixo x estão dispostas as dissimilaridades entre as seqüências de proteínas para as espécies  $i$  e  $j$ . Neste gráfico estão representados as dissimilaridades entre as histonas h3, citocromos c, mioglobinas e lisozimas. Para cada um dos quatro gráficos temos 840 pontos

tabela 3.1, a primeira classe de proteína da série (histona h3), possui a menor dissimilaridade média DISS=5.298(ou uma maior identidade) e a última classe possui a maior dissimilaridade média DISS=35,166 (menor identidade) entre as seqüências para as quatro classes pesquisadas. O aumento da dissimilaridade na série de proteínas, nos sugere, pelos resultados de trabalho anterior[50], que tais classes de proteínas se encontrem em condições hidrofóbicas diferenciadas.

Para se analisar a relação entre dissimilaridade e similaridade foram escolhidas seis escalas hidrofóbicas diferentes :

- **Escala Consensus**
- **Escala Chothia**
- **Escala Kyte-Doolittle**
- **Escala Myiazawa-Jerningan (MJ)**
- **Escala Pi-r (Roseman)**
- **Escala Radzscika**

Cada uma dessas escalas teve seus índices hidrofóbicos normalizados segundo a regra 3.2

$$h_i^{norm} = \frac{h_i - h_{min}}{h_{max} - h_{min}}, \quad (3.2)$$

onde  $h_i^{norm}$  se refere ao valor do índice hidrofóbico do aminoácido  $i$  normalizado,  $h_i$  se refere ao valor do índice hidrofóbico do aminoácido  $i$ ,  $h_{max}$  se refere ao maior valor da escala hidrofóbica e por último,  $h_{min}$  se refere ao menor valor do índice hidrofóbico. Desta maneira, as escalas hidrofóbicas foram normalizadas entre um intervalo entre zero e um. As escalas hidrofóbicas normalizadas são apresentadas na tabela (3.2) Utilizando a normalização apresentada em 3.2, para seis escalas hidrofóbicas, foram avaliadas através de uma simulação simples as hidrofobicidades médias, de cada classe de proteína estudada, através de 3.3 :

$$\overline{h_j^{esc}} = \frac{\sum_{i=1}^{N=41} \sum_{j=1}^{20} h_{ij}^{esc} \cdot f_i}{N}, \quad (3.3)$$

de modo que, a hidrofobicidade média da proteína  $j$  na escala  $esc$   $\overline{h_j^{esc}}$  é dada pelo somatório sobre todos os vinte tipos de aminoácidos do produto do valor  $h_i^{esc}$  da hidrofobicidade do aminoácido do tipo  $i$  para a escala  $esc$  pela fração  $f_i$  de fração de aminoácidos do tipo dividido pelo número  $N$  de seqüências diferentes, que no nosso caso é 41. Também foi encontrado a

aminoácidos	Eisenberg	Kyte	Chothia	MJ	Roseman	Radzicika
Ala	0.806	0.700	0.627	0.391	0.698	0.000
Arg	0.000	0.000	0.000	0.202	0.000	0.140
Asn	0.448	0.111	0.186	0.125	0.328	0.152
Asp	0.417	0.111	0.237	0.105	0.022	0.232
Cys	0.721	0.778	0.830	0.819	0.675	0.232
Gln	0.430	0.111	0.102	0.151	0.426	0.300
Glu	0.458	0.111	0.288	0.115	0.167	0.298
Gly	0.770	0.456	0.593	0.252	0.635	0.451
His	0.545	0.144	0.271	0.354	0.532	0.630
Ile	1.000	1.000	1.000	0.967	0.928	0.707
Leu	0.918	0.922	0.746	0.908	0.923	0.708
Lys	0.264	0.066	0.033	0.000	0.190	0.733
Met	0.810	0.711	0.661	0.987	0.790	0.772
Phe	0.951	0.811	0.830	1.000	1.000	0.810
Pro	0.677	0.322	0.289	0.151	0.794	0.811
Ser	0.601	0.411	0.356	0.187	0.436	0.858
Thr	0.634	0.422	0.373	0.254	0.474	0.858
Trp	0.854	0.400	0.440	0.775	0.977	0.935
Tyr	0.714	0.355	0.237	0.483	0.871	1.000
Val	0.923	0.966	0.898	0.770	0.844	1.000

Tabela 3.2: **Tabela dos índices hidrofóbicos normalizados para as seis escalas utilizadas**

distância de similaridade, ou dissimilaridade média  $\overline{DISS}$  3.4, calculando se a média aritmética sobre todas as dissimilaridade possíveis  $\overline{DIST}$  3.4.

$$\overline{DIST} = \frac{\sum_{i=1}^{41} \sum_{j>i}^{41} D(i, j)}{M}, \quad (3.4)$$

Os gráficos apresentados na figura 3.2 apresentam no eixo y a hidrofobicidade média e no eixo x as distância de similaridade para o citocromo c ,histona H3, lisozima e mioglobina. A cada uma das grandezas apresentadas foram avaliados os erros que estão apresentados nos gráficos da figura 3.2. Para Nos histogramas os pontos ciano, vermelho,verde e azul representam os valores obtidos da energia média para o banco de dados de 41 espécies diferentes.

Conforme é possível notar, mesmo sofrendo variações entre as diferentes escalas, existe uma correlação positiva entre a hidrofobicidade e a dissimilaridade, indicando assim que à medida que a distância de similaridade,ou melhor, a dissimilaridade aumenta, ocorre um aumento na hidrofobicidade média. Estes resultados sugerem que famílias de proteínas com alta hidrofobicidade possuem uma menor similaridade entre as seqüências, ao passo que, proteínas que tem baixa hidrofobicidade têm uma maior similaridade de seqüências.

Para analisarmos a significância dos valores médios encontrados em nosso trabalho, foi avaliada a hidrofobicidade média para um conjunto de 2865 seqüências de proteínas não redundantes. Este banco de dados é constituído por 1330 seqüências de enzimas e o restante por diversos tipos de proteínas. Os resultados do cálculo da hidrofobicidade média são apresentados na figura 3, composta pelos seis histogramas de hidrofobicidade média.Os pontos ciano, vermelho, verde e azul representam os valores médios obtidos respectivamente, para a Histona, Citocromo C, Lisozima e Mioglobina.

Os dados obtidos das hidrofobicidades média para as quatro classes de proteínas, com os valores das hidrofobicidades médias do banco de dados estão representados nos histogramas das hidrofobicidades. Os círculos de cores ciano, vermelho, verde e azul representam os valores das hidrofobicidades médias da respectivamente para a histona H3, citocromo c, mioglobina e lisozima para as respectivas escalas hidrofóbicas. Observa-se que dentro de um espectro de hidrofobicidade que , para um conjunto de 2865 sequências de proteínas, os valores das hidrofobicidades médias para as quatro classe de proteínas correspondem a um amplo espectro de energia, demonstrando assim que as variações nos valores de hidrofobicidade são significativas.

Os bancos de dados do UNIPROT(SWISS-PROT) e o PIR são bancos de dados secundários de seqüências de proteínas. Estes bancos de dados priorizam a informação sobre a estrutura primária, entretanto eles também disponibilizam e fornecem diversas informações sobre as funções e estruturas das proteínas. Pesquisando as estruturas das proteínas de nosso banco de dados foi encontrada uma desigualdade entre o número de seqüências e o número de folds diferentes para todas as classes diferentes.

Nestes bancos de dados foram obtidas as 41 estruturas para cada uma das quatro classes de proteínas. De posse do conjunto de estruturas obtidas calculou-se a dissimilaridade das quarenta e uma estruturas para cada uma das classes de proteína pesquisadas

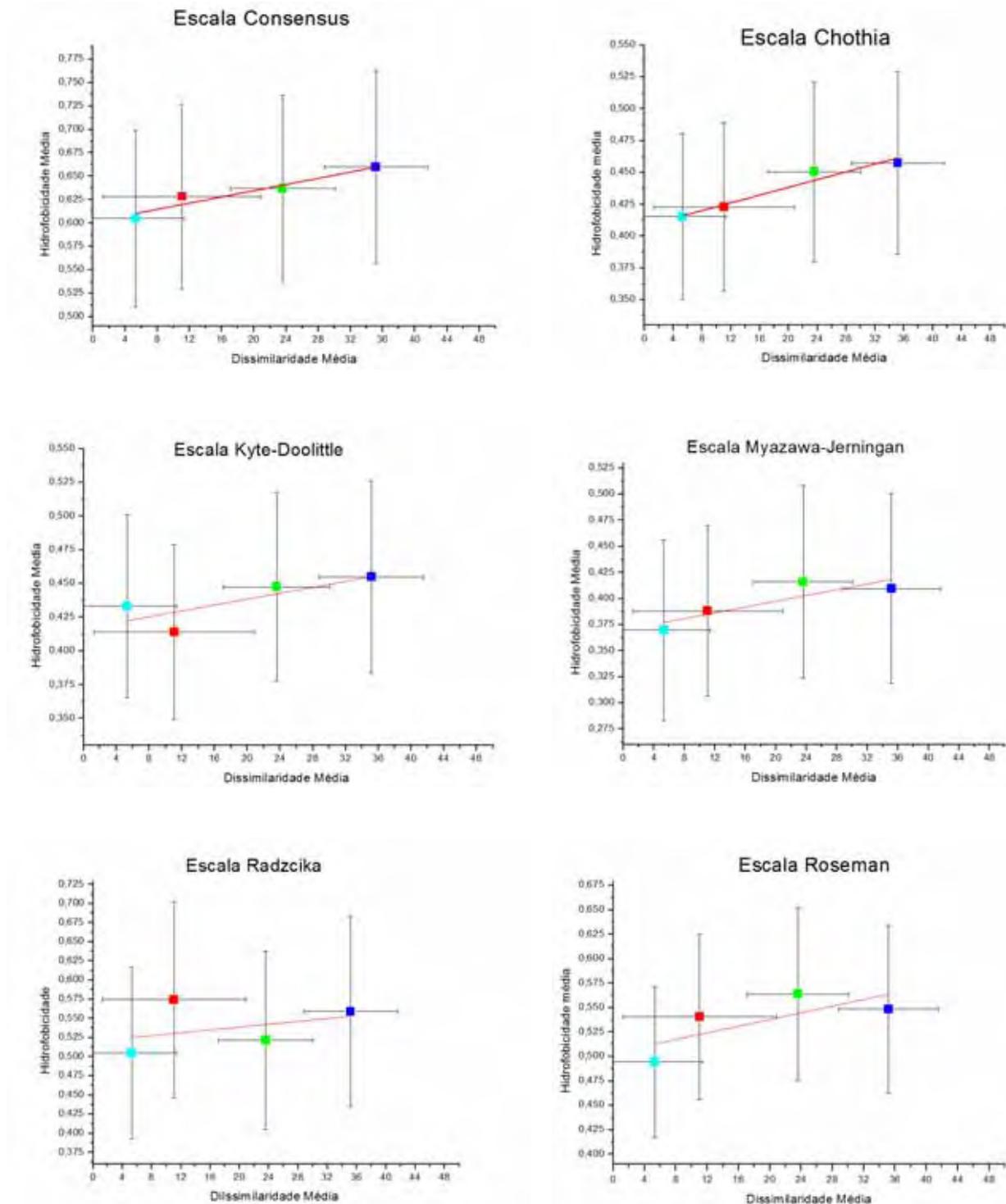


Figura 3.2: Relação entre distância de similaridade(Dissimilaridade) para as classes de proteínas pesquisadas e hidrofobicidade média para seis escalas hidrofóbicas diferentes. As cores ciano, vermelho, verde e azul se referem as grandezas das histonas H3, citocromo C, mioglobina e lisozima respectivamente. Pode se notar um correlação positiva entre os valores médios de hidrofobicidade e dissimilaridade

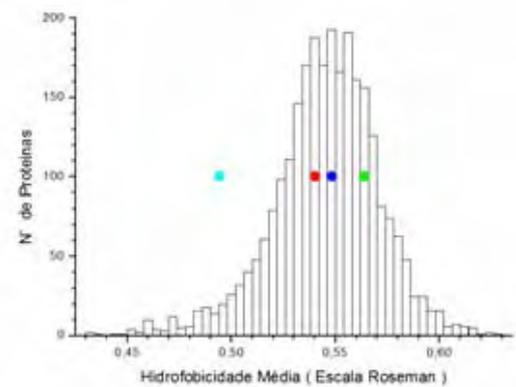
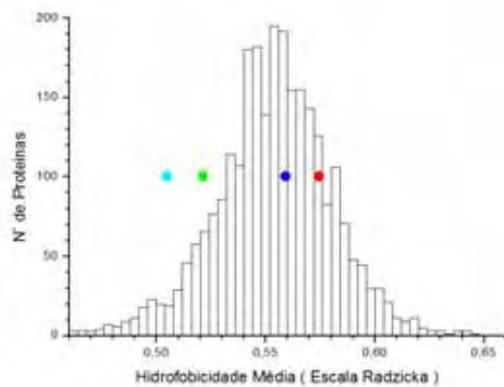
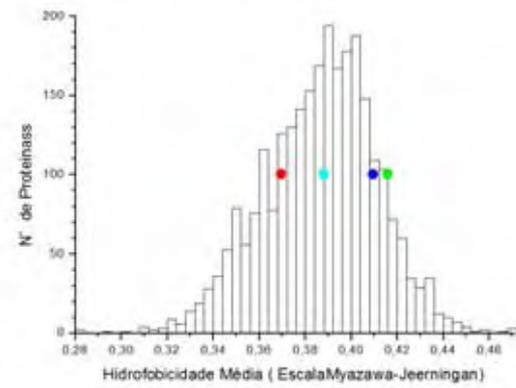
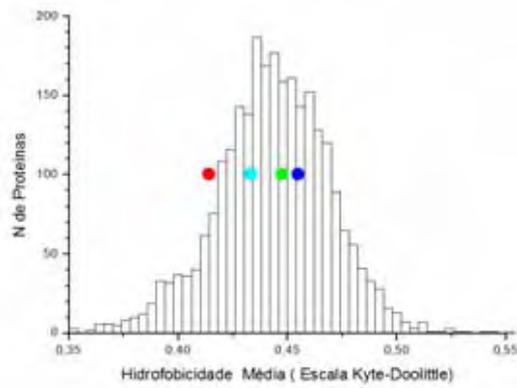
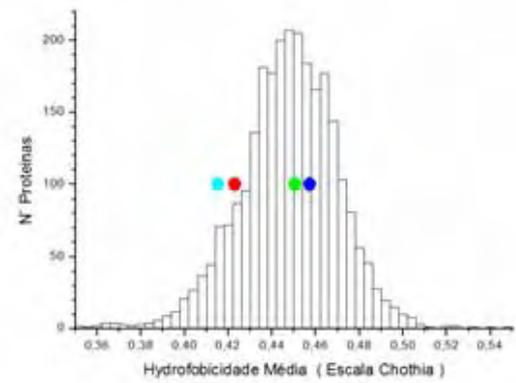
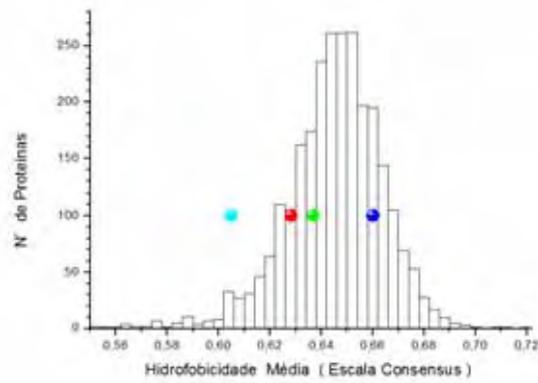


Figura 3.3: Histogramas das hidrofobicidades para as seis escalas utilizadas para 2865 proteínas. Os pontos em todas as escalas se referem aos valores de hidrofobicidade encontradas para as 41 seqüências das quatro classes de proteínas. Os pontos ciano, vermelho, verde e azul representam os valores médios obtidos respectivamente, para a Histona, Citocromo C, Lisozima e  $\epsilon_6$  Mioglobina

através do programa **MAMMOTH-mult**. Este que está disponível no endereço eletrônico <http://ub.cbm.es/mammoth/multi>. Tal programa realiza o cálculo do RMSD das estruturas de proteínas considerando somente as coordenadas dos carbonos alfa das proteínas. A saída, do programa MAMMOTH-multi fornece o RMSD(i,j) entre as estruturas i e j para um conjunto de N estruturas de proteínas.

Assumindo que,

$$DISS(i, j) = RMSD(i, j), \quad (3.5)$$

disimilaridade entre duas estruturas seja igual ao valor do RMSD(i,j) fornecido pelo programa **MAMMOTH-mult**. Submetidas as estruturas encontradas para cada uma das quatro classes de proteínas pesquisadas, foi encontrado como repostas o RMSD(i,j) entre todas as estruturas diferentes possíveis. A partir, dos dados de RMSD fornecido pelo programa **MAMMOTH-mult** foi gerada uma matriz de RMSD(41,41) para cada classe de proteína, de modo que cada matriz representasse as dissimilaridades entre todas as quarenta e uma estruturas encontradas. As matrizes de similaridade de cada classe foram submetidas ao programa SPIN. Este programa realiza a reordenação da matriz de entrada da ordem NxN. A saída do programa SPIN nos fornece uma forma visual, através de uma escala de cores, de se observar a matriz reordenada. O intervalo de cores, representado pela escala de cores, fornece uma referência para se avaliar as regularidades encontradas na matriz reordenada. Para cada uma das quatro matrizes das dissimilaridades o algoritmo neighborhood do programa SPIN foi aplicado. Como resultado, temos os gráficos apresentados na figura 3.4. Ao lado de cada figura, representado as matrizes, está disposta a escalas de cores para a respectiva matriz reordenada. As cores das partes inferiores da barra de cores refletem os menores valores de dissimilaridades entre as estruturas, ao passo que as cores da parte superior da barra de cores representam as maiores valores de dissimilaridades entre as estruturas.

Para análise dos dos resultados do programa SPIN deve se desconsiderar a diagonal principal, pois esta reflete a dissimilaridade  $DISS(i, i)$ . Observando, primeiro o gráfico das dissimilaridades das estruturas das histonas da figura 3.4, nota se a formação de dois quadrados azuis escuros e um pequeno para 28 estruturas. Seguindo a escala de cores estes quadrados refletem pequenos valores de dissimilaridades entre as estruturas das histonas H3, desta forma estes quadrados azuis escuros representam grupos de estruturas quase idênticas. Olhando agora para o gráfico das dissimilaridades dos citocromos c, de acordo com sua barra de cores, nota se a formação de seis quadrados azuis escuros, de tamanhos variados, que refletem a formação de grupos de estruturas quase idênticas. Já, quando observamos o gráfico das dissimilaridades de estruturas para a mioglobina observamos a formação de 12 quadrados azuis escuros, que refletem a formação de grupos de estruturas quase idênticas.

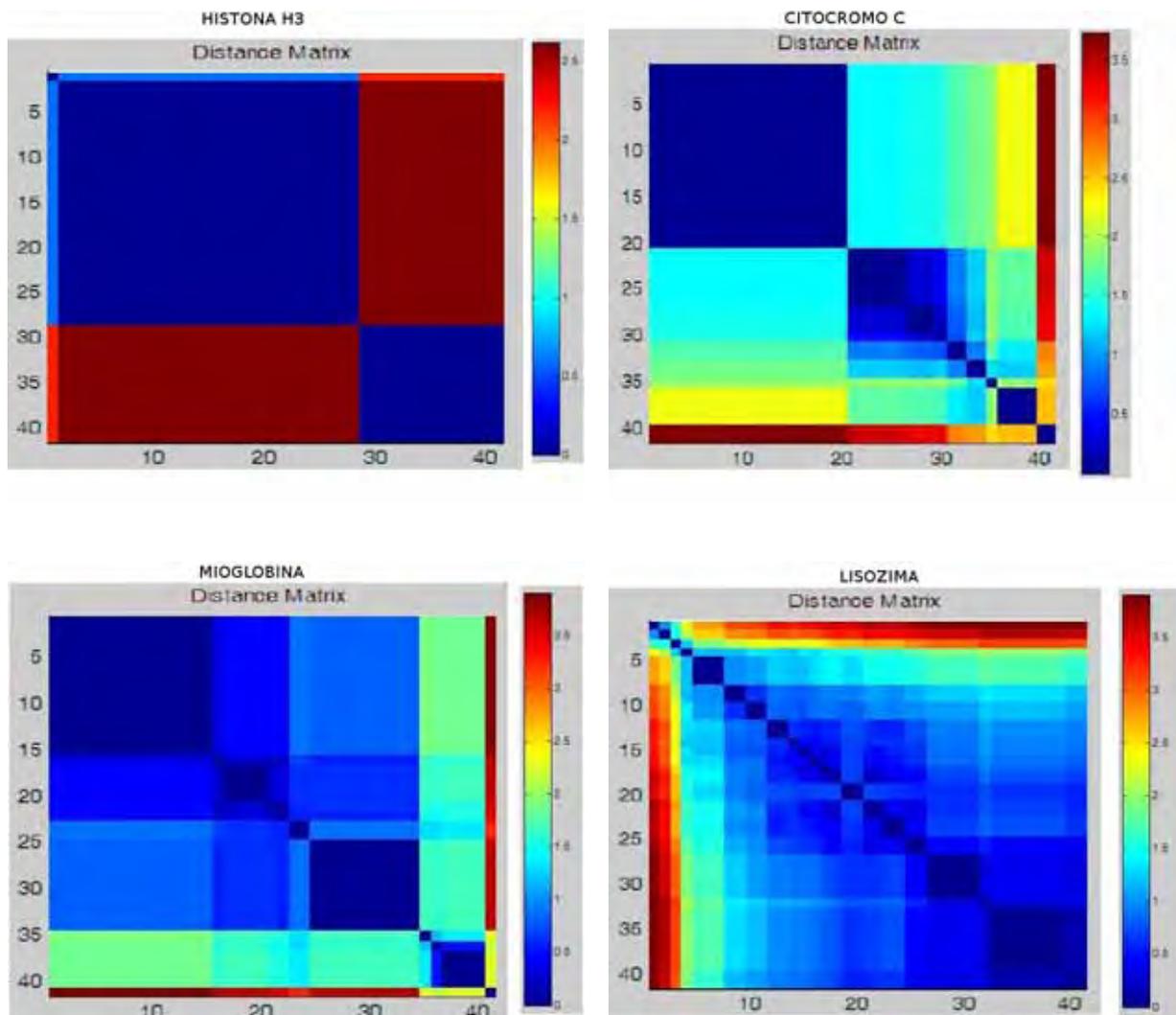


Figura 3.4: Resultado do programa SPIN. Analisando os gráficos, de acordo com as escalas de cores, primeiramente pode se observar a formação de quadrados azuis, que refletem a existência de dois grupos com estruturas quase idênticas. Para o gráfico das dissimilaridades das estruturas dos citocromos C observamos a formação de oito grupos de estruturas quase idênticas. No gráfico das dissimilaridades das estruturas das mioglobinas percebe-se a formação de 11 grupos de estruturas idênticas. Por fim, nota-se a formação de pelo menos 20 grupos de estruturas quase idênticas para a matriz de dissimilaridade das estruturas da Lisozima

Por último verificamos a formação de no mínimo 22 quadrados azuis escuros para o gráfico das dissimilaridades das estruturas das lisozimas. De fato estes dados podem ser resumidos na tabela (). Na tabela () na primeira coluna estão dispostos os dados de número de folds encontrados por classe de proteína e na segunda coluna se refere ao número de sequências que a estrutura melhor projetável possui.

proteína	$N^\circ$ de estruturas	Projetabilidade
Histona	3	28
Citocromo C	6	23
Mioglobina	12	14
Lisozima	22	5

**Tabela 3.3: Tabela que fornece na primeira coluna o número de estruturas diferentes encontradas nos bancos de dados para cada classe de proteína. A segunda se refere ao maior número de sequências que possuem a mesma estrutura terciária. Pode se notar facilmente nesta tabela que o número de estruturas diferentes encontradas varia largamente de proteína para proteína**

Se cruzarmos os dados sobre a formação de grupos de estruturas quase idênticas de proteínas com as hidrofobicidades médias das proteínas podemos observar 3.5 certa correlação entre estas duas grandezas. Esta correlação é expressa mais claramente na escala Consensus, entretanto nas outras escalas pode se observar uma certa correlação positiva entre o número de estruturas diferentes encontradas em cada classe e sua hidrofobicidade média.

Para encerrar nossos resultados sugere que proteínas com baixa hidrofobicidade possuem baixa variabilidade de seqüências e estruturas de proteínas, por outro lado, proteínas com alta hidrofobicidade possuem alta variabilidade de seqüências e estruturas, enfim parece que existe uma regra de número de folds e hidrofobicidade para as famílias de proteínas.

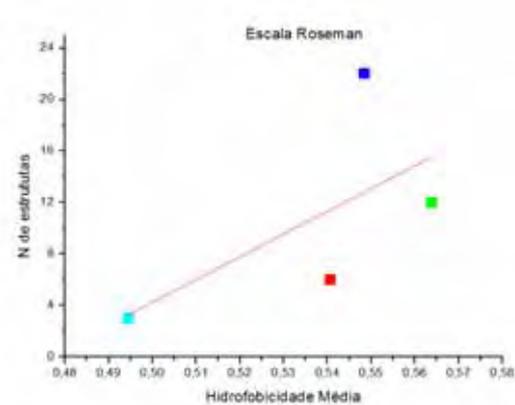
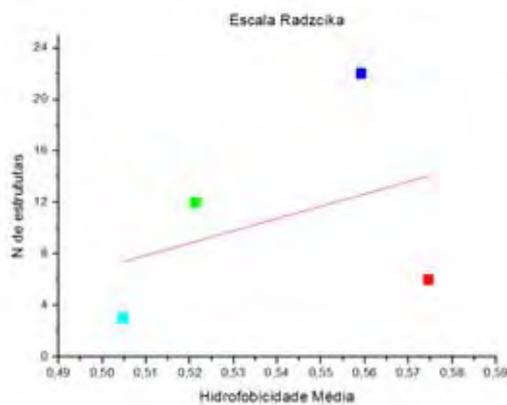
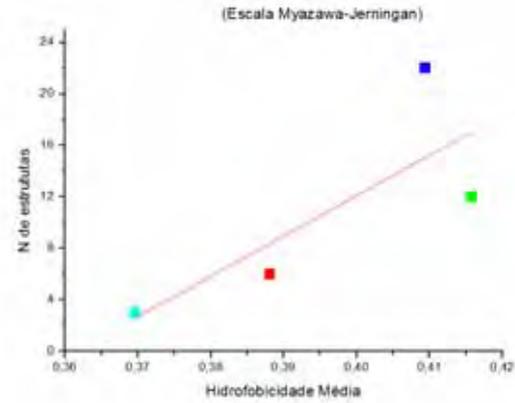
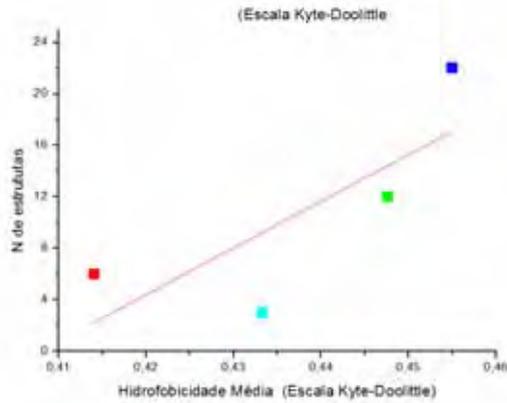
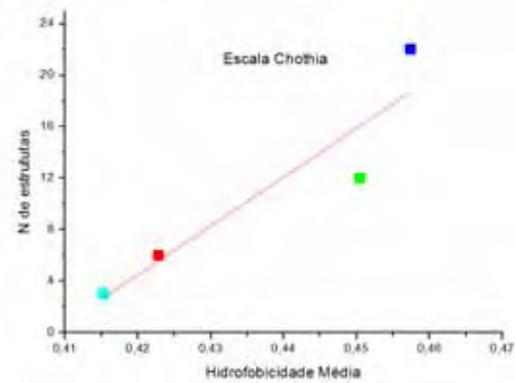
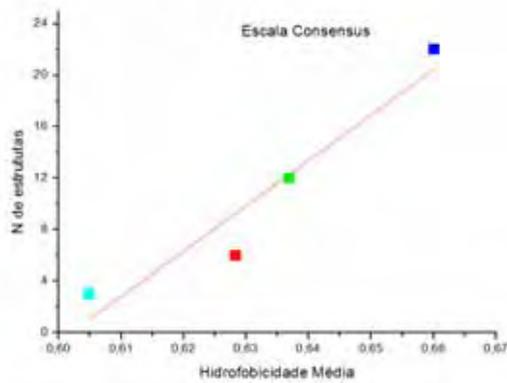


Figura 3.5: Gráficos do número de estruturas vs. a hidrofobicidade para cada classe de proteínas nas escalas hidrofóbicas. Nestes gráficos as cores ciano, vermelho verde e azul, representam respectivamente os dados das histonas, citocromos c, mioglobinas e lisozimas. Observando estes gráficos é fácil notar, que existe uma regra de que quando se aumenta a hidrofobicidade média de uma classe admite-se uma maior uma maior quantidade de estruturas diferentes, ou seja a proteína torna-se mais robusta a mutações

## Capítulo 4

# Discussão e Considerações finais

Sempre se desejou compreender quais mecanismos são preservados na evolução das proteínas. Naturalmente, hoje, através de resultados de simulação computacional e de experimentos, já se sabe que uma proteína deve se enovelar rapidamente em sua estrutura nativa termodinamicamente estável. Além disso, sabe-se que existem dois cenários distintos para o enovelamento de proteínas. No primeiro, a proteína se encontra em alta hidrofobicidade. Neste regime a proteína sofre um rápido colapso seguido de um lento rearranjo estrutural até atingir sua estrutura nativa. No segundo cenário, a proteína está em um regime de baixa hidrofobicidade. Neste regime, a proteína se enovela diretamente em sua estrutura nativa.

Parece natural pensar que os diferentes mecanismos de enovelamento, mencionados no parágrafo anterior, imponham consequências diferentes para a evolução das famílias de proteínas. Nosso trabalho analisou a relação entre hidrofobicidade e dissimilaridade para quatro classes de proteínas. nossos resultados sugerem fortemente deve existir uma regra entre hidrofobicidade e homologia em uma família de proteínas. Foi observado que quanto maior a hidrofobicidade maior será a variabilidade (dissimilaridade) encontrada entre as sequências. Ou de outro modo, quanto menor a hidrofobicidade menor a variabilidade das sequências. Outro resultado interessante observado é que o número de folds ou motifs encontrados varia de acordo com o regime hidrofóbico da classe de proteína. Foi encontrado que quanto menor a hidrofobicidade de uma classe menos folds são encontrados. Na medida que ocorre o aumento da hidrofobicidade ocorre também um aumento no número de folds distintos.

Entretanto, deve se considerar que a hidrofobicidade de classe deve depender do tipo de função que a família de proteína possui. Existem proteínas, como o citocromo c e a histona, que para serem eficientes em sua função devem suportar mais mutações durante a evolução. Por outro lado, existem proteínas para serem eficientes em sua função deve suportar uma grande

quantidade de mutações durante a evolução. Nossos resultados indicam que proteínas com funções que suportam poucas mutações possuem baixa hidrofobicidade e proteínas que sua função pode suportar uma grande quantidade de mutações tem alta hidrofobicidade.

Nosso trabalho ao se comparar os dados de hidrofobicidade para as quatro classes de proteínas com o banco de dados das 2865 sequências não redundantes mostrou-se que as variações que ocorrem no cálculo da hidrofobicidade são significativas e assim pode-se distinguir entre alta e baixa hidrofobicidade. Dessa forma, pode se considerar que alta e baixa hidrofobicidade se referem a mecanismos diferentes de envelhecimento.

Recentemente B. Shakhovich (*i Genome Res.* **15**:385-92) tentou correlacionar a projetabilidade de uma família de proteína com o número de estruturas diferentes calculando a densidade de contatos das proteínas. Ele encontrou famílias pequenas de proteínas com baixa densidade de contato. No entanto encontrou famílias pequenas e grandes com altos valores de densidade de contatos. É nossa hipótese, que o estudo sobre a evolução das proteínas deve considerar as condições hidrofóbicas de cada família de proteína. Pois esta informação pode ajudar a indicar se esta ou aquela família de proteína tem uma maior ou menor variabilidade de sequência e conformação.

Para perspectivas futuras esperamos analisar a correlação entre o tamanho de uma família de proteína com sua hidrofobicidade média e sua variabilidade de sequências e estruturas. Outro ponto que merece atenção é a correlação entre hidrofobicidade estrutura/função.

# Referências Bibliográficas

- [1] D. VOET, J. VOET . **Biochemistry**.1990. New York, Jonh Wilie & Sons.1223 p.
- [2] STRYER L.**Biochimistry**.1988.New York, Frerman, , 1089 p.
- [3] M. DUANE.*Molecular Biophysics*.1999..Oxford.english edition.
- [4] T. E. CREIHHTON.**Proteins:strutures and Molecuar properties**.1995.New York. W.H. Freman and Company. :507p
- [5] D.L. NELSON, M.M. COX.**Princípios de Bioquímica. Sarvier Editora**.2002.São Paulo, ed 3 975 pp.
- [6] L. Pauling, R. B. Corey and H. R. Branson.**The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain***Proc. Natl.Acad. Sci. USA*.1951.**37** 205-211
- [7] J. D. Watson and H. C. Crick.**Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid***Nature*.1953.**171**:737-738
- [8] J.N. ONUCHIC , Z.LUTHEY-SCHULTEN ,P.G. WOLYNES.: **Theory of protein folding: the energy landscape perspective***Ann. RevPhys.Chem*.1997.**48** pp. 545-600.
- [9] C. B. ANFINSEN ,E. HABER ,M. SELA ,F WHITE Jr.*Proc. Natl. Acad. Sci. USA*.1961.**vol. 47** p. 1309
- [10] C. LEVINTHAL.*J. Chim. Phys*.1968.**vol65**:44-45.
- [11] W. KAUZMANN.**Some forces in interpretation of protein denaturation***Adv. Protein Chem*.1959.**vol.16** p. 1-16.
- [12] J. D. Bryngelson,J. N. Onuchic,N. D Socci,P.G.Wolynes.**Funnels, pathways and the energy landscape of protein folding: A synthesis. Proteins: Struct. Funct. Genet**.1995.**21**:167-195

- [13] P. E. Leopold; M. Montal and J. N. Onuchic. **Protein folding funnels: A kinetic approach to the sequence-structure relationship.** *Proc. Natl. Acad. Sci. USA.* 1992. **89**:8721-8725.
- [14] P. E. LEOPOLD and E. I. SHAKHNOVICH. **Protein folding kinetics in the dense phase. System Sciences Proceeding of the Twenty-Sixth Hawaii International Conference.** 1993.
- [15] N.D. SOCCI, J.N. ONUCHIC. 1994 **Time Vs. Temperature, Kinetics Vs. Thermodynamics Using Simple-Models To Explore The Protein Folding Problem** *Biophysical Journal.* 1994. (vo. **166(2)**):A398
- [16] N.D. SOCCI, J.N. ONUCHIC. **Folding Kinetics Of Protein Like Heteropolymer** *J. Chem. Phys.* 1994. (vol. **101(2)**):445-455
- [17] C.L. BROOKS III, M. GRUEBELE, J.N. ONUCHIC, P.G. WOLYNES. **Chemical physics of protein folding.** *Pro. Natl. Acad. Sci. USA.* 1998. **95**:11037-11038.
- [18] SOCCI N.D., ONUCHIC J.N., WOLYNES P.G. **Difusive dynamics of reaction coordinate for protein folding funnels** *J. Chem Phys.* 1996. **104**:5860-5868
- [19] R.HELLING, H. LI, R. MÈLIN., J MILLER ., N.WINGREEN , C. ZENG , C. TANG. **The designability of protein structures.** *Journal of Molecular Graphics and Modelling.* 2001. **19**;p157-167.
- [20] A.G. MURZIN , S.E. BRENNER , T. HUBBARD , C. CHOTHIA **SCOP: A Structural Classification of Protein Database for the investigation of Sequences and Structures** *J. Mol Biol.* 1995. **247**:536-540.
- [21] C.A. ORENGO , A.D. MICHIE, D.T. JONES , M.B. SWINDELLS , J.M. THORNTON **CATH: A Hierarchic Classification of Protein Domain Structures.** *Structure.* 1997. **5**: 1093-1108
- [22] C. CHOTHIA, **Proteins: one thousand families for the molecular biologist** *C. Nature.* 1992. **357**:543-544
- [23] H.S. CHAN, K.A DILL. *Physics Today.* 1993. **FEBRUARY** : 24-32.
- [24] K.A DILL , S. BROMBERG , K. YUE. **Principles of Protein Folding A Perspective from Simple Exact Models** *Protein Science.* 1995. vol. **4**: 561-602.

- [25] H. TAKETOMI, Y. UEDA, N. GO. **Studies on Protein Folding, unfolding and fluctuations by computer simulations.** *Int. J. Peptide Res.* 1975. **7**:445-455.
- [26] Y. UEDA, H. TAKETOMI, N. GO. **Studies on Protein Folding, unfolding and fluctuations by computer simulations II. A Three-Dimensional Lattice Model of Lysozyme.** *Biopolymers.* 1978. **7**:71531-51548.
- [27] H. S. CHAN and K. A. DILL. 1989. **Compact polymers.** *Macromolecules* **22**:4559-4573
- [28] J.N. ONUCHIC, P.G. WOLYNES. **Theory of Protein folding** *Curr. Opin. Struc. Biol.* 2004. **14**:70-75
- [29] A.V. FINKELSTEIN , A.M. GUTUN. **Why are the same protein folds used to performed different functions?** *FEBS Lett.* 1993. **325**:23-28.
- [30] SHAKHNOVICH E.I. **Protein design: a perspective from simple tractable models.** *Fold. Dis.* 1998. **3**:R45-58.
- [31] P.G WOLYNES. **Symetry and energy landscape of biomolecules** *Proc. Nat. Acad Sci. USA.* 1996. **93**:14249-55.
- [32] H. LI , M. HELLING , C. TANG , N. S. WINGREEN. **Emergence of preferred structures in a simple model of protein folding** *Science.* 1996. **49**:666-669.
- [33] H. LI, R. HELING C. TANG, N.S. WINGREEN. **Are protein folds atypical ?** *PNAS USA.* 1998. **95**:4987-4990.
- [34] R. MÉLIN, H. LI, N.S. WINGREEN, C. TANG. **Designability, thermodynamics stability and dynamics in protein folding; a lattice model study** *J. Chem. Phys.* 1999. **110**:1252-1262.
- [35] H. S. CHAN and K. A. DILL. **Compact polymers.** *Macromolecules* 1989. **22**:4559-4573
- [36] J.N. ONUCHIC, P.G. WOLYNES. **Theory of Protein folding** *Curr. Opin. Struc. Biol.* 2004. **14**:70-75
- [37] Finkelstein A. V., GUTUN A.M. 1993. **Why are the same protein folds used to performed different functions?** *FEBS Lett.* **325**:23-28.
- [38] E.I. SHAKHNOVICH. **Protein design: a perspective from simple tractable models.** *Fold. Dis.* 1998. **3**:R45-58.

- [39] K.B. ZELDOVICH , E.I SHAKHNOVICH. **Understanding Protein Evolution: From protein physics to Darwinian Selection** *Annu. Rev. Phys. Chem.*.2008.**59**:105-157.
- [40] J.L. ENGLAND ,E.I. SHAKHNOVICH. **Structural determinant of protein designability.** *Phys. Rev. Lett.*.2003.**90**:218101.
- [41] A. SALI ,E.I. SHAKNOVICH, M. KARPLUS. **Knetics of protein folding. A lattice model study of a requerements for folding to the native state** *J.Mol. Bio.*1994.**235** pp. 1614-1636.
- [42] T.S. CHIU ,R.A. GOLSTEIN **Compaction and Folding in Model Proteins** *J.Chem Phys.*1997.**107**:4408
- [43] I.S. MILLET,D.J. SEGEL ,S. DONIACH, D.BAKER ,K.W. PLAXO **Exploring Protein Structure and Dynamics under Denaturing Conditions by Single-Molecule FRET Analysis** *Nat Struct Biol.*1999.**6**:554
- [44] Klimov D.K.,Thirumalai, D. : **A criterion that determines the foldability of proteins.**1996. *Phys. Rev. Lett.* 76:4070-4073
- [45] D.THIRUMALAI,D.K. KLIMOV. **Deciphering the time scales and mechanisms of protein folding using minimal off-lattice models.**1999. *Curr. Opin. Struct. Biol***9**: 197-207.
- [46] A.M. GUTIN,V.I. ABKEVICH, E.I. SHAKNOVICH. **Is burst hydrophobic collapse necessary for protein folding?** *Biochemistry.*1995.:3066-3076.
- [47] SHAKHNOVICH BE,DEEDS E,DELISI C, SHAKHNOVICH E. **Protein structure and evolutionary history determine sequence space topology** *Genome Res* .2005. **15**:385-92.
- [48] B.G. MIRKIN, T.I. FENNER, M.Y. GALPERIN,E.V. KOONIN. **Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes.** *BMC Evol. Biol* .2003.**3**:2
- [49] J.D. BLOOM ,D.A. DRUMOND ,F.H. ARNOLD,C.O. WILKE. **Structuraldeterminants of the rate of protein evolution in yeast.** *Mol.Biol. Evol.*2006.**23**:1751-1761.
- [50] L.C. OLIVEIRA ,R.H.T. SILVA,V.B.P. LEITE ,J. CHAHIENE. **Frustration and hydrofobicity interplay in protein folding** *J. Chem. Phys.*2008. (125):1-6.

- [51] RENNELL D., BOUVIER S.E., HARDY L.W., POTEETE A.R. **Systematic mutation of bacteriophage T4 lysozyme.** *j. Biol. Mol.* 1991. **222**:67
- [52] J. SONDEK, D. SHORTLE. **Structural and energetic differences between insertions and substitutions in staphylococcal nuclease.** *Proteins.* 1992. **13**:132
- [53] L.H. Weaver, M.G. GRUTTER, S.E. REMINGTON, T.M. GRAY, N.W. ISACS, B.W. MATTHEWS. *J. Mol. Evol.* 1985. **21**:97
- [54] KEIFHABER T. **Kinetics traps in lysozyme.** *Proc. Natl. Acad. Sci USA.* 1995. *92*:9029-9033.
- [55] M. MUCKE, F.X. SMCHMID **A protein folding intermediate of ribonuclease T1 characterized at high resolution by 1D and 2D real-time NMR spectroscopy** *J. Mol. Biol.* 1994. **239**:713
- [56] KHORASANIZADEH S., PETERS I.D., BUTT T.R., RODER H. 1993. *Biochemistry* **31**:7054.
- [57] GILLESPIE B., PLAXO K.W. *Proc. Natl. Acad. Sci USA.* 2000. **96**:12014.
- [58] .XIONG, J. **Essential Bioinformatics Cambridge University Press.** 2006. New York:362p.
- [59] .ATWOOD T.K., PARRY-SMITH, D.J. **Introduction to Bioinformatics Cambridge University Press.** 2006. New York:362p.
- [60] BAXEVANIS, A.D., OULLETTE, B.F. **Bioinformatics** *Jonh Wiley and Sons.* 2001. New York:489p
- [61] M.O. DAYHOFF, R. SCHWARTZ, B.C. ORCUTT. **A model of change in proteins** *Atlas of Protein Sequence and Structure.* 1978. **vol. 5**:345-352.
- [62] S. HENIKOFF, J.G. HENICOFF. **Automated assembly of protein blocks for database searching.** *Nucl. Acids. Res.* 1991. **19**:6565-6572.
- [63] S. HENIKOFF, J. G. HENICOFF. **Amino acid substitution matrices from protein blocks.** *Proc. Nat. Acad. Sci. USA.* 1992. **89**:10915-10919.
- [64] S. HENIKOFF S., J.G. HENICOFF. **Performace evaluation of amino acid substitution** *Proc. Nat. Acad. Sci. USA.* 1993. **89**:10915-10919.
- [65] A.K. JAIN, R.C. DUBES. **Algorithms for Clustering Data** *Prentice Hall* 1988.

- [66] P. BERKIHIN. **Survey of clustering data mining techniques**. Technical report  
*itAccrue Software, San Jose, CA.*(2002)
- [67] A. K. JAIN, M. N. MURTY, P. J. FLYNN. **Data clustering: a review**. *ACM Computing Surveys*.1999.**31(3)**:264-323
- [68] J. SANDER J., X QIN X., Z. LU, N. NIU, A. KOVASSKY. **Automatic extraction of clusters from hierarchical clustering representations**. In *PAKDD Pacific-Asia Knowledge Discovery and Data Mining of LNAI Springer-Verlag*.2003.**2637**:75-87
- [69] D. TSAFRIR et al. **Data analysis and Visualization by Ordering Distance Matrix**  
*Bioinformatics*.2005.**21**: 2301-2308.
- [70] T. KOOPMANS, M BECKMAN. **Assignment problems and location of economic activities**.*Econometrica*.1957.**27**:57-76.
- [71] E. A. DINIC, M. A. KRONROD. **An algorithm for the solution of assignment problem***Sovit. Math. Dokl.*1969.**10**:1324-1326
- [72] G. GALLI **Dissecting hydrophobicity***PNAS*.2007.**104(8)**:2257-2258.
- [73] D. CHANDLER **Interfaces and driving force of hydrophobic assembly***Nature*.2005.**vol.437**:640-646
- [74] M. CHARTON , B.I. CHARTON **The Structural Dependence of Amino Acid Hydrophobicity Parameters***J. Theor. Biol.*1982.**99**:629-644.
- [75] ROSE G.D., WOLFENDEN R.1993**Hydrogen bonding, hydrophobicity, packing, and protein folding***Annu.Rev. Biomol. Struct.***22**:381-415.
- [76] K.M. BISWASS, D.R. DEVIDO, J.G. DORSEY *Chromatography A*.2003.**1000**:637-655.
- [77] Y. NOZAKI, C. TANFORD. **The solubility of amino acids and related compounds in aqueous urea solutions** *J. Biol. Chem.*1971.**246**:2211.
- [78] J.H. FENDLER, F. NOME , J. NAGYVARY. **Compartmentalization of amino acids in surfactant aggregates. Partitioning between water and aqueous micellar sodium dodecanoate and between hexane and dodecylammonium propionate trapped water in hexane.** *J. Mol. Evol.*1975. **6**:215-232.
- [79] R . WOLFENDEN, L. ANDERSSON, P.M. CILLIS, C.C.B SOUTHGATE. **Affinities of amino acid side chains for solvent water** *Biochemistry*.1981.**20**:849-855.

- [80] B.Y. ZASLAVSKY, N.M. MESTECHKINA, L.M. MIHEEVA, S.V. ROGOZHIN. *J. Chromatogr.* 1982. **240**:21.
- [81] RADIZICKA A., WOLFENDEN R. **Comparing the Polarities of Aminoacids: Side-Chain Distribution Coefficients between the Vapor Phase, Cyclohexane, 1-Octanol, and Neutral Aqueous Solution.** 1988. *Biochemistry* **27**:1664-1670.
- [82] KYTE J., DOOLITTLE R.F. **A Simple Method for Displaying the Hydrophobic Character of a Protein.** *J. Mol. Biol.* 1981. **157**:105-132
- [83] ROSEMAN M. A. 1988. **Hydrophilicity of Polar Amino Acid Side Chain is Markedly Reduced by Flanking Peptide Bonds.** *J. Mol. Biol.* **200**:513-522.
- [84] CHOHIA C. **The Nature of Accessible and Buried Surfaces in Proteins.** 1976. *J. Mol. Biol.* **105**:1-14.
- [85] R.S. HODGES, B.-Y. SHU, N.E. ZHOU, C.T. MANT, **Reversed-phase liquid chromatography as a useful probe of hydrophobic interactions involved in protein folding and protein stability** 1994. *J. Chromatogr A* **676**:3-15.
- [86] A.A. ABODERIN **An empirical hydrophobicity scale for  $\alpha$ -amino-acids and some of its application** *Int. J. Biochem.* 1971. **2**:537.
- [87] V. PLISKA, M. SCHMIDT, J.L. FAUCHERE *J. Chromatogr.* 1982. **216**:79-92.
- [88] PLASS M., VALKO K., ABRAHAM M.H. **Determination of solute descriptors of tripeptide derivatives based on high-throughput gradient high-performance liquid chromatography retention data** *J. Chromatogr. A* .1998. **803**: 51-60.
- [89] YUTANI K, OGASAHARA K., TSUJITA T., SUGINO Y. **Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit.** *Proc. Natl. Acad. Sci. USA.* 1987. **84**: 4441-4444. bibitem BULL74 BULL H .B, BREESE K. **Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues** *Arch. Biochem. Biophys.* 1974. **161**:665
- [90] D. EKISENBERG, E. SCHWARZ, M. KOMAROMY, R. WALL, **Analysis of membrane and surface protein sequences with the hydrophobic moment plot** *J Mol Biol.* 1984. **179**: 125-142.
- [91] C. CHOHIA 1976 **The Nature of Accessible and Buried Surfaces in Proteins.** *J. Mol. Biol.* **105**:1-14.

- [92] S. Myiazawa, R. Jerningan **Estimation of Effetive Interresidue Contact Energies from protein Cristal Structure:Quasi-Chemical Aproximation.***Macromolecules.*1985.**18**:534-552.

## Capítulo 5

## Apêndice A

Citocromo c	Lisozima	Mioglobina	Histona H3	Organismo
P99999	P61626	P02144	P84243	Human
P99998	P61627	P02145	XP5188	Chimpanzee
P00002	P79239	P02148	Q5RCC9	Orangutan
Q5RFH4	P30201	P02149	A5A612	Reshus
Q6WUX8	P79179	P02147	B5FW76	Gorilla
Q52V10	P79294	P02155	A9X1D9	Squirrel Monkey
Q7YR71	P67979	P02150	A9RA98	Leaf Monkey
P00004	P11376	P68082	NP0010	Horse
P68097	P11375	P02178	186433	Donkey
P00011	P37714	P68083	UPI000	Zebra
Q37430	P81708	P63113	721930	Dog
P68098	P37713	A6N8S4	Q5E9F8	Goat
P62896	O97723	P02160	Q71LE2	Guanaco
P62894	P17607	P02190	BOLRN3	Sheep
P00008	Q27996	P02196	P68432	Cow
P62895	P16973	P02170	70749P	Rabbit
P68099	P12067	P02189	CAA361	Pig
483111	P37712	Q2MJN4	229423	Dromedary
P62897	P00713	P04249	Q64400	Guinea Pig
P68096	P00697	Q9QZ76	276613	Rat
Q56A15	P17897	P04247	484530	Mouse
P00012	Q659U5	P02191	Q4S4R3	Chital
P00022	Q7LZQ1	P68080	B3FVA6	SEAL
P00014	P51782	P02202	Q6QN07	Chinese turtle
P00007	A5HKM9	P02194	A5PK61	Kangaroo
P00017	P00699	P84997	B5FZ56	Hippopotamus
B5FXZ1	P24364	P02199	B5G4Q1	Penguin
P67881	P00698	Q7LZM5	P84247	Night hawk
P67882	P00703	Q7LZM5	P84247	Pheophila
849951	P00700	Q7LZM2	P02297	Turkey
P00019	P00704	Q7LZM4	P84229	Guinea Fowl
P00021	Q7LZQ2	P02196	A4UU65	Duck Bill
496339	P30805	Q7LZM1	P84230	Himalalian Pheasant
AAA487	Q7LZP9	P02197	Q92068	Ring Naked Pheasant
P00024	P00702	A6MHQ6	70755P	Westrn Clowed Frog
Q64OU4	Q5M8GO	Q9DEP1	P84232	Carpa
P00026	Q9IBG5	Q7T044	HSTR31	Trout
ACH706	P11941	Q9DGJ0	B5DG71	Salmo Salar
Q4SG99	P61944	Q701N9 92	Q4S9GO	Green Puffer
Q6IQM2	Q90YS5	Q6VN46	Q6PI20	Zebra Fish