CAIO JÚLIO MARTINS VELOSO

DESVELANDO O SISTEMA DE REDES DAS ESTRUTURAS PROTÉICAS
UMA CONTRIBUIÇÃO PARA UM MODELO FORMAL DE PROTEÍNAS

Belo Horizonte
18 de dezembro de 2007

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Bioinformática

DESVELANDO O SISTEMA DE REDES DAS ESTRUTURAS PROTÉICAS UMA CONTRIBUIÇÃO PARA UM MODELO FORMAL DE PROTEÍNAS

Projeto de tese apresentado ao Curso de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

CAIO JÚLIO MARTINS VELOSO

Belo Horizonte
18 de dezembro de 2007

Resumo

Tradicionalmente as proteínas têm sido vistas como um constructo baseado em elementos de estruturas secundárias e em seus arranjos no espaço tridimensional. Procurando ir além desta perspectiva, este trabalho propõe uma abordagem mais abrangente das proteínas, onde estas são vistas como sistemas complexos. Como prova de conceito, duas famílias distintas de proteínas são modeladas como redes de interações não covalentes entre átomos. Tais redes são analisadas buscando identificar indícios de que o padrão estrutural destas proteínas seja análogo àqueles comungados por outros sistemas complexos observáveis no mundo real, e estudado em outros trabalhos. O conceito fenomenológico das redes complexas é aplicado às estruturas tridimensionais das proteínas revelando como os resíduos de aminoácidos podem ser conectados uns aos outros produzindo uma rede de propriedades próprias impossíveis de serem totalmente previstas *a priori*. Os resultados sugerem que as proteínas devam apresentar propriedades emergentes únicas. O reconhecimento de que as proteínas podem ser modeladas e estudadas como sistemas complexos aparece como um paradigma promissor para o entendimento da robustez destas moléculas em seu estado funcional, mesmo em face as mutações aleatórias e sugere uma via alternativa para a investigação dos determinantes estruturais, dinâmicos e de regulação destas moléculas.

Abstract

Traditionally proteins have been seen as a construct based on elements of secondary structures and their arrangements in three-dimensional space. Going beyond this perspective, this work suggests a more comprehensive approach of proteins, where they may seen as complex systems. As proof of concept, two families of different proteins were modeled as networks of non-covalent interactions among atoms. Such networks were analyzed aiming to identify signs that the structural patterns of proteins are similar to those shared by other complex systems observable in the real world, and studied in other works. The phenomenological concept of complex networks is applied to three-dimensional structures of proteins showing how amino acids residues may be connected to each other producing a network which properties are impossible to be planned *a priori*. The results suggest that the protein should present unique emerging properties. The recognition that the proteins can be modeled and studied as complex systems arises as a promising paradigm for understanding the strength of these molecules in its functional state, even in face random mutations and suggests an alternative route for investigating the structural, dynamic and regulation determinants of these molecules.

"... questo grandissimo libro [della natura scritto da Dio] che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto ..."

Galileo Galilei,
Il Saggiatore

Agradecimentos

Muitos há a quem devo meus agradecimentos. Sinto me em débito com o amigo Carlos Henrique da Silveira pelas conversas e discussões inspiradoras relativas aos problemas de estruturas de proteínas, construção dos modelos de redes dentre outras. Tenho muito a agradecer ao meus orientadores e amigos Prof. Wagner Meira Jr. e Prof. Marcelo Matos Santoro pela paciência, pela amizade, pelas sugestões auspiciosas e julgamento crítico ao longo de todo este tempo. Agradeço também ao amigo Prof. Marcos Augusto dos Santos pelas discussões relativas às ferramentas de análise espectral. Agradeço ainda ao Prof. Mohammed Javeed Zaki, à Profa. Gloria Regina Franco, aos colegas Sérgio Nogueira Júnior, Cristina Ribeiro, Raquel Cardoso de Melo pelo apoio e incentivo. Agradeço à Pontificia Universidade Católica de Minas Gerais a concessão da bolsa de apoio através do Programa Permanente de Capacitação Docente.

Contudo, agradeço principalmente a Octávio, Isabella e Liz, meus filhos, e à Leda, minha mulher, pelo sacrifício que estes anos de trabalho representaram para eles e como ainda assim eles me apoiaram neste esforço. Por fim, agradeço a Deus por ter me encaminhado nesta direção, por ter colocado estes valiosos amigos em meu caminho e por ter me dado inspiração e suporte para perseguir este objetivo.

Sumário

1 Introdução **1**

2 Proteínas como Sistemas **10**

2.1 Os Princípios da Teoria Geral dos Sistemas 11

2.1.1 Conectividade e Interdependência 14

2.1.2 Co-evolução 14

2.1.3 Sistemas dinâmicos 15

2.1.4 Exploração do Espaço de Possibilidades 17

2.1.5 Realimentação 18

2.1.6 Causalidade 19

2.1.7 Auto-Organização 20

2.1.8 Abordando a Estrutura dos Sistemas Complexos 21

2.2 Proteínas como Sistemas Complexos 22

3 Redes Complexas **27**

3.1 Conceitos e Modelos de Redes Complexas 30

3.1.1 Redes Regulares 31

3.1.2 Redes Randômicas 32

3.1.3 Redes "*small-world*" 33

3.1.4 Redes Livres de Escala 36

3.2 Pontos Fracos das Redes Complexas 38

3.3 Métricas Básicas 40

3.3.1 Distâncias entre Nodos 40

3.3.2 Coeficiente de Aglomeração 42

3.3.3 Distribuição dos Graus de Conectividade 43

3.3.4 Vulnerabilidade da Rede 44

3.4 Redes Complexas Ponderadas 45

3.4.1 Conceito de peso em redes ponderadas 45

3.4.2 Conectividade em redes ponderadas : *Strength* 46

3.5 Análise Espectral de Grafos 47

3.5.1 Decomposição de Valores Singulares 48

3.5.2 Matrizes relacionadas às redes 51

3.5.3 Análise Espectral da Matriz de Adjacências de um Grafo 51
3.5.4 Distribuição dos Valores Espectrais 52

4 Materiais e Métodos **59**
4.1 Proteínas em estudo 59
4.2 Adoção do modelo de redes na análise da estrutura das proteínas 61
4.3 Representação Gráfica da Estrutura das Proteínas 63
4.4 Cálculo de Oclusão entre Átomos 65
4.4.1 Introdução 65
4.4.2 A Necessidade de Solvatação 67
4.4.3 O Problema da Oclusão 67
4.4.4 Análise das RGPSs 81
4.4.5 Perfil de distribuição das arestas dos átomos e resíduos de aminoácidos 81
4.5 Análise do Ganho de Informação 82

5 Resultados e Discussão **84**
5.1 Comparação entre os resultados obtidos com e sem o uso do método de oclusão entre átomos 84
5.2 Análise Estatística das Estruturas das Proteínas 88
5.2.1 Perfis de Distribuição das Interações Atômicas 88
5.2.2 Índices de Transitividade 98
5.3 Análise Espectral das Estruturas das Proteínas 117
5.4 Identificação dos Nodos Concentradores 128
5.4.1 Identificação dos Resíduos mais conectados nas globinas 128

6 Conclusões e Perspectivas **151**

Referências Bibliográficas **157**

Lista de Figuras

2.1 Bifurcação 16

3.1 Exemplo de modelos topológicos de redes complexas apresentados na literatura 30
3.2 Estruturas regulares com diferentes valores de interação com sua vizinhança . . . 31
3.3 Evolução de um grafo randômico 32
3.4 Rede regular, "*small-world*" e Randômica 34
3.5 Coeficiente de aglomeração C e distância média entre vértices l no modelo WS . 35
3.6 Coeficiente de aglomeração C e distância média entre vértices l no modelo "*small-world*" 35
3.7 Padrões de "*rewiring*" nos modelos "*small-world*" 36
3.8 Exemplos de distribuições associadas às redes do mundo real 37
3.9 Evolução para uma rede livre em escala 38
3.10 Efeito da remoção dos nodos em uma rede inicialmente conectada 39
3.11 Tamanho relativo do maior algomerado na Internet quando uma fração dos domínios é removida 40
3.12 Espectro de cores da luz 49
3.13 Esquema da aplicação da SVD em uma matriz $[M]$. 50
3.14 Função $\delta_\epsilon(x)$ 53
3.15 Distribuição de densidade - $\rho(\lambda)$ 54

4.1 Níveis de vizinhanças em redes 62
4.2 Interações de longa distância na estrutura de uma proteína 63
4.3 Um átomo qualquer considerado como uma partícula carregada. 68
4.4 Linhas de força de duas cargas de sinais contrários em interação próxima. 69
4.5 Esquema mostrando a oclusão entre átomos 70
4.6 Esquema mostrando a oclusão do átomo A_j pelo átomo A_k como " visto" por A_i. 71
4.7 Comparação das áreas expostas e oclusas em funções da distância entre átomos . 72
4.8 Sobreposição das projeções dos átomos k_1 e k_2 sobre a área visível do átomo j . 73
4.9 Projeções de dois átomos k_1 e k_2 sobre átomo j mapeadas sobre uma matriz. . . 74
4.10 Situações possíveis quanto à posição relativa entre os átomos A_j e A_k. 77
4.11 Situação de sobreposição parcial entre os átomos A_j e A_k. 77
4.12 Área retangular que contém a área de A_j ofuscada por A_k. 79

5.1 Freqüências das distâncias para ligações oclusas e não oclusas 85
5.2 Freqüências das distâncias para ligações oclusas e não oclusas 86
5.3 Curva de tendência das distâncias entre átomos para ligações oclusas e não oclusas 87
5.4 Números de contatos por átomo sem critério de oclusão para globinas 89
5.5 Números de contatos por átomo sem critério de oclusão para serinoproteases . . 90
5.6 Distribuições de $f(N_{CA})$ e N_{CA} para as globinas solvatadas e com oclusão 91
5.7 Distribuições de $f(N_{CA})$ e N_{CA} para as serinoproteases solvatadas e com oclusão 92
5.8 Número de coordenação médio - N_c, versus densidade do arranjo de partículas - ρ 93
5.9 Relacionamento entre a média das energias de interação e o número de interações por átomo 95
5.10 Distribuições, para as globinas, da freqüência de níveis de energia por átomo $f(E_A)$ 96
5.11 Distribuições, para as serinoproteases, da freqüência de níveis de energia por átomo $f(E_A)$ 97
5.12 Distribuições dos índices de aglomeração por resíduo na estrutura primária para globinas 99
5.13 Distribuições dos índices de distância média entre átomos para as globinas 102
5.14 Distribuições de distância média entre átomos para globinas monoméricas 107
5.15 Distribuições de distância média entre átomos para globinas multiméricas 108
5.16 Distribuição do número de átomos presentes dentro de um raio r para as globinas 109
5.17 Distribuição da distância média entre átomos dentro de um raio r para as globinas 110
5.18 Distribuição dos índices de clusterização por resíduo para as serinoproteases . . . 111
5.19 Distribuição da distância média entre átomos, por resíduo, para as serinoproteases 114
5.20 Distribuição do número de átomos presentes dentro de um raio r para as serinoproteases 115
5.21 Distribuição da distância média entre átomos dentro de um raio r para as serinoproteases 116
5.22 Distribuição de densidade espectral de três modelos de rede 118
5.23 Distribuição de densidade de três modelos de rede com os parâmetros das globinas 119
5.24 Distribuição de densidade de três modelos de rede com os parâmetros das serinoproteases 120
5.25 Distribuição de freqüências da média dos valores singulares para as globinas . . . 121
5.26 Distribuição de freqüências da média dos valores singulares para as serinoproteases 121
5.27 Distribuição dos valores singulares com base nas interações não-covalentes 123
5.28 Distribuição espectral das interações ponderadas pelas energias potenciais 126
5.29 Resíduos pertencentes ao grupo-núcleo das globinas estudadas 127
5.30 Resíduos pertencentes ao grupo-núcleo das serinoproteases estudadas 127
5.31 Distribuições dos níveis de energia por átomo para as proteínas solvatadas com oclusão 129
5.32 Distribuições dos níveis de energia dos resíduos distantes 130
5.33 Distribuições $f(E_A)$ e E_A dos resíduos distantes 131
5.34 Alinhamento estrutural das globinas 132

5.35 Resíduos "*hubs*" em regiões espacialmente alinhadas nas globinas 135
5.36 Mapeamento dos resíduos "*hubs*" destacado o grupo **HEME** 136
5.37 Mapeamento dos resíduos "*hubs*" sobre a globina 1A6G 136
5.38 Resíduos "*hub*" na estrutura 3D da mioglobina 1A6G 137
5.39 Resíduos "*hub*" presentes no **locus 1** da mioglobina 1A6G 138
5.40 Representação do sítio de ligação de uma globina 138
5.41 Vistas do **locus 2** no alinhamento estrutural conjunto das globinas 139
5.42 Vistas de topo (b) e frente (a) do **locus 2** 140
5.43 Vista do **locus 3** 141
5.44 Vista do **locus 4** 142
5.45 Sobreposição de todas as águas de solvatação 143
5.46 Sobreposição de todas as águas de solvatação sobre o núcleo hidrofóbico 144
5.47 Distribuição das águas de solvatação para a mioglobina PDBID 1A6G 144
5.48 Distribuição das águas de solvatação sobre o núcleo hidrofóbico da mioglobina 1A6G145
5.49 Resíduos "*hub*" do **locus 1** mantidos próximos por águas de solvatação 145
5.50 Quadro de 159 mutantes da mioglobina PDBID 1A6G 149

Lista de Tabelas

5.1 Parâmetros representativos distribuições $f(N_{CA}) \times N_{CA}$ 94
5.2 Parâmetros representativos distribuições $f(E_A) \times E_A$ 98
5.3 Coeficiente médio de aglomeração para as globinas 100
5.4 *Alguns coeficientes para redes biológicas já estudadas.* 101
5.5 Coeficiente médio de distância entre átomos para as globinas 103
5.6 globinas do aglomerado"1-5" 104
5.7 globinas do aglomerado"1-25" 105
5.8 Coeficiente médio de aglomeração médio para as serinoproteases em estudo. . . 111
5.9 Coeficiente médio de distância entre átomos para as serinoproteases em estudo . 112
5.10 Seqüências resultantes do alinhamento estrutural das globinas 132
5.11 Posições mais freqüentes responsáveis pela estabilidade do **locus 2**. 140
5.12 Posições mais freqüentes responsáveis pela estabilidade do **locus 3**. 141
5.13 Posições mais freqüentes responsáveis pela estabilidade do **locus 4**. 142
5.14 Posições "*hub*", identificadas no alinhamento estrutural das globinas 146

Capítulo 1

Introdução

"*... Lasciate ogne speranza, voi ch'intrate.*"
Dante Alighieri,
La Divina Commedia - Inferno, Canto III, 9

CONCEPÇÃO do Mito da Caverna narrada por Platão no livro VII de *A Republica* é talvez, uma das mais poderosas metáforas já imaginadas para descrever a situação geral em que se encontra a humanidade e as suas ciências. Para Platão, a humanidade está em uma condição onde cada um de nós, ao vermos "sombras" à nossa frente, irá inexoravelmente tomá-las como sendo a expressão da "verdade". Esta crítica, feita há quase 2.500 anos, mantém-se apropriada e inspira ainda inúmeras reflexões, notadamente quando nos deparamos com a produção do conhecimento científico.

Tal como no "Mito da Caverna", a percepção subjetiva da "realidade", ou "visão de mundo"[1], que trazemos conosco é sempre fragmentada e sectária, tendo sido construída à medida que passamos pelos processos formais que atribuem legitimidade às opiniões que emitimos e trabalhos que realizamos no âmbito da comunidade tecno-científica em que habitualmente atuamos [Durkheim (2003)]. O termo "paradigma" foi usado por Kuhn [Kuhn (1962)], para referenciar tais estruturas e/ou concepções mentais do mundo, adotadas por comunidades científicas ou em outros contextos epistemológicos.

Paradoxalmente, os mesmos paradigmas, consolidados com os anos de práticas e experiências e, que auxiliam os estudiosos de diferentes áreas do conhecimento em suas realizações, podem ser, também, grandes fardos para aqueles que desejam se aventurar para

[1] O conceito de 'visão de mundo', originário da antropologia cultural, corresponde à organização fundamental da mente de um indivíduo (ou sociedade). Essa organização mental é moldada por um conjunto de pressupostos que dirigem os seus atos, seus pensamentos, suas disposições e seus juízos. Esses pressupostos têm origem nos grupos sociais de referência e nas experiências do indivíduo. Ao mesmo tempo, esses pressupostos têm caráter ontológico e epistemológico, determinando quais idéias ou crenças são, subjetivamente, consideradas válidas e relevantes. Esses pressupostos são centrais na elaboração do pensamento indivídual, sendo referências subjetivas numa grande variedade de contextos [Cobern (1996)].

além de suas disciplinas de origem em direção à "Terra Média"[2] da transdisciplinaridade[3], onde o que importa é a dinâmica gerada pela concomitante interação de diferentes paradigmas (ou níveis da "Realidade"). Enquanto a pesquisa disciplinar diz respeito a fragmentos de um único e mesmo nível de "Realidade"[Nicolescu (2000)], a postura transdisciplinar busca pelo conhecimento nos interstícios interdisciplinares, de tal forma que as pesquisas disciplinares e transdisciplinares não são mais antagonistas mas complementares. A transdisciplinaridade é uma abordagem que passa entre, além e através das disciplinas, numa busca de compreensão da complexidade. Sendo complementar à abordagem disciplinar, a transdisciplinaridade emerge da busca da unificação semântica e operativa das acepções através e além das disciplinas. Sem esquecer dos rigores do método científico, a transdisciplinaridade pressupõe uma racionalidade aberta, atentando para a relatividade das noções de "definição" e de "objetividade". O excessivo formalismo, a rigidez das definições e a absolutização da objetividade, incluindo-se a exclusão do sujeito, conduzem a uma ética empobrecida, onde o saber insulado priva o homem da compreensão compartilhada, fundamentada no respeito absoluto às alteridades unidas pela vida comum.

Ao abordar a complexidade, a abordagem transdisciplinar busca, em última instância, entender e conjugar todas as facetas de um fenômeno, levando ao entendimento *holístico* do mesmo, necessidade que já havia sido levantada por Ludwig von Bertalanffy [Mulej et al. (2004)]. Ao introduzir, na década de 1940, o conceito de Teoria dos Sistemas, Ludwig von Bertalanffy chamou a atenção para a necessidade de uma nova forma de construção do conhecimento, contrária à abordagem reducionista. Com isto, Bertalanffy enfatizava a necessidade de reviver a visão de unicidade da ciência. Porém, tal fundamento epistemológico acaba remetendo a um contexto onde o pesquisador deveria ter competências diversificadas que permitissem a ele abordar um mesmo problema com as visões de múltiplas disciplinas.

Apesar de instigante, a abordagem transdisciplinar de qualquer fenômeno não é uma tarefa trivial. A maior dificuldade está em compreender e harmonizar os paradigmas e ontologias das diferentes áreas pelas quais perpassa um trabalho desta natureza. Quem se propõe a tal empreitada, deve estar ciente de que apreender "as visões de mundo" basilares das áreas do saber estranhas às suas origens é algo difícil e que não ocorre prontamente.

Vemo-nos então tal como um peregrino, que deixa sua família, seus ofícios e sua cultura, e se lança em direção a terras das quais só ouviu falar, numa jornada onde ele busca o conhecimento de que necessita, mas que não se encontra em sua cultura de origem. Para sobreviver, este peregrino dever ter consigo somente o necessário para a sua sobrevivência, despojando-se de tudo que possa vir a ser um estorvo para a sua jornada. Para entender o que está por vir, devemos então esquecer o que éramos em nossa origem e despir-nos das arraigadas "verdades" pelas quais vemos as "sombras da realidade". Desta forma, tal como crianças, estaremos prontos para escutar e apreender os novos conceitos e novas "visões de mundo" que são compartilhadas pelos estudiosos de além. Sem isto, a viagem será tediosa,

[2]Terra Média é o nome para a terra antiga de John R. R. Tolkien, onde a maioria dos contos do seu imaginário ocorrem.

[3]O termo Transdisciplinaridade foi apresentado por Jean Piaget em um colóquio da UNESCO, de 1972, sobre interdisciplinaridade.

sofrida e inútil.

Começando dos princípios, será possível aprender o novo. Depois, recuperando o que já era sabido, faz-se o amálgama dos conhecimentos oriundos das múltiplas áreas. Muito do conhecimento revolucionário que alçou a humanidade a novos níveis sociais, de qualidade de vida e de conhecimentos sobre o universo e a vida, deve-se a estudiosos que permearam o árduo caminho sintetizado nesta metáfora.

Segundo Kuhn [Kuhn (1962)], uma revolução científica define-se pelo aparecimento de novos esquemas ou "paradigmas"[4]. Estes põem em evidência aspectos que não eram anteriormente vistos nem percebidos, ou eram mesmo suprimidos na ciência geralmente aceita e praticada no momento. Por conseguinte há um deslocamento nos problemas observados e estudados e uma mudança das regras da prática científica, comparável à troca de percepção psicológica de um fenômeno já conhecido.

Ainda segundo Kuhn[Kuhn (1962)], as primitivas versões de um novo paradigma são na maioria das vezes toscas, resolvem poucos problemas e as soluções dadas às diferentes instâncias dos problemas estão longe de serem perfeitas. Ocorre então uma profusão e competição de teorias, cada uma das quais limitadas no que tange ao número de problemas a que se refere e à solução dos que são levados em consideração. Contudo, a virtude do novo paradigma está em permitir abarcar novos problemas, em especial aqueles que anteriormente eram tratados como fora de escopo ou como sendo metafísicos.

Bioinformática é uma, dentre as emergentes áreas do conhecimento transdisciplinar, onde novos paradigmas têm tomado lugar. Como área de produção de conhecimento e técnicas, a Bioinformática tem sido vista com interesse por diferentes segmentos da sociedade. Ao conjugar conhecimentos advindos das áreas biológicas, da computação, da física e da matemática, a Bioinformática vêm, na última década, auxiliando a alavancar o desenvolvimento das pesquisas biológicas em variadas frentes, como na pesquisa genética e na biologia molecular. Os potenciais ganhos científicos, políticos e financeiros decorrentes da aplicação dos conhecimentos e técnicas deriváveis destas novas pesquisas vem estimulando os investimentos na formação de pesquisadores e em pesquisas puras e aplicadas. Na atualidade, a Bioinformática está, preferencialmente, voltada para as pesquisas em problemas originários da Bioquímica e Biofísica.

Tanto para a Bioquímica quanto para a Biofísica, o problema da estabilidade estrutural e dinâmica das proteínas, apesar de não ser recente, tem sido objeto de renovado interesse. Nos sistemas vivos, as proteínas são as operárias por excelência. Nestes sistemas, as proteínas exercem papéis cruciais em todos os processos biológicos. Sua importância e notável gama de atividade as levam a exercer diferentes funções: catalítica e enzimática; transporte e armazenamento; estrutural e mecânica; protetora; geração e transmissão de impulsos; controle do metabolismo, do crescimento e da diferenciação celular, dentre outras.

Tal diversidade de funções faz destas um elemento de relevada importância, visto que a regulação de sua expressão e atividade pode permitir que as mesmas sejam alvo de novas

[4] Kuhn usou o termo "paradigma" para se referir às estruturas e/ou compreensões mentais do mundo adotadas por comunidades científicas ou outros contextos epistemológicos.

drogas, possam ter suas ações reguladas ou serem usadas para atribuir novas propriedades a plantas e animais. Muito das funções exercidas pelas proteínas deve-se à inerente plasticidade de sua conformação tridimensional. Ao mesmo tempo, sabe-se que a função de uma proteína está diretamente relacionada à sua conformação tridimensional [Lehninger et al. (2007)]. O conhecimento necessário para ativar, regular ou inativar uma proteína, demanda o conhecimento das mudanças pelas quais a proteína de interesse passa ao longo do seu ciclo de vida.

Anfinsen [Anfinsen (1973)] conjectura que a função biológica de uma proteína está mais relacionada à sua geometria que aos detalhes químicos. Ainda segundo Anfinsen, somente a geometria de uma proteína e seu sítio ativo necessitam ser conservados para que a proteína mantenha sua função biológica. Diante desta constatação, deduz-se que a manutenção da estabilidade estrutural de uma proteína assume importância central. Por outro lado, conhecer como uma proteína se mantém estável, ao mesmo tempo em que é capaz de mudar sua conformação para atender às demandas de função e contexto em que atua, exige o conhecimento das forças e interações que definem sua estrutura tridimensional.

Em todos os organismos vivos, partindo da informação contida no RNA ou DNA, uma cascata de processos envolvendo transcrição ("leitura" do DNA e produção do RNA mensageiro) e tradução ("leitura" do RNA mensageiro pelos ribossomos e produção da seqüência de resíduos de aminoácidos), chega-se à estrutura primária da proteína. O termo "seqüência primária" é adotado para indicar que a proteína neste estágio mostra-se como um filamento totalmente amorfo. À medida que ela interage consigo e com o seu ambiente, uma nova seqüência de processos, cujos detalhes ainda são pouco conhecidos, faz com que este filamento vá se enovelando até atingir uma forma tridimensional estável, a qual é denominada "estrutura terciária" da proteína.

Uma proteína, na sua forma tridimensional, não é uma estrutura estática. Mesmo apresentando alta densidade, uma proteína mostra alguma fluidez em sua porção mais externa. Isto leva alguns estudiosos a considerá-la como sendo um material que apresenta altíssima viscosidade [Iben et al. (1989)]. À semelhança de outros sistemas físicos, uma proteína enovelada continua apresentando agitação molecular, já que sua temperatura encontra-se bem acima do 0 Kelvin. De fato, a estrutura terciária da proteína oscila em torno de uma conformação média estável, a qual acredita-se ser a observada nos ensaios cristalográficos, com uma freqüência da ordem de *femtosegundos*($10^{-15}s$) [Zhu et al. (1994)].

Quando enovelada, a proteína encontra-se em condições de exercer suas funções. À medida que esta interage com seu ambiente (interagindo com ligantes ou sendo exposta a variações de pH, dentre outras situações possíveis), uma proteína sofre uma mudança conformacional como uma resposta adaptativa às condições em que se encontra naquele momento. Tais mudanças podem alterar o comportamento da proteína, até que uma nova mudança de conformação ocorra. Desta forma, a função de uma proteína pode ser contingenciada tanto pelo contexto presente, quanto pela seqüência de estados pelos quais ela passou anteriormente. De fato, as nuances dos processos pelos quais passam as diferentes proteínas não são totalmente

conhecidas. Do processo de síntese, até os fenômenos de modulação alostérica[5], muito ainda precisa ser entendido. Em especial, a estabilidade termodinâmica das proteínas emerge como uma propriedade associada à robustez inerente à rede de interações atômicas que as formam.

Muitas destas propriedades, contudo, não são observáveis diretamente. Semelhante a outras ciências onde o estudo dos fenômenos de elevada complexidade só é possível com o uso de modelos, por meio dos quais o comportamento dos sistemas de interesse pode ser simulado por meios computacionais. Para que os fenômenos biofísicos relativos a estas macro moléculas possam ser melhor estudados com o auxílio de ferramentas computacionais, um modelo formal das proteínas se faz necessário. A despeito da sofisticação e da capacidade apresentada pelos métodos correntes de simulação computacional da dinâmica molecular, tais métodos são genéricos e aplicáveis a diferentes tipos de moléculas, não sendo um modelo específico que possa ser usado para explicar os fenômenos típicos das proteínas nem prever comportamentos ainda não observados para as mesmas.

Acreditamos que, para entender estes fenômenos, a visão das proteínas deve ir além da percepção das mesmas como "macro moléculas", devendo estas ser vistas como "sistemas". A questão não é só de terminologia mas sim ontológica. Vista como sistema, a proteína herda as propriedades comuns já descritas para os sistemas em geral. À semelhança de outros estudos relativos aos sistemas, este deve iniciar pela busca do entendimento da estrutura das proteínas. O entendimento de um sistema começa com a identificação dos seus diferentes elementos e as diferentes relações que estes elementos estabelecem entre si. A diversidade de elementos e relações que constituem o sistema acabam por definir a sua estrutura.

Esta tese foca o entendimento da estabilidade estrutural e dinâmica das proteínas como problema central. Longe de querer dar solução a este problema, esta tese enseja perscrutar as proteínas sob a óptica sistêmica. A relevância deste esforço reside na expectativa que o entendimento futuro das questões relativas à estabilidade termodinâmica e modulação conformacional das proteínas, venha contribuir para avanços em outras áreas tais como a da saúde humana e animal, e da produção de alimentos.

Ao contemplar a proteína como um sistema espera-se, neste estudo, observar nestas macromoléculas as mesmas propriedades apresentadas por outros sistemas complexos existentes no mundo real. Para estudar este arranjo de interações entre os constituíntes das proteínas, a adoção dos métodos de estudo, já consolidados, relativos às redes complexas aparece como uma opção natural.

A propriedade do uso dos modelos de redes complexas, no estudo das proteínas, decorre do fato de que, ao longo da última década, o estudo das redes complexas tem contribuído para melhorar o entendimento de diferentes fenômenos do mundo real. Ao mesmo tempo, tem sido significativo o avanço na elaboração dos modelos teóricos, os quais têm sido objeto de renovada atenção por parte da comunidade acadêmica internacional. Nestes anos muito têm sido feito relativo aos aspectos estruturais dos sistemas complexos. Isto fica evidente pelo número de publicações devotadas a este tema. Nos campos do saber afetos aos sistemas

[5]Mudança na forma e nas propriedades de uma proteína que se segue como conseqüência da interação de uma outra molécula com esta proteína.

sociais, naturais ou artificiais, a modelagem computacional de sistemas tornou-se amplamente aceita como tendo validade científica. A adoção desta abordagem para o estudo dos sistemas biológicos é inquestionável, já que na própria noção de vida a idéia de complexidade tem importância central [von Bertalanffy (1950)].

A adoção da abordagem sistêmica das proteínas, remete à necessidade inicial de identificar as relações que os constituintes das proteínas estabelecem entre sí, para depois entender como do arranjo conjunto destes elementos emerge a estrutura tridimensional estável das proteínas. Similar à abordagem adotada no estudo de outros sistemas reais, esta tese centra na identificação e caracterização da rede formada pelas interações não-covalentes[6] existentes entre os átomos/resíduos que constituem as proteínas. A justificativa para a abordagem sistêmica das proteínas, apresentada nesta tese, reside na constatação de que, a despeito da existência de vários estudos a respeito dos aspectos dinâmicos e da modulação conformacional das proteínas, um estudo sistemático que vincule a rede de interações subjacente à estrutura terciária, com estes aspectos dinâmicos, ainda está por ser feito.

Contudo, a identificação da rede de interações, que estabiliza a estrutura tridimensional das proteínas, demanda a correta caracterização das próprias interações, existentes entre os átomos/resíduos que constituem as proteínas. O interesse neste estudo recai sobre as interações de natureza não covalente, visto que as interações de natureza covalente definem o encadeamento seqüencial dos resíduos de aminoácido nas proteínas, não são determinantes para a estabilidade tridimensional da proteína, após esta ter se enovelado, com exceção das ligações entre resíduos de cisteína. Adotando uma visão de granulação mais fina, as proteínas são inicialmente vistas neste trabalho como sendo uma malha de interações entre átomos, deixando a abordagem no nível dos resíduos de aminoácidos, para um momento posterior. Esta abordagem inicial justifica-se visto que nesta granulação é possível captar cada uma das interações entre átomos que ocorrem dentro das proteínas. Desta identificação é possível deduzir quais apresentam os atributos físicos e químicos que permitam classificá-las como sendo factíveis. A partir dos mesmos atributos, é possível calcular a energia associada a cada uma destas interações.

A técnica aqui proposta permite fazer uma associação mais realística de fatores de relevância física aos pesos das arestas para a construção dos modelos de rede. Um tratamento computacional exaustivo foi necessário para identificar as energias das interações para todos os pares atômicos que formam as estruturas das proteínas analisadas. A identificação da rede de interações não-covalentes entre os átomos, em uma proteína, passa pela identificação de cada uma destas interações, o que requer alguns cuidados. Contudo, é preciso considerar que os átomos só podem estabelecer interações se estes forem mutuamente acessíveis. Esta acessibilidade tem a ver tanto com a distância euclidiana entre um par de átomos, mas também é influenciada pela densidade de átomos na vizinhança deste par. Em outras palavras, um dado par de átomos não pode estabelecer interação se existir, entre eles, um outro átomo que obstrua esta interação. Neste trabalho, este problema é solucionado, estudando o fenômeno

[6] Três os tipos de ligações não-covalentes ocorrem nas proteínas: ligações iônicas, ligações de hidrogênio, interações de van der Waals.

da oclusão entre os átomos. Ao mesmo tempo, cada um dos modelos de proteína estudados foi solvatado. Isto evita que o modelo considere interações entre os átomos das cadeias laterais, dos resíduos expostos na superfície da proteína, que não existem na realidade. Sem a solvatação, estas interações seriam identificadas no modelo, o que na realidade não ocorre.

Outro aspecto da análise das interações não-covalentes entre os átomos é a determinação da energia potencial associada a cada uma delas, considerando os potenciais de Coulomb e de Lennard-Jones. Isto permite que seja associado um peso à cada uma destas interações. Ao mesmo tempo, os modelos resultantes apresentam uma semântica física melhor que a correntemente adotada.

Esta abordagem, difere conceitualmente das abordagens adotadas nos trabalhos onde o mesmo problema foi contemplado até agosto de 2007 [Greene e Higman (2003), Amitai et al. (2004), Atilgan et al. (2004), Rao e Caflisch (2004), Bagler e Sinha (2005), Brinda e Vishveshwara (2005), Kundu (2005), del Sol e O'Meara (2005), Kundu (2005), Aftabuddin e Kundu (2006), Alves e Martinez (2006), Higman e Greene (2006), del Sol et al. (2006b), del Sol et al. (2006a), Atilgan et al. (2007), Jiao et al. (2007)]. Enquanto estes trabalhos adotam limiares arbitrários, e variados, para as distâncias entre os átomos, sem considerar as eventuais interferências estéricas entre estes átomos, a abordagem que propomos, inova ao não estipular nenhum limiar fixo para as distâncias entre os átomos, condição que neste caso emerge naturalmente como conseqüência da existência destas interferências estéricas. Mesmo que estas eventuais interferências não obliterem totalmente as interações entre os átomos, elas podem influir nas mesmas, induzindo perturbações que irão diminuir a energia de interação entre estes átomos, condição contemplada nos métodos aqui adotados.

Como resultado, mostramos a existência de uma rede interações não-covalentes comum às globinas (página 128), pode-se estimar que tal propriedade deve ser comum às demais famílias topológicas, sendo que cada uma deve apresentar um padrão próprio. Como conseqüência, especula-se sobre a possibilidade futura de criar um modelo estrutural típico para cada família topológica, e em dedicar futuros estudos aos aspectos dinâmicos das proteínas. O valor destes conhecimentos reside na possibilidade de que os mesmos possam contribuir futuramente para a melhoria do entendimento dos fenômenos funcionais das proteínas, entendimento que faculte o projeto e uso racional deste admirável material para os mais diversos fins.

Ao colocar o entendimento da estabilidade estrutural e dinâmica das proteínas como problema central, este trabalho objetiva, de forma geral, mostrar rigorosamente que a rede de interações não-covalentes, entre átomos, subjacente a uma proteína, é específica para cada família topológica, ao mesmo tempo em que a caracterização destas interações demanda cuidados na eleição de critérios de análise.

Busca-se ainda neste trabalho, de forma mais específica:

Desenvolver um método não supervisionado de identificação de interações não-covalentes entre átomos dentro de uma proteína, e que se apóie em fundamentos físicos e químicos bem definidos, ponderando as interações não-covalentes entre átomos com base na energia potencial associada à cada uma das interações;

Identificar as características topológicas da rede de interações não-covalentes ponderada comum a todas as proteínas de uma mesma família topológica, e identificar qual a similaridade desta rede com relação aos modelos canônicos de redes complexas;

Identificar nesta rede comum a todas as proteínas de uma mesma família topológica, os grupos de átomos topologicamente conservados que respondem pela deflagração dos sinais indutores de alterações alostéricas;

Identificar nesta rede comum a todas as proteínas de uma mesma família topológica, os átomos topologicamente conservados que apresentam os maiores valores de energia potencial, e que atuem como vértices com papel estrutural relevante nestas redes;

Identificar para estes átomos, a que resíduos eles pertencem, em que posições nas seqüências primárias das proteínas eles se encontram e identificar se estas posições conservam alta energia potencial nas proteínas de uma mesma família topológica;

Identificar para estas posições, quais são os atributos evolutivamente conservados.

Visando apresentar todas as etapas seguidas no decorrer do desenvolvimento dos estudos, este trabalho está organizado da seguinte forma:

O capítulo 2 discute a visão das Proteínas como sistemas complexos, apresentando uma leitura detalhada dos fundamentos da Teoria Geral dos Sistemas e da Cibernética, e como os mesmos são aplicáveis ao entendimento holístico das proteínas.

No capítulo 3 é apresentada uma revisão dos conceitos desenvolvidos nos estudos sobre redes complexas, apresentando os fundamentos que permitirão o posterior estudo das redes de interações não-covalentes entre os átomos das proteínas.

O capítulo 4 apresenta as proteínas selecionadas para este estudo e como esta seleção foi efetuada. Este capítulo apresenta, ainda, todo o desenvolvimento matemático feito para a definição dos métodos de identificação das interações atômicas existentes nas proteínas em análise. Para este trabalho, uma série de proteínas, cujas seqüências de resíduos de aminoácidos apresenta baixo grau de homologia mútua, foi selecionada para a família das globinas e para as serinoproteases. As redes de interações não-covalentes entre os átomos de cada uma destas proteínas foram identificadas, com o uso de um método de análise desenvolvido, especificamente para este fim, onde a energia potencial associada a estas interações foi usada como um atributo relevante para as análises. Por fim, neste mesmo capítulo, são discutidos outros cuidados tomados com relação aos dados, para não comprometer a validade dos resultados.

No capítulo 5 são apresentados e discutidos os resultados obtidos na identificação das redes de interações não-covalentes entre os átomos das proteínas. Tais resultados mostraram a existência, do ponto de vista topológico, de resíduos e propriedades evolutivamente conservadas e de uma estrutura hierárquica entre os diversos resíduos das seqüências, a qual parece influenciar na forma e na estabilidade das proteínas.

Este estudo de caráter exploratório visa contribuir para a melhoria do entendimento das características estruturais das proteínas. Ao adotar um novo método de análise e identificação

das interações não-covalentes entre os átomos no seio das proteínas (o qual será detalhado no capítulo 4), foi possível identificar as interações mais plausíveis e quantificar a energia potencial inerente a estas interações. Observadas em conjunto, estas interações mostram a existência de uma rede entre os elementos (átomos/resíduos) constituintes das proteínas estudadas, onde esses elementos apresentam uma hierarquia baseada na conectividade e na energia das interações. Tais redes hierárquicas apresentam propriedades notáveis como alto grau de aglomeração, pequena distância média entre vértices, alta resiliência a mutações. Ao mesmo tempo, tudo sugere que as redes subjacentes às proteínas apresentam propriedades de transmissão de sinais alostéricos bem característicos, onde o fluxo da informação mostra ser direcionado, sendo os sítios de ligação os pontos de deflagração destes sinais.

Visto que estudos desta ordem ainda constituem uma novidade no estudo das proteínas, acreditamos que os problemas relativos à estabilidade estrutural e à dinâmica das mesmas são tópicos que ainda têm muito a ser investigado e para os quais respostas mais satisfatórias devem ser identificadas. Os achados apresentados neste trabalho sugerem que a visão das proteínas como sistemas, aqui proposta, constitui um paradigma pertinente. Em conjunto, estes achados demonstram que as proteínas apresentam comportamentos similares aos observados em outros sistemas complexos encontrados no mundo real. Ao mesmo tempo, esses resultados contibuem para demonstrar a pertinência da elaboração, e uso de modelos formais para as proteínas. Potencialmente, estes modelos poderiam auxiliar os avanços futuros para aprimorar o conhecimento das proteínas e no avanço do uso destas para a melhoria da qualidade de vida da humanidade.

Capítulo 2

Proteínas como Sistemas

Mais les parties du monde ont toutes un tel rapport, et un tel enchaînement l'une avec l'autre, que je crois impossible de connaître l'une sans l'autre et sans le tout.

Pascal

Penseés

Desde os primeiros trabalhos de von Bertalanffy [von Bertalanffy (1950)] e posteriormente com Norbert Weiner [Wiener (1948)], o entendimento dos problemas complexos tem sido um tema recorrente nas ciências biológicas. Entretanto, o renovado interesse hodierno por estes problemas decorre dos progressos observados na biologia molecular que estão permitindo amealhar um extenso conjunto de dados sobre os sistemas biológicos. O poder computacional disponível na atualidade, vem permitindo a análise dos sistemas biológicos em diferentes níveis, resultando em um contínuo e amplo espectro de conhecimentos.

A abordagem sistêmica da biologia, ou Biologia Sistêmica, é baseada em duas proeminentes características [Dhar et al. (2004)]. A primeira diz respeito ao fato de que ela é construída sobre conhecimentos obtidos a partir da biologia experimental. Segundo, as tecnologias da informação permitem a manipulação e a integração do vasto volume de dados observáveis destes sistemas. Desta forma, o intuito nesta abordagem deixa de ser a mera apresentação dos dados e passa para a busca da descrição integrada dos vários níveis em que a vida se expressa, entendendo-a nos seus aspectos estruturais, dinâmicos e de regulação.

A chave para tal está em encontrar representações apropriadas para descrever quantitativamente os fenômenos biológicos. De fato, muito dos processos biológicos podem ser melhor descritos com o uso das metáforas oriundas dos sistemas de informação, o que é mais difícil de ocorrer com o uso dos formalismos matemáticos. Isto é de especial importância quando os estudiosos estão mais habituados a lidar com conhecimentos fenomenológicos, do que com descrições mecanísticas e quantitativamente precisas dos fenômenos [Dhar et al. (2004)]. Neste caso incluem-se as redes de regulação, de transdução de sinais e formação de padrões, todos resultantes dos processos de desenvolvimento e evolução dos organismos. A importância de abordagem computacional na biologia sistêmica está no fato de que ela provê uma descrição efetiva dos sistemas em diferentes níveis. Ao contrário de outras abordagens,

onde as influências macro (ou micro) sistêmicas são ignoradas ou simplificadas, a abordagem computacional permite, ao menos potencialmente, incorporar estas influências nos fenômenos estudados.

Dentre os fenômenos relacionados aos sistemas biológicos, o entendimento do processo de enovelamento e a manutenção da estabilidade das proteínas permanece sendo um dos grandes problemas. Estes fenômenos têm sido enfocados de maneiras diversas ao longo das últimas décadas. Apesar de alguns progressos notáveis feitos em direção à elucidação dos mesmos, estes constituem um grande desafio que adentra este novo século. De forma poder tratar com estes fenômenos, diferentes abordagens têm sido utilizadas, trazendo algumas pistas sobre os atributos intrínsecos da estrutura das proteínas. Entretanto, as complexidades estruturais e dinâmicas inerentes aos aglomerados atômicos, que caracterizam as proteínas, mostram a exemplo de outros problemas similares, quão difícil é lidar com os arranjos sistêmicos que ocorrem na natureza.

Neste capítulo uma nova perspectiva é endereçada, onde a proteína é vista sob a óptica do paradigma sistêmico. Ao contemplar a proteína como um sistema, infere-se que a mesma seja dotada das propriedades inerentes a todos os sistemas complexos observados no mundo real. Para expor com mais detalhes tal abordagem, este capítulo apresenta uma introdução sucinta aos princípios da Teoria Geral dos Sistemas, para depois apresentar uma discussão sobre esta visão mais abrangente do que pode ser uma proteína.

2.1 Os Princípios da Teoria Geral dos Sistemas

> *As ciências modernas têm se caracterizado por sua constante especialização muitas vezes decorrente do grande acúmulo de conhecimento e crescente complexidade de técnicas e de estruturas teóricas aplicáveis nos seus respectivos campos de estudos. Desta forma, as ciências vêm se dividindo em inumeráveis novas disciplinas. Como conseqüência, os pesquisadores de cada área tendem a se encapsular cada vez mais nos domínios dos respectivos campos de estudos tornando cada vez mais difícil o intercâmbio de conhecimento entre estes casulos.*

O texto extraído do clássico *General Systems Theory* apresenta a percepção de Bertalanffy [von Bertalanffy (1975)] quanto às dificuldades com que se depara aquele que se aventure a explorar temas que se encontram nos limites de disciplinas científicas já sedimentadas. Segundo Bertalanffy [von Bertalanffy (1975)], a observação da Natureza deixa patente o fato de que os elementos e fenômenos observáveis no Universo são estudados, em quase sua totalidade, por diversas disciplinas da ciência clássica as quais tendem a isolar estes elementos e fenômenos do Universo observável de forma que os mesmos apresentem uma complexidade reduzida e possam desta forma ser mais facilmente entendidos. Contudo, ao serem contemplados de forma conjunta, os conceitos e experimentos originados destas áreas "estanques", não formam um modelo coeso.

Na década de 1940, Ludwig von Bertalanffy apresentou a Teoria Geral dos Sistemas, chamando a atenção para a necessidade de uma nova forma de construção do conhecimento diversa da tradicional abordagem reducionista. Em seu trabalho, von Bertalanffy enfatizou o conceito onde um sistema pode ser constituído de inúmeros sub-sistemas que apresentam uma dinâmica interna, ao mesmo tempo em que mantêm interações com seus vizinhos e com seu ambiente. Assim, ao invés de reduzir um sitema às propriedades de suas partes, a Teoria Geral dos Sistemas [von Bertalanffy (1975)] foca no arranjo e nas relações entre estas partes, as quais apresentam-se conectadas formando um todo. Desta maneira, estes sistemas evoluem no tempo, apresentanto novas propriedades que emergem do amálgama dos processos endógenos e das influências exógenas. Contrariando a abordagem tradicional de reduzir um sistema à soma das propriedades de suas partes, a Teoria Geral dos Sistemas [von Bertalanffy (1975)] busca focar no arranjo e nas relações entre estas partes, as quais apresentam-se conectadas formando um todo.

A Teoria Geral de Sistemas [von Bertalanffy (1975)] surge como uma forma de pensar, de caráter transdisciplinar, cujo objeto de estudo são os sistemas e os seus fenômenos, tendo como modelo os níveis mais abstratos e abrangentes de organizações. Estas organizações são estudadas independente de sua substância, tipo ou escala temporal e espacial. A investigação recai sobre princípios comuns a todas as organizações complexas e modelos, usualmente matemáticos ou computacionais, adotados para descrevê-las. Ao elaborar suas hipóteses von Bertalanffy [von Bertalanffy (1975)] concebe os sistemas como um conjunto de elementos e suas interrelações. Nenhuma hipótese é feita quanto a natureza do sistema, de seus elementos ou quanto às relações entre os mesmos. Partindo desta definição várias propriedades foram inferidas [von Bertalanffy (1975)], algumas sendo expressas em termos bem definidos em vários campos do conhecimento, outras definidas em termos antropomórficos ou metafísicos. Assim, von Bertalanffy propõe que o paralelismo das concepções gerais ou mesmo de "leis" especiais em diferentes campos do conhecimento decorre como conseqüência do fato de serem os fenômenos entendidos como pertinentes a "sistemas" sendo que estes "princípios" identificados seriam aplicáveis a quaisquer sistemas independente de sua natureza e dos elementos envolvidos. Assim, por mais complexo ou diverso que seja o mundo que se experimente, nele sempre serão encontrados diferentes tipos de organização as quais podem ser descritas a partir dos conceitos e princípios que são independentes do domínio específico a partir do qual são observados [von Bertalanffy (1975)]. Desta premissa seria então possível inferir que, ao menos em tese, descobertas as "leis" gerais que regem quaisquer sistemas, seria possível analisar e resolver problemas sistêmicos em qualquer domínio do conhecimento. A Teoria Geral dos Sistemas seria a "Ciência do Todo" [von Bertalanffy (1975)]. Em sua clássica citação *"The whole is more that the sum of its parts"* [von Bertalanffy (1975)], von Bertalanffy expressa o entendimento de que as características constitutivas de um sistema não são, *a priori*, explicáveis a partir das características das partes quando isoladas. As características deste "complexo" apareceriam como "novas" ou "emergentes". Entretanto, ressalta Bertalanffy [von Bertalanffy (1975)], se uma soma é concebida como sendo composta de forma gradual, deve-se conceber um sistema como se ele fosse composto instantaneamente e não

como o incremento gradual das partes. Ocorreria aí algo similar a uma "mudança de fase".

Não existe na atualidade, uma Teoria Unificada da Complexidade, mas algumas teorias emergem de vários estudos relativos aos fenômenos complexos relacionados a diferentes áreas do conhecimento como a biologia, química, computação, ecologia, matemática e física. Estas teorias incluem os trabalhos conduzidos ao longo das últimas quatro décadas como: Stuart Kauffman [Kauffman (1993), Kauffman (1996), Sole et al. (2005)] John Holland [Holland (1995), Holland (1998)], Chris Langton [Waldrop (1992)] e Murray Gell-Mann [Gell-Mann (1995)] do Santa Fe Institute, sobre sistemas complexos adaptativos (CAS), bem como os trabalhos de estudiosos europeus como Peter Allen [Allen (1997)] e Brian Goodwin [Goodwin (1995), Webster e Goodwin (1996)]; Axelrod (Axelrod (1990), Axelrod (1997), Axelrod e Cohen (2000)); Casti [Casti (1998)], Bonabeau et al (Bonabeau et al. (1999)), Epstein e Axtel (Epstein e Axtell (1996)) e Ferber [Ferber (1999)] sobre modelagem e simulação computacional; nos trabalhos de Ilya Prigogine [Prigogine e Stengers (1984), Nicolis e Prigogine (1989), Prigogine (1990)] sobre sistemas e estruturas dissipativas; Humberto Maturana, Francisco Varela [Maturana e Varela (1992)] e Niklaus Luhman [Luhman (1990)] sobre autopoiesis; bem como nos trabalhos de Gleick sobre a teoria do caos [Gleick (1987)].

Os trabalhos mencionados podem ser sumarizados em quatro áreas de pesquisas: (a) sistemas complexos adaptativos; (b) estruturas dissipativas; (c) autopoiesis; (d) teoria do caos. A observação conjunta destes trabalhos permite contemplar a existência de dez princípios genéricos relativos à complexidade os quais serão discutidos neste capítulo: auto-organização, emergência, conectividade, interdependência, realimentação, não-equilíbrio, espaço de possibilidades, coevolução, evolução temporal, dependência de trajetória.

Cabe aqui uma discussão a respeito do conceito da existência de princípios genéricos, no sentido de que estes princípios, ou características seriam comuns a todos os sistemas complexos encontrados no mundo real. Uma forma de olhar para os sistemas biológicos é observar as características genéricas dos sistemas naturais e considerar se estas características são relevantes ou apropriadas para o sistema biológico em questão. Contudo, esta abordagem apresenta uma limitação onde tal analogia é simplesmente um ponto de partida para a análise do sistema e não um mapeamento obrigatório.

Esta limitação deve ser enfatizada por duas razões:

- apesar de ser desejável que uma explicação válida em um domínio seja consistente com a explicação válida em um outro domínio, e que estas explicações adiram ao Princípio da Consistência [Hodgson (2001)], características e comportamentos de diferentes sistemas não podem ser mapeadas diretamente de um domínio em outro, sem que haja um processo rigoroso de teste da propriedade e relevância destes mapeamentos. Não somente a unidade de análise entre os sistemas pode ser diferente, mas também os domínios científicos podem também apresentar certas diferenças fundamentais que podem invalidar o mapeamento direto.

- deve-se considerar que os princípios da complexidade podem, dependendo da situação, ser somente metáforas e analogias que são limitadas e também limitam (podendo

inclusive atrapalhar), o entendimento do sistema em estudo. Isto não significa que as metáforas e analogias não devem ser usadas. Contudo, estas devem aparecer como mecanismos de auxilio para que seja possível fazer a transição de um domínio do conhecimento para outro, principalmente quando se defronta com novas idéias ou conceitos.

Estes princípios provêem uma base racional para o estudo da complexidade, auxiliando no entendimento da natureza do mundo e das organizações que nele são observadas. Contudo, tanto as teorias da complexidade quanto os seus princípios comuns provêem um arcabouço conceitual, uma forma de pensar e ver os fenômenos.

2.1.1 Conectividade e Interdependência

O comportamento complexo emerge do inter-relacionamento, da natureza das interações e da inter-conectividade dos elementos de um sistema, e do sistema com seu ambiente. Nestes sistemas, a conectividade e a interdependência implicam que uma perturbação induzida por um (ou em um) elemento do sistema tem o potencial de afetar os demais elementos relacionados ao elemento perturbador e o próprio sistema. Contudo, tais efeitos induzidos não são iguais para todos os elementos do sistema nem têm impactos sistêmicos uniformes. Ao contrário, estes efeitos e impactos variam em função do "estado" momentâneo de cada um dos elementos afetados, tornando-os mais ou menos sensíveis a estas perturbações. Tanto o estado de um elemento, quanto do sistema, são determinados (em maior ou menor grau) pela sua história e pela sua constituição, o que inclui sua organização e sua estrutura.

Quanto a interconectividade dos elementos constituintes do sistema, as teorias da complexidade mostram que o crescimento monotônico da interconectividade media dos elementos do sistema leva a graus indesejáveis de interdependência. Em outras palavras, uma grande interdependência dos elementos faz com que qualquer perturbação percebida por um dos elementos seja propagada para todos os demais elementos do sistema. Graus elevados de dependência entre os elementos, usualmente trazem efeitos indesejáveis para todo o sistema. Assim, se um dos elementos buscar uma condição melhor dentro do sistema, isto pode resultar na piora das condições de outros elementos. Assim, a melhoria alcançada por cada um dos elementos pode impingir "custos" associados a outros elementos ou ao próprio sistema. Estendendo o raciocínio, as mudanças percebidas por um dos elementos do sistema não são capazes de afetar o resto do sistema, se consideradas de forma isolada. A contribuição de tais mudanças para o sistema vai depender dos demais elementos naquele contexto. Esta dependência contextual influencia, direta ou indiretamente, como os elementos inter-relacionados irão responder às perturbações neles induzidas.

2.1.2 Co-evolução

Uma maneira de descrever co-evolução é aquela onde a evolução de um domínio ou de uma entidade é parcialmente dependente da evolução de outros domínios ou entidades ele relacionado [Kauffman (1993), Kauffman (1996)]. A noção de co-evolução coloca ênfase na evolução das interações e na evolução recíproca. Para Kauffman, a co-evolução tem lugar

dentro de um ecossistema e não pode ocorrer de forma isolada. Tal processo pode ocorrer relacionado com os elementos que compõem o ecossistema, mas também pode ocorrer nos relacionamentos e interações entre as entidades que co-evoluem. Embora exista uma distinção conceitual entre sistema e seu ambiente, é importante notar que não existe dicotomia ou uma fronteira bem definida entre estes termos, sendo que um sistema sempre reage de forma a se adaptar às mudanças de seu ambiente, ou ele se desagrega.

Neste contexto, as mudanças percebidas por um sistema devem ser vistas como sendo um processo de co-evolução associada aos demais sistemas com os quais ele compartilha o mesmo ecossistema. Nesta perspectiva, a co-evolução ocorre quando entidades relacionadas mudam ao mesmo tempo.

Os efeitos das mutações, por outro lado, propagam-se por todo ecosistema como uma função do grau de conectividade e interdependência entre os elementos do sistema, sendo que esta propagação ou influência não ocorre de forma uniforme, mas variando conforme o grau de conectividade dos diversos elementos.

2.1.3 Sistemas dinâmicos: Estruturas Dissipativas, Longe-do-equilíbrio e História

O conceito de sistema dinâmico nasce da necessidade de um modelo geral que explique os sistemas que evoluem segundo uma regra que liga o estado presente aos estados passados. Conceitualmente, um sistema dinâmico apresenta alternância de estados, os quais estão restritos a um conjuto finito de estados alcançáveis por este sistema [Martin et al. (1973)], conjunto este que define o seu Espaço de Fase. Um sistema dinâmico evolui regido por um conjunto de variáveis determinantes que se alteram ao longo do tempo, ou seja, definem o fluxo do sistema ao longo do espaço de fases. Assim, o estado do sistema, regido por n variáveis, é definido pela combinação destas variáveis, podendo ser representado por um ponto no espaço de fase. A trajetória do sistema é definida, então, como sendo a curva descrita no espaço de fase pela sucessão de estados pelos quais o sistema passou ao longo do tempo.

Para os sistemas dinâmicos estáveis, o sistema acaba por alternar sua trajetoria ao longo de um pequeno conjunto de estados característicos, denominados "atratores". Já nos sistema ditos "dissipativos", a trajetória do sistema, dentro do seu espaço de fases, tende a apresentar soluções alternativas em função das oscilações apresentadas pelas variáveis determinantes, no momento da mudança de fase. Tal fenômeno é genericamente denominado de "bifurcação". Contudo, neste contexto, o termo "bifurcação" não implica dizer que somente duas rotas alternativas podem ser seguidas pelo sistema. Ao contrário, a maioria dos sistemas dissipativos costuma apresentar variadas soluções possíveis [Nicolis e Prigogine (1989)].

Ilya Prigogine, em seus trabalhos sobre estruturas dissipativas, apresenta uma reinterpretação da segunda lei da termodinâmica, onde a entropia nem sempre leva um sistema à dissolução. Ao contrário,

> "under certain conditions, entropy itself becomes the progenitor of order. ... under non-equilibrium conditions, at least, entropy may produce, rather than degrade,

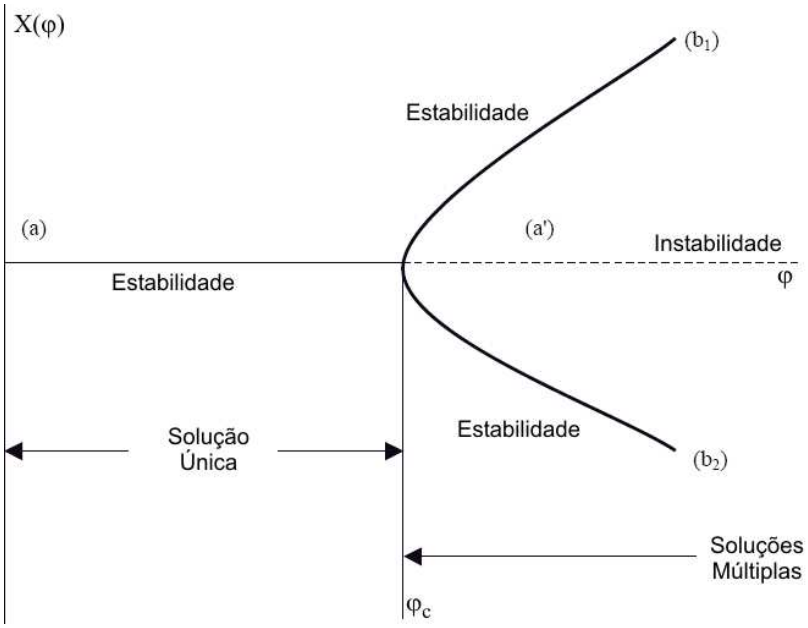


Figura 2.1 *– Diagrama de bifurcação mostra como o estado de um sistema* X(φ)*, dentro de um espaço de fases, é afetado quando o parâmetro de controle* φ *varia. O sistema, na condição onde* $\varphi \leq \varphi_c$*, estagia na região de estabilidade (*a*), onde tem-se condição de solução única, onde ele apresenta sempre a mesma fase. Além do limiar crítico* $\varphi = \varphi_c$*, o sistema sai da condição de estabilidade em que estava. Para estas nova situação onde* $\varphi > \varphi_c$*, o sistema pode derivar para novas configurações equiprováveis (*$b_1 \vee b_2$*) alcançáveis dentro do seu espaço de fases, as quais são estáveis nestas novas condições.*

> order (and) organisation ... If this is so, then entropy, too, loses its either/or character. While certain systems run down, other systems simultaneously evolve and grow more coherent ..."

[Prigogine e Stengers (1984)]. Desta forma, uma característica distintiva dos sistemas complexos, é a propriedade de derivar para novos estados de ordem interna.

Ao seu turno, o fenômeno de quebra de simetria ocorre quando a homogeneidade da ordem corrente é quebrada e novas estruturas aparecem. A quebra de simetria é um fenômeno que pode ser entendido como sendo um gerador de "informação", no sentido que quando um padrão de homogeneidade sistêmica é quebrado, as novas formas apresentadas podem ser vistas como sendo informações apresentadas pelo sistema [Prigogine e Stengers (1984)].

No caso dos sistemas dissipativos, ao ocorrer a mudança de fase, não é possível *a priori* predizer que estado emergirá;

> "...only the chance will decide,through the dynamics of fluctuations. The system will in effect scan the territory and will make a few attempts, perhaps unsuccessful at first, to stabilize. Then a particular fluctuation will take over. By stabilizing it the system becomes a historical object in the sense that its subsequent evolution depends on this critical choice ..."

[Prigogine e Stengers (1984)]. Em uma escala totalmente diferente, as noções de possibilidade

e história são usadas por Kauffman [Kauffman (1993)] para descrever uma perspectiva da biologia evolucionária que vê

> "... organisms as ultimately accidental and evolution as an essentially historical science. In this view, the order in organisms results from selection sifting unexpected useful accidents and marshalling them into improbable forms. In this view, the great universals of biology-the genetic code, the structure of metabolism and others-are to be seen as frozen accidents, present in all organisms only by virtue of shared descent ..."

.

A existência de estados de não-equilíbrio pode permitir que um sistema evite a degeneração entrópica e transforme parte da energia passada a ele pelo ambiente em um comportamento ordenado de um novo tipo, uma nova estrutura dissipativa caracterizada pela quebra de simetria pela possibilidade da ocorrência de múltiplas escolhas. Na química, o fenômeno de autocatálise mostra comportamentos similares, e a reação de Belousov-Zhabotinski, sob determinadas circunstâncias de não-equilíbrio mostra que a quebra de simetria, auto-organização, soluções múltiplas possíveis, e histerese (o trajeto específico entre os estados, que pode ser seguido, depende da história passada do sistema) [Prigogine e Stengers (1984), Kauffman (1993), Kauffman (1996)].

Além disso, a auto-reprodução, uma propriedade fundamental da vida biológica, é "o resultado de um ciclo autocatalítico no qual o material genético é replicado pela intervenção de proteínas específicas, sendo elas mesmas sintetizadas com com base nas instruções contidas no material genético ..." [Nicolis e Prigogine (1989)]. Em certo sentido, complexidade está associada aos sistemas em que a evolução e a história exercem ou exerceram um papel importante, sejam eles sistemas de caráter biológico, físico, ou químico.

2.1.4 Exploração do Espaço de Possibilidades

A complexidade sugere que para sobreviver e prosperar, uma entidade necessita explorar seu espaço de possibilidades para gerar variedade. A complexidade também sugere que a busca de estratégia única de condição "ótima" pode não ser possível e nem desejável. Toda a estratégia somente pode ser ótima sob determinadas circunstâncias, e quando aquelas circunstâncias mudam, a estratégia pode já não ser ótima.

Para sobreviver, uma entidade necessita fazer a varredura constante de seu ambiente, ao mesmo tempo em que atua tentando diferentes estratégias. Tal entidade pode necessitar ter "em mãos" diversas micro-estratégias que possam evoluir antes que os seus recursos estejam comprometidos com uma única estratégia. Isto reduz o risco de que esta entidade tenha à sua disposição uma única estratégia de sobrevivência demasiado cedo, a qual pode não ser a melhor, mas que conduza sua evolução sensível às mudanças do ecosistema. Em essência, a sobrevivência em ambientes instáveis requer abordagens flexíveis baseadas na variedade de requisitos [Ashby (1969)].

Ao pesquisar o espaço de fases, não é possível que todas as possibilidades sejam exploradas. Contudo, é possível considerar mudar um passo além do estado já existente. Assim, a adaptação pode ser considerada uma exploração do conjunto de estados adjacentes possíveis [Kauffman (1996)]. De acordo com Kauffman, a persistente introdução da "novidade" nos sistemas já estabelecidos acontece com a exploração das possibilidades adjacentes. A taxa de descoberta ou de mutação, entretanto, é restringida por seleção para evitar possíveis catástrofes que poderiam destruir o sistema. Os sistemas biológicos sustentáveis apresentam, usualmente, uma taxa de mutação bem abaixo do limite "erro-catástrofe", a partir do qual a transição de fase leva o sistema a configurações insustentáveis. Parece haver um contrapeso entre a descoberta e o que o ecossistema pode eficazmente sustentar. Os ecossistemas parecem ter "*os mecanismos endógenos que bloqueiam a exploração do possível adjacente de tal forma que, na média, tais explorações encontram com sucesso maneiras novas de fazer uma vida ...*" [Kauffman (1996)]. As adaptações que ocorrem na biosfera, são selecionadas pela seleção natural, em uma taxa que seja sustentável.

2.1.5 Realimentação

A retro-alimentação é vista tradicionalmente em termos de regulação positiva e negativa, sendo também descritos como mecanismos que atuam "reforçando (isto é amplificando) ou balanceando". Retro-alimentação positiva orienta as mudanças evolutivas. Já a retro-alimentação negativa atua balanceando, moderando, ou atenuando as influências sobre o sistema, mantendo a estabilidade deste sistema.

Em condições longe do equilíbrio, as relações não lineares prevalecem, e o sistema torna-se extremamente sensível às influências externas. Assim, pequenos estímulos podem levar a grandes e devastadores efeitos [Nicolis e Prigogine (1989)]. Isto faz com que todo o sistema se reorganize. Parte deste processo surge como resultado dos mecanismos de realimentação de caráter positivo. "Em condições longe do equilíbrio, percebe-se que perturbações ou flutuações muito pequenas podem ser amplificadas em ondas que irão desintegrar o sistema"[Nicolis e Prigogine (1989)].

Uma razão para as intervenções que criam condições longe-do-equilíbrio, pode ser a inoperância dos processos existentes de realimentação. Este pode ser o caso quando os processos de realimentação negativa, que até então eram aptos a ajustar ou influenciar o comportamento do sistema, já não podem produzir os resultados desejados. Quando esforços destinados a aprimorar o desempenho do sistema falham continuamente, e quando as mudanças incrementais já não são eficazes, os sistemas podem "recorrer" a intervenções essenciais num esforço para produzir uma mudança radical. Entretanto, estas intervenções podem falhar também, e o sistema pode tornar-se constrangido em um ciclo constante de reestruturações ineficazes. Em algumas situações, tais falhas são resultados do uso constante dos mecanismos de ajuste baseados em ciclos de realimentação negativa que operaram no passado.

Ao mesmo tempo, face um ambiente turbulento, o ecossistema inteiro pode mudar, e nestes casos os novos comportamentos não podem ser mais uma mera extrapolação das experiências passadas. A emergência de novas formas de comportamento e de novas estruturas

pode ser necessária, ao mesmo tempo em que estas podem depender dos novos processos de realimentação positiva, ou estabelecerem-se completamente a partir destes.

A co-evolução, dentro do sistema, pode também depender da realimentação das influências recíprocas entre entidades. Contudo, resta saber como o grau de conectividade e de realimentação influência a co-evolução. Como pode a estrutura de um ecossistema afetar a co-evolução? Kauffman ressalta que a estrutura de um ecossistema governa a co-evolução [Kauffman (1993)]. Esta assertiva é baseada em simulações computacionais. Desta forma os processos de realimentação podem conseqüentemente ter uma influência no grau de conectividade (em todos os níveis), influenciando a estrutura do ecossistema, e nos processos de co-evolução.

2.1.6 Causalidade

Segundo Morin [Morin (1986)] a óptica sob a qual a ciência clássica vê o mundo é caracterizada pela premissa onde *"em toda parte, sempre, nas mesmas condições, as mesmas causas produzem sempre os mesmos efeitos"*. Esta expressão formal denota o princípio da predicabilidade: sendo conhecido o efeito E_{f_1} de uma causa C_1 e se é conhecida uma outra causa C_2 que seja igual a C_1, então pode-se prever que C_2 terá um efeito E_{f_2} igual a E_{f_1}. Por outro lado, se forem conhecidos dois efeitos E_{f_2} e E_{f_1} e sabendo que E_{f_1} foi causado por C_1, então que a causa de E_{f_2} (C_2) é igual à causa C_1. Como corolário desta expressão temos a assertiva que:

"causas distintas levam a efeitos distintos".

Por esta premissa, se duas situações são inicialmente distintas, elas permanecerão distintas ao longo de suas evoluções futuras, e deveriam ter sido distintas durante toda sua evolução prévia [Morin (1986)]. Ou seja:

> "quando uma causa varia, ou é trocada por uma causa diferente daquela em qualquer aspecto, o efeito a ela associado irá variar".

Com relação às definições apresentadas, se o termo "igual" for tomado com o sentido de "idêntico" tem-se o resultado tautológico em que uma causa é idêntica a ela mesma e então os seus efeitos devem ser idênticos a si mesmos. Desta forma, se um dado estado inicial for cambiado por um outro estado que não seja idêntico ao seu antecessor, então os estados resultantes não serão idênticos.

Entretanto, na realidade observável o termo "igual" só pode ser entendido como "similar": duas causas ou eventos usualmente são diferentes de algum modo. A partir do advento da física das partículas e dos posteriores progressos no campo dos sistemas dinâmicos não lineares, foi possível observar que os sistemas dinâmicos são usualmente caóticos[1] [Ueda (1979),

[1] De acordo com a Teogonia (A origem dos Deuses) de Hesíodo, Caos era o Nada a partir do qual os primeiros seres da existência apareceram. Estes primeiros seres, descritos como "filhos do Caos"sozinho, eram Gaia (a Terra), Tártaro (o Submundo), Eros (o amor sexual), Érebo (a escuridão do alma) e Nix (a escuridão da noite). Destes seres e da primeira geração de seres criados a partir deles, Hesíodo estabelece a linhagem dos deuses da antiga Grécia.

Bak et al. (1988)]. Na Matemática e na Física, a Teoria do Caos descreve o comportamento de sistemas dinâmicos não lineares que, sob condições específicas, exibem dinâmicas sensíveis às condições iniciais. Como resultado desta sensibilidade, o comportamento destes sistemas parece ser aleatório. Isto acontece mesmo no caso destes sistemas serem determinísticos, no sentido em que o estado futuro destes sistemas pode ser definido por suas condições iniciais, não havendo elementos aleatórios envolvidos. Neste sentido, mesmo as diferenças mais ínfimas que venham a ocorrer nos estados iniciais de um evento podem levar a diferenças enormes nos estados finais uma vez que, a rigor, as possíveis similaridades entre os estados iniciais foram perdidas.

Nestes casos, não se pode predizer com exatidão como o sistema se comportará. Na verdade pode-se estimar, com maior ou menor grau de precisão, as probabilidades dos eventos que se seguirão em função da precisão com que se conhece as condições de contorno do seu estado inicial. Quanto mais instável for a situação de equilíbrio inicial deste conjunto mais sensível ele será às variações de sua vizinhança e maior será o grau de variação do comportamento deste sistema às variações do ambiente com o qual ele interage em maior ou menor grau. Mesmo que macroscopicamente o comportamento deste sistema possa mostrar-se aleatório, de fato ele estará seguindo o princípio da causalidade: a mais leve diferença nas condições iniciais leva a resultados diferentes. O princípio da causalidade, na prática, somente tem sentido quando a distinção ou similaridade das condições de contorno sob as quais o fenômeno se desenvolve pode ser feita de forma macroscópica, sendo possível ignorar as diferenças microscópicas.

2.1.7 Auto-Organização

A auto organização é um processo de evolução onde os efeitos do ambiente sobre o sistema são mínimos, ou seja, onde o desenvolvimento de novas e complexas estruturas tem sua origem primeira no próprio sistema. O fenômeno da auto organização pode ser entendido com base nos mesmos processos evolutivos onde a variação e a seleção natural são orientadas pelas variações ambientais. O processo de auto organização é normalmente disparado por processos de variações internas, usualmente denominadas "flutuações" ou "ruídos". O fato de estes processos produzirem uma nova configuração sistêmica de forma ordenada e seletiva tem sido denominado como o princípio da "*order from noise*" de Heinz von Foerster ou mecanismo da "*order through fluctuations*" de Ilya Prigogine.

O incremento de organização em um sistema pode ser mensurado mais objetivamente como um decréscimo de sua entropia. Segundo Morin [Morin (1986)] os sistemas fechados e as organizações não ativas só podem evoluir no sentido do crescimento de sua entropia. Entretanto, em um sistema ativo ou "produtor-de-si"' [Morin (1986)] o seu trabalho ininterrupto mantém o seu nível de entropia constante enquanto tal sistema perdurar. Tal equilíbrio entrópico apareceria como o balanceamento entre a tendência natural de crescimento do grau de entropia deste sistema e a auto reprodução constante deste mesmo sistema. Este grau de entropia negativa que o sistema auto organizador conhece como decorrência dos trabalhos efetuados pelos seus mecanismos auto reguladores é denominado Neguentropia . Tal conceito

foi apresentado por Schröedinger [Schrödinger (1944)], e como grandeza assume, nas organizações ativas, um papel antagônico frente à tendência natural dos sistemas de evoluírem para estados de entropia crescente. Entropia e Neguentropia, embora constituindo o caráter positivo e negativo de uma mesma grandeza, correspondem a processos antagônicos do ponto de vista da organização. Entretanto os sistemas vivos [Morin (1986)] só são capazes de produzir neguentropia às custas da obtenção de energia e de informação do meio em que se encontram. Desta forma o sistema consegue aumentar o seu grau de organização interna e, simultaneamente, reduzir o grau de incerteza quanto ao seu estado atual. Segundo Morin [Morin (1986)] o sistema negentrópica, para se perpetuar, necessita manter e incrementar o seu nível interno de informações.

Um sistema auto organizante que tenha a capacidade de decrementar o seu nível de entropia interna deve necessariamente (como conseqüência da segunda lei da termodinâmica) exportar (ou dissipar) tal entropia para o ambiente em que se encontra, tal como notado por von Foerster e por Prigogine (op. cit.). Tal sistema que possua a capacidade de exportar entropia de forma contínua para o seu ambiente foi denominada por Prigogine de "Estruturas Dissipativas".

As estruturas auto organizantes são usualmente associadas com fenômenos mais complexos, de caráter não linear, ao contrário daquelas imbuídas de processos relativamente simples de difusão de calor. Todos atributos sistêmicos intrinsecamente associados à não linearidade (comportamento quase-caótico, sensibilidade às condições de contorno dos estados iniciais e intermediários dos processos, estruturação dissipativa, dentre outros) podem ser entendidos pela interação entre os ciclos de retroalimentação positiva e negativa onde algumas variações tendem a reforçá-los e outras tendem a mitigá-las. Ambos tipos de retroalimentação atuam como alimentadores do processo de seleção natural: as retroalimentações positivas tendem a aumentar o número de configurações possíveis para o sistema, retroalimentações negativas tendem a estabilizar as variações possíveis de configurações. Ambos proporcionam configurações que apresentam vantagens seletivas quando comparadas às configurações concorrentes pelos mesmos recursos. As interações entre estas configurações, onde as variações podem ser reforçadas em algumas direções e reduzidas em outras, podem criar forma intrincados e imprevisíveis (caos), as quais podem desenvolver-se rapidamente em direção a uma configuração estável (atrator).

2.1.8 Abordando a Estrutura dos Sistemas Complexos

Ao longo da última década, o estudo dos sistemas complexos tem sido objeto de renovada atenção por parte de parte da comunidade acadêmica internacional. Tal como ocorreu nos primeiros anos da IA[2], alguns estudiosos dos sistemas complexos têm, há tempos, feito promessas e levantado expectativas difíceis de serem atendidas.

Todavia, ao longo destes anos muitas descobertas interessantes têm sido feitas sobre os aspectos estruturais dos sistemas complexos. Isto fica evidente não só pelo número de pu-

[2]Inteligência Artificial

blicações devotadas a este campo, mas também pelas mudanças de postura observadas nas universidades. Em todos os campos do conhecimento, relativos às ciências voltadas para os sistemas sociais, naturais ou artificiais, a modelagem computacional dos sistemas em estudo tornou-se amplamente aceita como uma atividade científica válida.

2.2 Proteínas como Sistemas Complexos

As proteínas, no campo da biologia molecular, aparecem como sendo as candidatas naturais para serem vistas como um arranjo de elementos que apresenta propriedades sistêmicas. Estas, ao mesmo tempo, apresentam algumas vantagens para estudos desta natureza [Frauenfelder (1994)]. Elas têm o tamanho certo, pequenas o suficiente para ser teoricamente modeladas e grandes o bastante para serem verdadeiramente complexas. Elas podem ser modificadas com precisão por meio de engenharia genética ao mesmo tempo em que sondas espectroscópicas já são parte natural de suas estruturas ou podem ser aí introduzidas [Frauenfelder (1994)]. Buscando evidenciar a pertinência da visão das proteínas como sistemas, esta seção apresenta uma interpretação das propriedades apresentadas pelas proteínas em geral e a relação destas com as propriedades dos sistemas complexos, já discutidas na seção anterior.

As proteínas são polímeros sintetizados com base em um conjunto restrito de 20 L-aminoácidos. Estes polímeros, de forma geral, passam por um processo de colapso e estabilizam-se tridimensionalmente sempre da mesma forma, mantidas as condições ambientais. Fazendo uma analogia das proteínas com os sistemas complexos, os aminoácidos (e seus respectivos átomos) quais podem ser vistos como os elementos constituíntes estes sistemas. Ao mesmo tempo, estes aminoácidos (elementos) estabelecem entre si, interações de diferente natureza como: ligações covalentes; interações eletrostáticas e de van der Waals; e ligações de hidrogênio. Dentro desta analogia, tanto a forma tridimensional de uma proteína, quanto seus atributos, podem ser vistos como propriedades emergentes, visto que estes não podem ser inferidos a partir da mera observação da sua seqüência de residuos. Ademais, vale ressaltar que a estabilidade tridimensional das proteínas, deve-se em grande parte às interações de caráter não covalentes que mantém espacialmente próximos os átomos e resíduos que, na seqüência linear, encontram-se bem afastados.

De pronto, percebe-se que existe, na proteína enovelada, uma grande rede de interações entre os diferentes átomos que constituem esta proteína. Constata-se o grande número de átomos existentes em uma proteína ($\approx$ 3600 átomos em média em uma globina típica) e o grande número de interações que estes átomos estabelecem entre si. A existência deste grande número de elementos e interações, além da diversidade da natureza das interações entre átomos, permite inferir que deve ocorrer, nesta situação, a emergência de comportamentos complexos. Indo um pouco mais além, estando as proteínas imersas no lumem citoplasmático, estas defrontam-se com um ambiente extremamente conturbado e variável. Nestas condições,

estas proteínas ao interagirem com o seu ambiente possivelmente irão apresentar comportamentos que em princípio não podem ser deduzidos a partir do conhecimento dos aminoácidos que irão compor esta proteína.

Analisando as hipóteses citadas, é inquestionável o fato de que, até esta data, ainda não é possível deduzir qual será a forma topológica que uma proteína irá apresentar, após enovelada, a partir da seqüência de resíduos que a compõem. Por outro lado, com relação às influências ambientais, a hemoglobina humana mostra ser um exemplo interessante. A capacidade de uma hemoglobina de transportar moléculas está relacionada não só à presença do grupo **HEME**, mas também á sua sensibilidade aos diferentes pHs existentes tanto nos alvéolos pulmonares quanto nos capilares do sistema circulatório. Estes dois exemplos, são uma ínfima amostra da diversidade de formas e da diversidade de comportamentos apresentados pelas proteínas.

Proteínas guardam ainda outras propriedades análogas àquelas associadas aos sistemas complexos:

- Conectividade e interdependência das partes;
- São hierarquicamente organizadas;
- Apresenta um fenômeno de co-evolução dos seus constituintes;
- Apresentam um variado conjunto de conformações estáveis;
- Os estados futuros que podem ser alcançados por uma proteína, dependem do estado presente;
- O resultado combinado das ações de uma proteína e das respostas ambientais, influencia sua próxima conformação;
- Como conseqüência da mutação em um ou mais resíduos, uma proteína pode apresentar novas organizações e/ou funções.

Guardando ainda uma grande aderência às propriedades apresentadas pelos sistemas complexos em geral, também nas proteínas a conectividade e a interdependência de átomos e resíduos faz com que perturbações induzidas por uma entidade do ambiente orgânico, sobre um átomo da proteína têm o potencial de induzir alterações por toda a estrutura da proteína. Da mesma forma, nas proteínas tais efeitos induzidos não são os mesmos para todos os átomos/resíduos das proteínas e nem têm os mesmos impactos sistêmicos. Assim, uma perturbação induzida no grupo **HEME** de uma mioglobina, induz alterações em toda a proteína que irão confinar ou liberar a molécula que está sendo transportada, em função das condições ambientais em que esta mioglobina se encontra. Da mesma forma, o estado presente de uma proteína tanto quanto o próximo estado alcançável por esta, são determinados pela sua trajetória ao longo do seu espaço de fase (história).

Outro indício significativo, remete ao postulado enunciado por Bertalanffy [von Bertalanffy (1975)]. Tanto um peptídeo quanto uma proteína são, em essência, polímeros de

aminoácidos. Contudo, estes grupos são funcionalmente bem distintos. Enquanto o comprimento dos peptídeos varia na faixa $8 \lesssim l \lesssim 30$ resíduos, uma proteína apresenta comprimento $l \gtrsim 100$ resíduos. Contudo uma proteína não é concebida como sendo um peptídeo que foi crescendo de forma gradual. Notavelmente, as propriedades que caracterizam uma cadeia polipeptídica como sendo uma proteína parecem emergir a partir de um limiar crítico no comprimento da cadeia como se de fato elas emergissem instantaneamente e não gradualmente, com o incremento monótono da seqüência. A partir deste ponto crítico, parece haver uma "mudança de fase" da cadeia polipeptídica.

Ao mesmo tempo, uma proteína apresenta outras propriedades sistêmicas que podem ser ressaltadas:

I- Para um dado número de aminoácidos, uma proteína pode assumir um grande número de conformações aproximadamente isoenergéticas;

II- Para as proteínas, simetrias estruturais podem ou não ser importantes dependendo da natureza dos processos com os quais elas estejam envolvidas.

Contudo, o número de estados alcançáveis no espaço de fase de uma proteína pode ser muito grande. Neste caso dois problemas emergem de imediato: a elucidação da organização do espaço de conformações e a investigação das transições desta proteína neste seu espaço de fases, sendo que tais transições correspondem às variações na estrutura deste sistema. Estes problemas ainda esperam por uma solução.

Já com relação à estrutura terciária das proteínas, tem-se a óbvia constatação que as forças que mantêm os aminoácidos da estrutura primária juntos são devidas à natureza covalente de suas ligações, as quais são quimicamente muito fortes. Tais ligações não são quebradas por flutuações térmicas. Ao mesmo tempo, a estrutura terciária é estabilizada por fracas forças decorrentes das ligações não-covalentes como ligações de hidrogênio, interações eletrostáticas e forças de van der Waals, as quais podem apresentar flutuações.

No que tange as questões de simetria, tudo indica que a Natureza "prefere" selecionar sistemas que apresentam simetria, por razões de economia e controle [Wolynes (1996), Goodsell e Olson (2000)]. Tanto Wolynes [Wolynes (1996)] quanto Olson [Goodsell e Olson (2000)] constatam a predominância da simetria estrutural no nível molecular e especulam que tal tendência de simetria se estenderia para os níveis celular e orgânico, pelas mesmas razões. Soluções desta ordem seriam preferencialmente selecionadas como conseqüência, em alguns casos, da predominante tendência para conformações de mínima energia, e em outros casos, devido à predominância das pressões funcionais impostas pelo processo evolutivo [Goodsell e Olson (2000)]. Ainda segundo Olson, a simetria estrutural exerceria um papel central na regulação alostérica das proteínas. Para tanto, a regulação alostérica requereria uma geometria molecular que permitisse, no caso das estruturas multiméricas, a passagem de mensagens de uma subunidade para as demais. Tanto as estruturas monoméricas quanto multiméricas, a regulação alostérica permitiria a adequação do comportamento funcional das proteínas aos diversos ambientes aos quais elas podem estar expostas [Goodsell e Olson (2000)].

Contudo, várias funções bioquímicas limitariam a simetria estrutural das proteínas como resultado de um "cabo de guerra" evolucionário [Goodsell e Olson (2000)] que leva um nicho funcional a um estado "ótimo". Como exemplo, processos que envolvem movimentação direcional limitam a simetria do maquinário molecular. Polimerases e ribossomos realizam processos direcionados de forma assimétrica o que as obriga a apresentar assimetria estrutural [Goodsell e Olson (2000)]. Ao mesmo tempo, proteínas monoméricas seriam fortemente assimétricas [Goodsell e Olson (2000)]. A assimetria dos L-aminoácidos dá origem a uma tendência preferencial na formação de hélices e folhas-β. O empacotamento destas unidades estruturais secundárias, as quais apresentam um pequeno número de modos preferenciais, dá origem a estruturas enoveladas de forma assimétrica. Casos com alto grau de simetria interna, tal como β-barril, são relativamente raros. Contudo, as assimetrias impostas pelas limitações dos L-aminoácidos parecem não se estender até o nível estrutural quaternário [Goodsell e Olson (2000)]. A assimetria, é o elemento chave para as interações alostéricas nas quais as unidades individuais podem adotar uma das diferentes conformações alternativas. Ao mesmo tempo, as quebras de simetria propiciam melhor adaptabilidade dos sistemas moleculares às pressões evolutivas devido à quebra das simetrias sinônimas, o que ocorre por meio de mutações e recombinações [Vera e Waelbroeck (1996), Itoh e Sasai (2006)].

Estes imbricados arranjos estruturais simétricos e assimétricos acabam por influenciar os aspectos dinâmicos e alostéricos das proteínas de forma não óbvia e não trivial. Desta forma, os aspectos estruturais de uma proteína determinam, em grande parte, sua dinâmica e alosteria. As implicações vão mais além. Uma vez que os processos dinâmicos e alostéricos são quase tautologicamente ligados [Kern e Zuiderweg (2003)], mudanças na estrutura de uma proteína implicam em alterações não óbvias nas propriedades dinâmicas das diferentes conformações do espaço de fases alostéricas da proteína o que influencia os valores de energia livre dos acoplamentos alostéricos por meio de efeitos entrópicos [Cooper e Dryden (1984), Kern e Zuiderweg (2003)].

Mesmo que por mais de quarenta anos, muitos esforços tenham sido engajados no entendimento dos mecanismos pelos quais informações são transmitidas entre pontos distantes dentro de uma proteína, este mecanismo ainda é pouco conhecido[Swain e Gierasch (2006)]. Tal processo de transdução de sinais através da estrutura das proteínas é a base da alostérica das proteínas. Este fenômeno determina, em grande parte, como, por exemplo, a entrada de um ligante em um sítio pode alterar a afinidade ou a eficiência catalítica em outro sítio distante na mesma proteína. Ao mesmo tempo em que se reconhece o intrínseco vínculo deste fenômeno com a estrutura das proteínas [Gunasekaran et al. (2004), Hardy e Wells (2004), Swain e Gierasch (2006)], praticamente nenhum trabalho se deteve ainda em explorar tal relação. Uma vez que esta propriedade possa ser manipulada de forma efetiva, a mesma poderá ser explorada para uma variedade de propósitos como ser aplicada para o desenvolvimento de biosensores e no projeto de drogas.

Pelo que foi apresentado, acreditamos ter muitos indícios que mostram a pertinência da abordagem sistêmica das proteínas. Na seqüência deste trabalho, com base nesta abordagem, alguns estudos propedêuticos serão conduzidos, no sentido de realizar uma avaliação inicial

das propriedades sistêmicas das proteínas.

Capítulo 3

Redes Complexas

"Puisqu'on ne peut être universel et savoir tout ce qui peut se savoir sur tout, il faut savoir peu de tout. Car il est bien plus beau de savoir quelque chose de tout que de savoir tout d'une chose ; cette universalité est la plus belle."

Pascal - Penseés

Nos anos recentes trabalhos significativos têm aplicado os princípios de redes para modelar sistemas complexos tais como processos epidemiológicos, processos sociais, redes metabólicas, dispositivos microeletrônicos, etc. Para redes de dezenas ou centenas de vértices, o esforço de elaborar um modelo representativo é algo relativamente fácil de ser conduzido. Ao mesmo tempo, é possível responder a outras questões específicas acerca da natureza da estrutura de tal rede por meio do simples exame de seu desenho. Desta forma alguns avanços significativos já foram possíveis no sentido de prover respostas relacionadas à caracterização e modelagem de estruturas de redes [Newman (2003c)]. Por outro lado, estudos relativos às influências da estrutura sobre o comportamento de tais redes mostram ser ainda incipientes.

A ubiqüidade das redes complexas nos fenômenos até agora observados na natureza conduz a uma série de questões relevantes sobre como a estrutura das redes facilita ou restringe o comportamento dinâmico das mesmas, questões estas que têm sido negligenciadas nas pesquisas conduzidas no seio das disciplinas tradicionais. Desta forma seria possível argüir como os arranjos sociais são capazes de mediar a propagação de doenças? Ou como as falhas em cascata podem se propagar em uma malha de transmissão de energia elétrica ou de telecomunicações em uma grande área geográfica? Qual a mais eficiente e robusta arquitetura para um sistema em particular que opera em um ambiente com grande mutabilidade e incerteza? Problemas desta ordem são corriqueiros mas relevantes ao mesmo tempo demandando respostas e soluções adequadas.

Tradicionalmente a modelagem de sistemas e processos físicos e não físicos tem sido feita assumindo-se, implicitamente, que as formas de interação entre os elementos individuais, no contexto dos sistemas ou processos subjacentes, seriam suficientemente bem representadas considerando, a cada momento, somente as interações par a par. Entretanto, no final da década de 1950 os matemáticos Erdös e Rényi (ER), descreveram uma rede com topologia complexa por meio do uso de grafos randômicos [Erdös e Rényi (1959)]. Este trabalho lançou

os fundamentos da teoria das redes randômicas seguido por uma série de estudos dos anos que se seguiram até a atualidade [Wang e Chen (2003)]. Mesmo que a intuição mostrasse que muitas das redes complexas do mundo real não eram nem completamente randômicas nem completamente regulares, o modelo ER era o único até então capaz de prover uma abordagem sensível e rigorosa o suficiente para permitir pensar sobre esta classe de problemas. Tal modelo predominou até recentemente, tanto pela carência de poder computacional suficiente para testar exaustivamente este modelo quanto pela ausência de informações sobre a topologia de redes de grande escala do mundo real [Wang e Chen (2003)].

A partir da última década do século passado, com a padronização e a produção em grande escala dos componentes de computadores, passou a ser possível a acumulação de grandes volumes de dados e o processamento de tais dados em escalas de tempo e custo aceitáveis. Com isto dados relativos às redes de grande escala existentes no mundo real puderam ser sistematicamente acumulados e tratados. Isto viabilizou o interesse em descobrir propriedades genéricas de diferentes tipos de redes complexas. Decorrente destes esforços, as descobertas do efeito "*small-world*" e da propriedade de independência de escala apresentados por muitas das redes existentes no mundo real representaram um marco no estudo da estrutura e da dinâmica dos sistemas complexos.

Em 1998, objetivando descrever como se processa a transição de uma rede regular para uma rede randômica, Watts e Strogatz (WS) descreveram o efeito de "*small-world*" (mundo-pequeno) apresentado pelas redes do mundo real [Watts e Strogatz (1998)]. Este nome surge de um fenômeno muito comum onde após conhecer uma pessoa estranha, alguém acaba por ficar surpreso ao descobrir que ambos têm um conhecido em comum: "- Que mundo pequeno!". Ainda mais interessante é o princípio dos "seis graus de separação" sugerido por Milgram [Milgram (1967)]. Mesmo sendo controverso este último ponto, o fenômeno do "*small-world*" tem sido presenciado em várias redes do mundo real. Um aspecto proeminente tanto no modelo ER quanto no modelo WS é a distribuição de conectividade da rede que após atingir um pico, decai exponencialmente. Tais redes são usualmente denominadas "redes exponenciais" ou "redes homogêneas", dado que a maioria dos nodos (ou vértices) tende a ter o mesmo número de conexões.

Outra significante descoberta foi a observação que muitas das redes complexas de grande tamanho apresentam um arranjo que é livre de escala, o que significa dizer que a distribuição de conectividade destas redes segue uma lei de potencia que é invariante com o tamanho desta rede [Albert e Barabasi (1999), Barabasi et al. (1999)]. Diferente de uma rede exponencial, a rede livre de escala é não homogênea por natureza, sendo que muitos dos nodos apresentam poucas conexões e poucos nodos apresentam muitas conexões. Assim, a descoberta destas propriedades das redes complexas permitiu avanços significativos na teoria das redes.

Barabasi e Albert [Albert e Barabasi (1999)] inauguram com seu artigo sobre estrutura de redes um momento na ciência onde as redes ditas complexas passam a ser adotadas como modelo básico para a análise de uma série de fenômenos em diferentes ramos do conhecimento que variam da Física à Psicologia. Este trabalho passa a ser referência para o estudo e descrição estática dos sistemas ditos complexos capazes de ser modelados com o uso de

grafos.

Dentre os aspectos característicos das redes complexas, talvez o mais referenciado seja aquele onde o grau de distribuição do numero de arestas por vértice, para muitos sistemas do mundo real modelados como grafos, seguem uma lei de potencia na forma $P(k) \sim k^{-\gamma}$, em contraste com a distribuição de Poisson esperada, caso as interações entre tais vértices ocorressem de forma aleatória. Para explicar este achado, foi necessário ir além da modelagem topológica da rede e endereçar o problema de como o grafo é formado e seu processo evolutivo.

Em particular, Barabasi e Albert propuseram um modelo onde uma rede cresce a partir de uma rede limitada formada por um numero reduzido de elementos agindo como estrutura semente. Nesta estrutura novos elementos são então adicionados e conectados àqueles já existentes segundo o princípio de conexão preferencial. De acordo com este princípio, a probabilidade de um novo vértice se conectar a um já existente é diretamente proporcional ao grau daquele vértice. Assim, poucos vértices muito conectados emergiriam como elementos atratores dentro da rede que se forma, os quais são referenciados como "*hubs*". Simulações numéricas posteriores confirmaram que estas estruturas evoluem para uma topologia invariante em escala que pode ser descrita por uma lei de potencia onde freqüentemente o expoente aparece com valores próximos a -3.

Existia então uma certa tentação em proclamar a universalidade desta relação e deste expoente, mas mais recentemente estudos têm concluído que este não é bem o caso [Goh et al. (2002)]. A evolução do escalonamento da arquitetura *livre de escala* apresenta além de outras propriedades aquela de assegurar robustez à rede contra falhas aleatórias dos nodos [Albert e Barabasi (2000)]. A robustez dinâmica destas redes complexas pode ser entendida como uma conseqüência direta da topologia livre de escala [Toroczkai e Bassler (2004)]. Ao que parece, a topologia livre de escala é resultado de agregações específicas direcionadas por processos evolucionários de caráter seletivo [Aldana e Cluzel (2003)] e parece ser também um padrão subjacente freqüente para muitos tipos de sistemas complexos encontrados na natureza.

Os princípios derivados da elucidação da forma destas redes têm sido aplicados aos domínios biológicos revelando que as redes de interações moleculares envolvidas nos processos de regulação celular, metabólica e transcricionais apresentam comportamentos descritos pelos modelos de redes complexas. Esforços têm sido feitos no sentido de aplicar os conceitos de redes complexas para estudar as estruturas de proteínas e nas transições de estado nos processos de enovelamento destas.

O propósito deste capítulo é apresentar as noções e os conceitos básicos relativos às redes complexas, enfatizando as relações existentes entre a topologia e o comportamento dinâmico destas redes.

3.1 Conceitos e Modelos de Redes Complexas

Todas as topologias de redes discutidas na literatura variam de estruturas completamente regulares a arranjos quase randômicos. Diferentes autores [Strogatz (2001), Newman e Watts (1999), Newman (2003c), Wang e Chen (2003)] concentram suas discussões em uma gama invariante de modelos topológicos: topologia regular (figuras 3.1a1 e 3.1a2), topologia randômica (figuras 3.1b), topologia "*small-world*" (figura 3.1c) e topologia livre de escala (figura 3.1d).

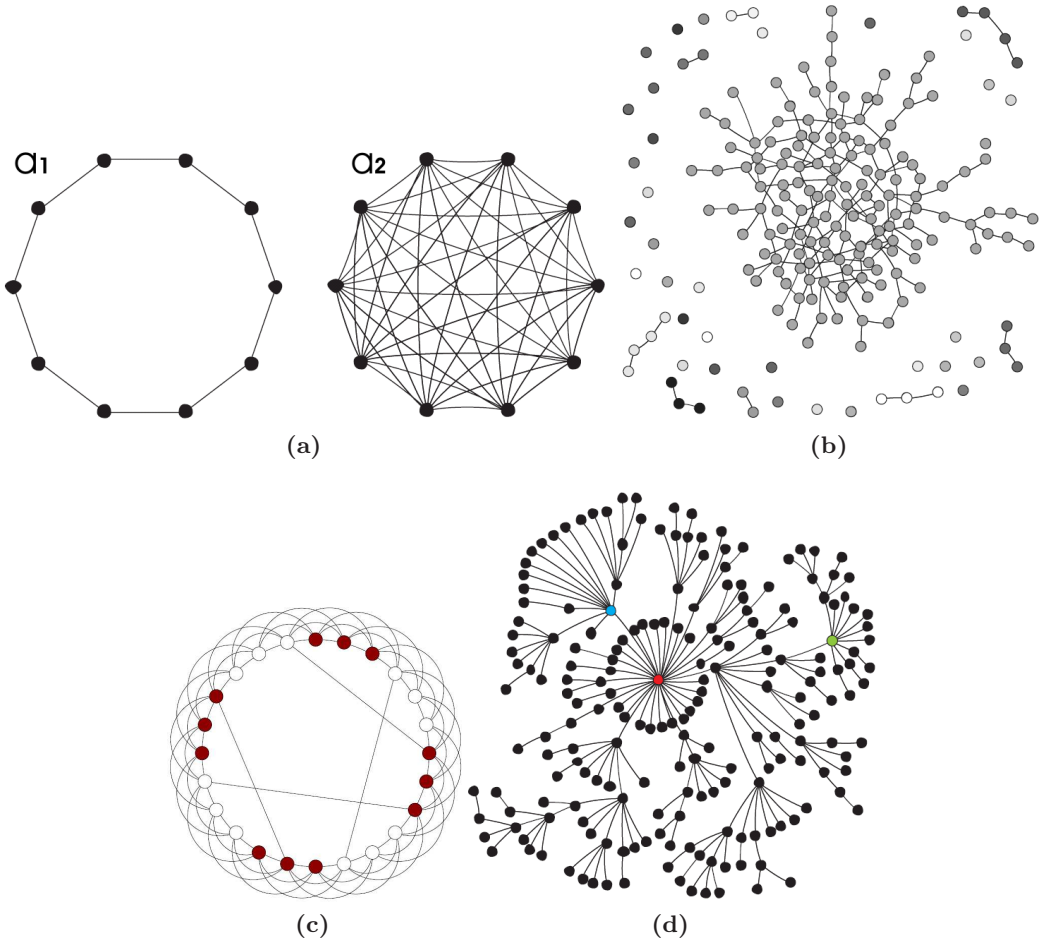


Figura 3.1 – *Exemplo de modelos topolólogicos de redes complexas apresentados na literatura*

Estes modelos simples permitem focar na complexidade causada pela dinâmica não linear apresentada pelos nodos, sem carregar a análise com qualquer outra complexidade adicional apresentada pela estrutura da rede [Strogatz (2001), Newman (2003c)].

Esquecendo nesta abordagem, os aspectos dinâmicos das redes, é possível colocar a atenção na descrição dos aspectos estruturais das arquiteturas complexas. A seguir serão discutidos os aspectos teóricos essenciais de cada um destes modelos, buscando mostrar que características os diferenciam e sua relevância para este estudo.

3.1.1 Redes Regulares

O modelo mais simples de rede regular é a rede com estrutura unidimensional onde os nodos são ligados aos seus vizinhos mais próximos, tal como uma fila de crianças unidas pelas mãos. Esta pode ser modelada por um grafo regular no qual todos os nodos estão unidos aos seus vizinhos mais próximos. Uma estrutura do tipo "vizinho mais próximo" pode apresentar condição periódica consistindo de N nodos arranjados em anel, onde cada nodo i é adjacente aos seus vizinhos imediatos, para $i = 1, 2, 3, \ldots, K/2$ onde K é inteiro e par (figura 3.2). À medida que o valor de K cresce, o coeficiente de aglomeração desta rede de vizinhos imediatos tende para $C = 3/4$.

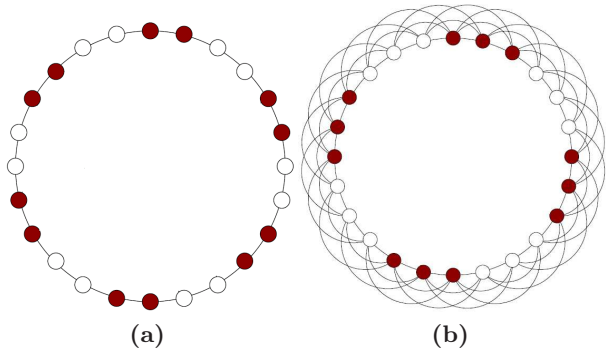


Figura 3.2 *– Estruturas regulares com diferentes valores de interação com sua vizinhança - K : (a) para K = 1, (b) para K = 2*

Quando $K = K/2$ a rede se torna globalmente conectada. Intuitivamente, esta rede apresenta a menor distância média entre nodos e o maior coeficiente de aglomeração. Apesar do modelo de rede globalmente conectada possuir os atributos de "*small-world*" e grande aglomeração apresentados por muitas redes do mundo real, ele não pode ser caracterizado como tal. Ao contrário, seu valor de distância média entre nodos é proporcionalmente grande e tende para o infinito quando N tende para o infinito. É fácil perceber que uma rede globalmente conectada, com N nodos, tem $N(N-1)/2$ arestas enquanto muitas das redes reais de grande escala apresentam-se esparsas, não tendendo a ser plenamente conectadas e com um número de arestas geralmente na ordem de N e não na ordem de N^2.

Segundo Wang e Chen [Wang e Chen (2003)], estas características ajudariam explicar porquê os processos dinâmicos que exigem coordenação global seriam difíceis de serem estabilizados (ou sincronizados) em tal rede formada somente com conexões locais. Contudo, o caso particular de rede em estrela é o caso onde uma rede regular apresenta-se esparsa, com nodos com coeficiente de aglomeração com alto valor relativo e caminho médio entre nodos pequeno. Nesta topologia, existe um nodo central e cada um dos outros $N-1$ nodos ligam-se somente àquele nodo mas não apresentam ligações mútuas. Neste arranjo topológico a distância média entre nodos $L \to 2$ e coeficiente de aglomeração $C \to 1$ à medida que $N \to \infty$ [Wang e Chen (2003)]. A rede em estrela captura bem os atributos de "*small-world*", de ser esparsa e alta aglomeração, além de outras propriedades interessantes do mundo real.

3.1.2 Redes Randômicas

Erdös e Rényi [Erdös e Rényi (1959)] estudaram como a topologia esperada de um grafo randômico, com N nodos e M arestas (figura 3.2b) variaria como função de M. Quando M é pequeno, o grafo aparece fragmentado em pequenos aglomerados de nodos, denominados *componentes*. À medida que M cresce, estes componentes crescem, primeiramente ligando nodos isolados e depois coalescendo com outros componentes. Uma transição de fase é perceptível quando $M = N/2$, onde vários aglomerados começam a se interligar espontaneamente para formar um único componente gigante. Para $M > N/2$ este componente gigante contém da ordem de N nodos (já que seu tamanho passa a variar linearmente com N à medida que $N \rightarrow \infty$, enquanto seus "rivais" mais próximos continuam variando seu tamanho em $O(log\ N)$ nodos. Em breve todos os componentes gigantes estarão conectados.

Uma das questões que podem ser levantadas quando se estuda sistemas com estruturas similares a um grafo randômico, é procurar saber a partir de que valores de probabilidade de conexão entre os nodos - p, uma propriedade particular deste grafo começa a ser observável. Erdös e Renyi [Erdös e Rényi (1959)] descobriram que propriedades importantes de um grafo randômico podem aparecer quase que subitamente.

Apresentando o problema de uma forma mais didática, consideremos o exemplo de um grafo ER randômico tal como ilustrado em [Wang e Chen (2003)]. Imaginemos um número grande ($N \gg 1$) de botões espalhados em uma mesa. Com a mesma probabilidade p amarramos cada um dos possíveis pares de botões com um pedaço de linha (figura 3.3). O resultado será uma estrutura similar a um grafo ER randômico com N nodos e $pN(N-1)/2$ arestas. Então pode-se perguntar: se um botão for levantado de mesa, quantos outros botões irão subir junto com ele? Erös e Renyi mostraram que a partir de um limiar crítico de probabilidade $p_c \sim (lnN)/N$, quase todo grafo randômico se mostra totalmente conectado. Da mesma forma, todos os outros botões da mesa irão ser levantados junto com o primeiro.

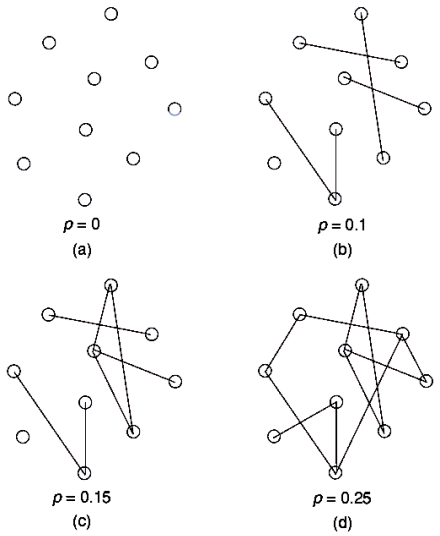


Figura 3.3 *– Evolução de um grafo randômico. Dado 10 nodos isolados em (a), cada par de nodos é conectado com probabilidade (b) p = 0.1,(c) p = 0.15 and (d) p = 0.25, respectivamente.*

O grau de conectividade médio de um grafo randômico é $\langle k \rangle = p(N-1) \sim pN$. Seja então L_{rand} a distância média entre nodos de uma rede randômica. É possível provar que aproximadamente $\langle k \rangle^{L_{rand}}$ nodos desta rede estão à distância L_{rand} uns dos outros. Portanto, tem-se que $N \sim \langle L \rangle^{L_{rand}}$ ou equivalentemente $L_{rand} \sim \ln N / \langle k \rangle$. Este crescimento logarítmico da distância média entre nodos como função do número de nodos da rede é típico do efeito de "*small-world*". Devido ao fato de $\ln N$ crescer lentamente com N, a distância média entre nodos permanece pequena mesmo em uma rede grande. Por outro lado, em uma rede completamente randômica, o coeficiente de aglomeração é $C = p = \langle k \rangle / N \ll 1$. Isto significa que mesmo uma grande rede randômica não apresenta valores de C consideráveis. De fato, para um valor grande de N, uma rede randômica apresenta uma distribuição de conectividade seguindo uma distribuição de Poisson.

3.1.3 Redes "*small-world*"

Apesar de Milgram ser freqüentemente associado à identificação do efeito "*small-world*" nas redes sociais [Milgram (1967)], a existência de tal efeito já havia sido especulada anteriormente pelo escritor húngaro Frigyes Karinthy e mais rigorosamente no trabalho dos matemáticos Pool e Köchen [*op.cit.* Newman (2003c)]. Atualmente este efeito tem sido estudado e verificado diretamente em um grande número de diferentes redes [Wang e Chen (2003), Newman (2003c)].

Como citado previamente, as redes regulares apresentam efeito de aglomeração apreciável, mas não exibem efeito "*small-world*" em geral. Por outro lado, grafos randômicos apresentam efeito "*small-world*", mas não apresentam aglomeração apreciável. Desta forma observa-se que tanto o modelo de rede regular como o modelo ER randômico falham em reproduzir algumas características importantes de algumas redes do mundo real. De fato, muitas das redes reais não são nem totalmente regulares nem totalmente randômicas. A realidade é que as pessoas conhecem seus vizinhos mas seu círculo de relacionamento não se restringe àqueles que vivem na casa ao lado. Por outro lado, os links entre as paginas Web na WWW certamente não foram feitos ao acaso, como o processo ER espera.

Objetivando descrever a transição de uma estrutura regular para uma estrutura randômica, Watts e Strogatz [Watts e Strogatz (1998)] apresentaram um modelo, referenciado como Modelo "*small-world*" de Watts e Strogatz - WS, cujo algoritimo de geração pode ser descrito como se segue:

1. Inicialmente cria-se uma rede de vizinhos mais próximos conectados, consistindo de N nodos arranjados em anel, onde cada nodo i está adjacente aos seus nodos vizinhos $i = 1, 2, \ldots, K/2$, com K sendo par;

2. Randomicamente cada aresta existente é redirecionada para outro nodo da rede com probabilidade p; variando p de tal forma que a transição entre a ordem ($p = 0$) e a aleatoriedade (p = 1) possa ser monitorada.

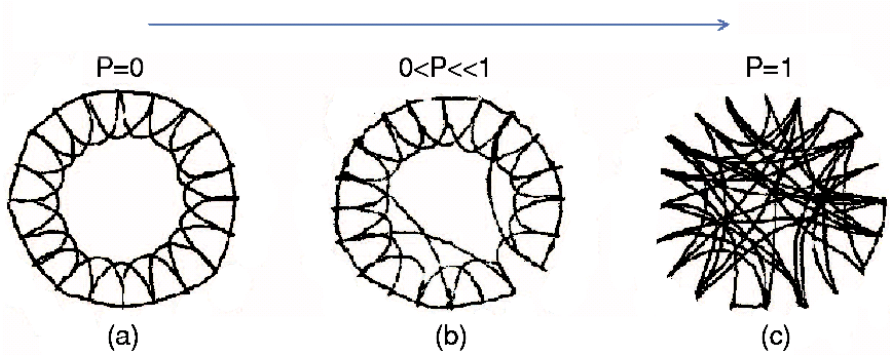


Figura 3.4 *– (a) Rede completamente regular com $K = 2$. (b) Rede "*small-world*". (c) Rede randômica.*

O rearranjo de nodos, no contexto do trabalho de Watts e Strogatz [Watts e Strogatz (1998)], significa deslocar uma ponta de uma aresta de um nodo para outro escolhido ao acaso, com probabilidade p, dentro da própria rede, com a restrição onde nenhum par de nodos pode ter mais que uma conexão entre eles e que nenhum nodo pode estar ligado a si mesmo. Este processo introduz $pNK/2$ arestas de ligação distante, as quais conectam nodos que de outra forma sempre seriam desconexos. O comportamentos tanto do coeficiente de aglomeração - $C(p)$, quanto da distância média entre nodos - $L(p)$, nas redes WS, podem ser considerados como funções da probabilidade - p, adotada para rearranjo dos nodos. Uma rede com arranjo regular ($p = 0$) mostra-se altamente aglomerada ($C(0) \cong 3/4$) mas apresenta uma distancia média entre nodos grande ($L(0) \cong N \gg 1$).

A seguinte definiçao de rede "*small-world*" pode ser adotada: *Se o número de nodos dentro de uma distância r de um vértice central típico, cresce exponencialmente com r, então o valor de l - distância média entre os vértices da rede, irá crescer na razão de* log n. O termo *"small-world effect"* pode então ser tomado com um sentido mais preciso: *Uma rede apresenta o efeito de "*small-world*" se o valor de l escala logaritimicamente ou de forma mais atenuada, com o grau de conectividade média da rede.* O escalonamento logarítmico pode ser provado para uma série de modelos de redes e, da mesma forma, tem sido observado em várias redes presentes no mundo real.

Observa-se que para valores pequenos de probabilidade de rearranjo – p, o coeficiente de aglomeração não difere muito do seu valor inicial ($C(p) \sim C(0)$), mas o valor da distância média entre nodos cai drasticamente para valores próximos aos apresentados por uma rede randômica ($L(p) \gg L(0)$) (Figura 3.5). Este processo é ilustrado na fig.3.7b.

Uma variação do modelo WS foi proposta por Newman e Watts [Newman e Watts (1999)], referida como modelo "*small-world*", no qual não existe a quebra de nenhuma conexão prévia existente entre os nodos vizinhos. Ao contrário, novas arestas são adicionadas, com probabilidade p, entre os possíveis pares de nodos da rede. Entretanto, os nodos continuam não podendo estabelecer mais que uma ligação por cada par nem ligar a si mesmos. Com valor de $p = 0$, o modelo NW reduz-se ao modelo regular original e com $p = 1$ ele se torna uma rede globalmente conectada.

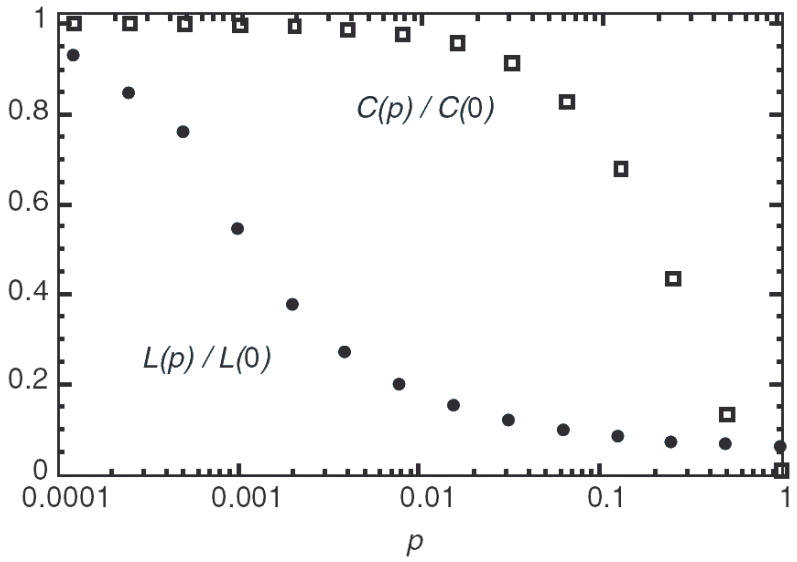


Figura 3.5 *– A distância média entre nodos e o coeficiente de aglomeração do modelo WS como uma função da probabilidade p [Watts e Strogatz (1998)]. Ambos estão com seus valores normalizados em função das redes regulares que lhes deram origem (p = 0).*

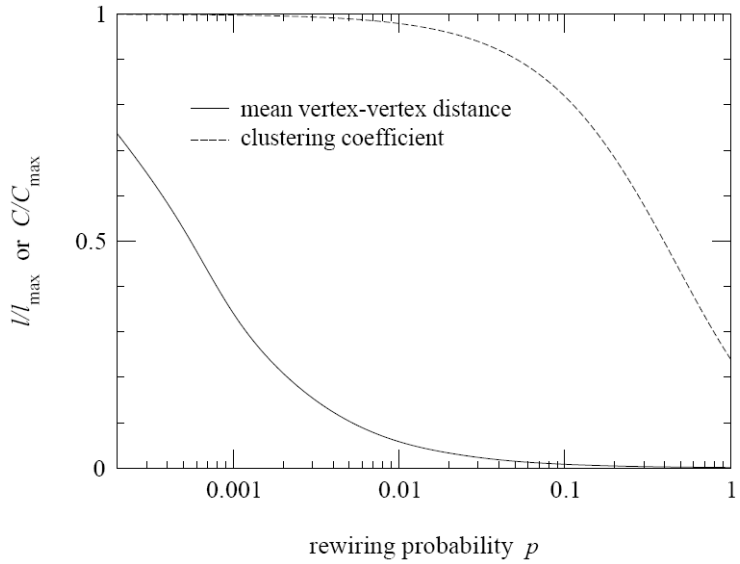


Figura 3.6 *– O coeficiente de aglomeração C e distância média entre vértices l no modelo "*small-world*" de Watts e Strogatz [Watts e Strogatz (1998)] como uma função da probabilidade de rearranjo das ligações p. Por conveniência, tanto C como l são divididos por seus valores máximos, o que ocorre quando p = 0. Entre os extremos p = 0 e p = 1, existe uma região na qual o índice de aglomeração é alto e a distância média entre vértices é simultaneamente pequena.*

Apesar das redes regulares e grafos randômicos serem idealizações úteis, a maioria das redes do mundo real encontram-se entre estes dois extremos de ordem e aleatoriedade [Strogatz (2001)]. Watts e Strogatz [Newman et al. (2001)] estudaram um modelo simples que pode ser ajustado nesta faixa intermediária: um arranjo regular onde as arestas originais são trocadas por outras geradas ao acaso com probabilidade $0 \leq p \leq 1$. Nesta estrutura, um pequeno rearranjo de arestas faz com que a rede passe a apresentar o fenômeno de "*small-world*",

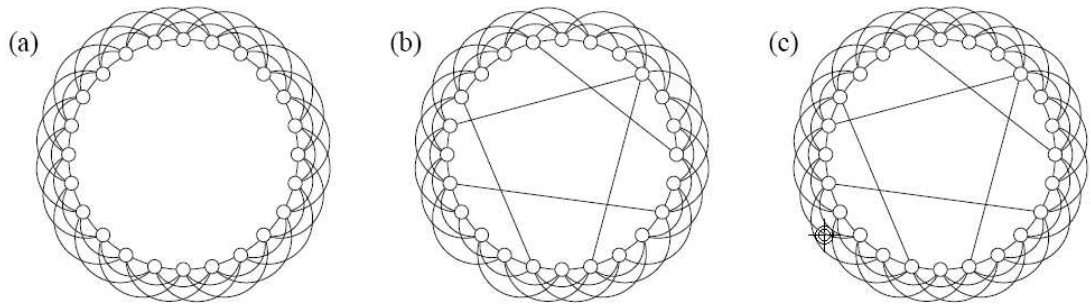


Figura 3.7 *– (a) Uma estrutura uni-dimensional com conexões onde os pares de vértices estão ligados aos seus k vizinhos mais próximos, neste caso com k = 3. (b) O modelo "*small-world*" [Watts e Strogatz (1998)] é criado escolhendo aleatoriamente uma fração p das arestas do grafo e movendo uma terminação desta aresta para um novo vértice escolhido aleatoriamente com probabilidade uniforme. (c) Uma pequena variação do modelo "*small-world*" [Newman e Watts (1999)] no qual atalhos são adicionados aleatoriamente entre os vértices, mas sem remoção de nenhuma aresta já existente esta estrutura.*

com alguns atalhos entre dois vértices quaisquer, mas apresentando um componente gigante como em um grafo randômico. Ao mesmo tempo, esta rede passa a apresentar um grau de aglomeração ("*clustering*") maior que o de um grafo randômico, ou seja, existe um aumento na probabilidade de que um vértice A qualquer esteja ligado a outro vértice C geodesicamente distante dele, propriedade também conhecida como *transitividade*.

3.1.4 Redes Livres de Escala

Uma característica comum aos grafos ER randômicos e os modelos WS "*small-world*" é que a distribuição de conectividade de ambas é homogênea [Newman (2003c), Wang e Chen (2003)], com pico em um valor médio e decaimento exponencial. Tais redes são denominadas redes exponenciais. Um fato importante sobre algumas redes do mundo real (incluindo a Internet, redes metabólicas dentre outras), é que estas são livres de escala e suas distribuições de conectividade seguem uma lei de potência.

Para explicar a origem desta distribuição em lei de potência, Barabási e Albert - BA, propuseram um outro modelo de rede [Albert et al. (1999), Albert e Barabasi (1999)]. O argumento é que muitos dos modelos existentes falham ao lidar com dois importantes atributos da maioria das redes reais.

Primeiro, as redes reais mostram-se abertas e são continuamente formadas pela adição de novos nodos, mas os modelos vigentes eram estáticos no sentido de que as conexões podiam ser rearanjadas, criadas e destruídas, mas o número de nodos mantinha-se fixo neste processo de formação. Por exemplo, a *World Wide Web* continuamente recebe e perde páginas; a literatura cientifica cresce constantemente à medida que novos artigos são publicados.

Segundo, tanto os grafos randômicos quanto os modelos "*small-world*" assumem uma probabilidade uniforme quando novas arestas são criadas, mas isto não é realístico. Por exemplo, as páginas web que já são mais referenciadas (Yahoo, Google, etc.) estarão mais propensas a serem ainda mais referenciadas; um novo artigo estará mais propenso a citar os

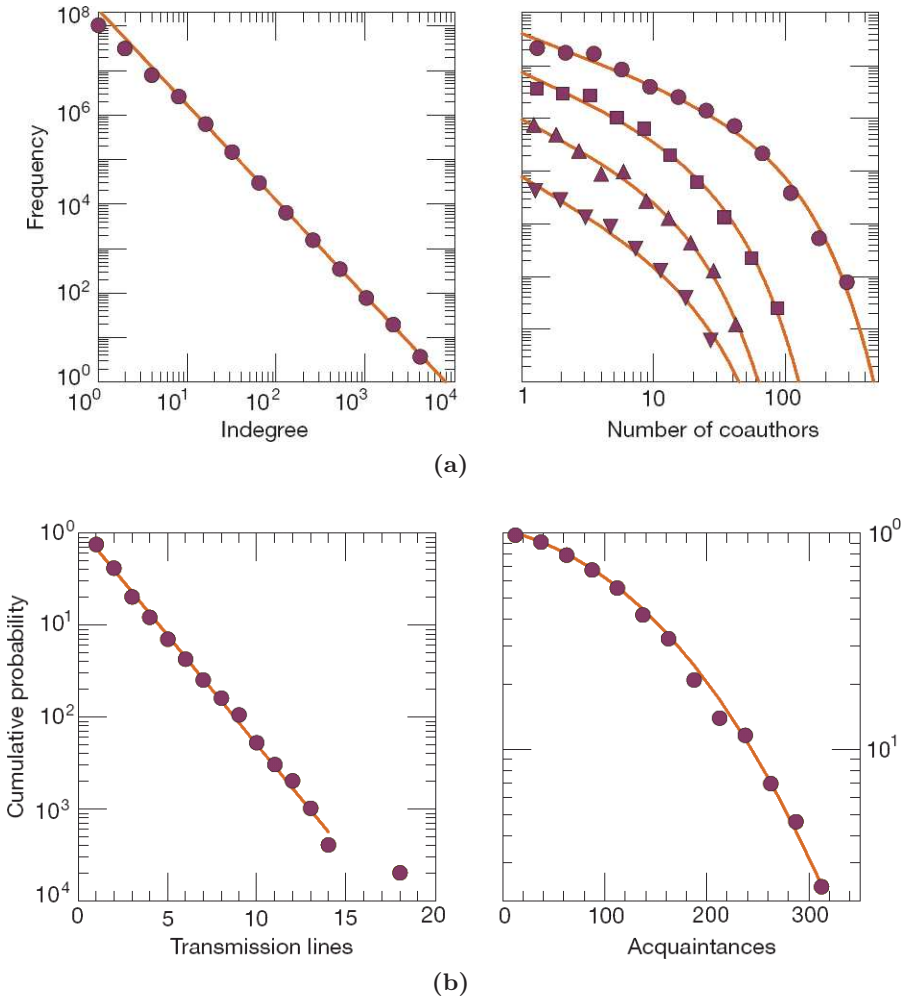


Figura 3.8 *– Alguns exemplos típicos de distribuições associadas às redes do mundo real. Em (a) os dados representam a* World-Wide Web. *Em (b) os dados representam redes de coautoria.*

artigos mais citados do que aqueles menos citados. Este fenômeno é do tipo "os ricos ficam cada vez mais ricos", o qual os demais modelos não levam em conta.

O modelo BA sugere que dois principais ingredientes das redes auto-organizadas em estruturas livre de escala crescem por agregação preferencial. Isto aponta para o fato que muitas das redes crescem continuamente pela adição de novos nodos, e estes novos nodos ligam-se preferencialmente aos nodos já existentes que apresentam um grande número de conexões. O algoritmo de geração de um modelo BA livre de escala é o seguinte:

1. Uma rede inicia com um pequeno número (N_0) de nodos. A cada momento um novo nodo é introduzido e conectado à N_t ($N_t \leq N_0$) nodos já existentes;

2. A probabilidade p_i que um novo nodo seja conectado ao nodo i (um dos N_t nodos já existentes) depende do grau k_i do nodo i, de forma que

$$p_i = \frac{k_i}{\sum_j k_j}. \tag{3.1}$$

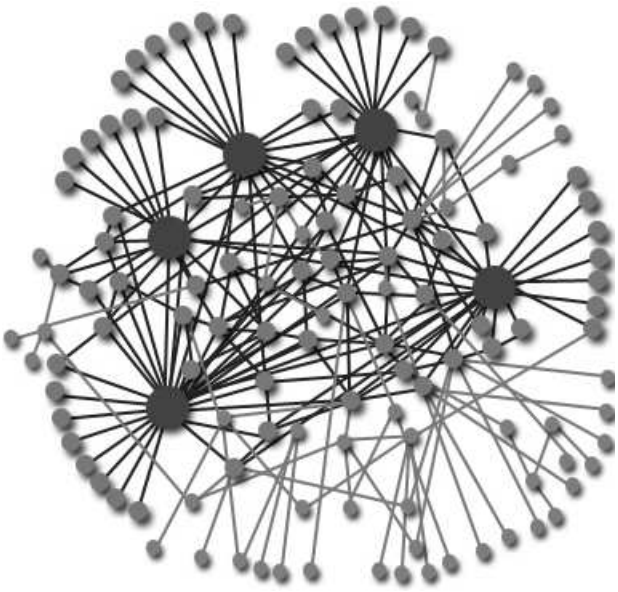


Figura 3.9 – *Uma rede livre de escala com 130 nodos, gerada com o algoritmo BA. Os cinco nodo mais conectados estão em contato com 60% dos demais nodos.*

Após t passos de tempo, este algoritmo resulta em uma rede com $N_t = t + N_0$ nodos e $N_t t$ arestas (figura 3.9). Crescendo de acordo com estas duas regras, a rede evolui para uma rede livre em escala. A forma da distribuição de conectividade dos nodos na muda com o tempo, não mudando com a escala para a qual a rede cresce.

Resultados numéricos têm indicado que, em comparação com um grafo randômico com o mesmo tamanho e o mesmo grau médio de conectividade, a distância média entre nodos na rede livre de escala é algo menor mas com coeficiente de aglomeração bem mais alto. Isto implica que a existência de uns poucos nodos "grandes" com um grau muito maior (com um número muito maior de conexões) exerce um papel decisivo em trazer os demais nodos da rede junto aos demais. Entretanto, não existem ainda nenhuma fórmula analítica capaz de predizer a distância média entre nodos e o coeficiente de aglomeração para os modelos livre de escala [Wang e Chen (2003)]. O modelo BA é um modelo mínimo que captura os mecanismos responsáveis pela distribuição em lei de potência. Segundo alguns autores [Wang e Chen (2003)], este modelo tem algumas limitações, quando comparado com algumas redes do mundo real. Albert e Barabási [Albert e Barabasi (2002)] apresentam uma revisão de algumas alternativas para o modelo BA puro.

3.2 Pontos Fracos das Redes Complexas

Um fenômeno interessante que ocorre com as redes complexas é o fato delas apresentarem vulnerabilidades, usualmente denominado "Calcanhar de Aquiles". Apesar de se apresentarem robustas com relação a "ataques" aleatórios, elas se mostram sensíveis quando o "ataque"

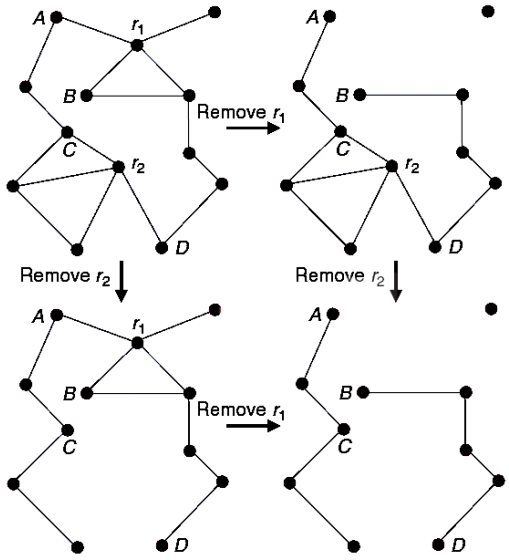


Figura 3.10 *– Ilustração dos efeitos da remoção dos nodos em uma rede inicialmente conectada. Inicialmente, $d_{AB} = d_{CD} = 2$. Após a remoção do nodo r_1 da rede original, $d_{AB} = 8$. Após a remoção do nodo r_2 da rede original, $d_{CD} = 7$. Após a remoção dos nodos r_1 e r_2, a rede quebra-se em três aglomerados isolados e $d_{AB} = d_{CD} = \infty$.*

ocorre de forma direcionada. A título de ilustração, consideremos uma rede grande e conectada. Se a cada momento um nodo for retirado (figura 3.10), as arestas ligadas a ele também são eliminadas havendo a destruição de alguns trajetos que antes existiam entre os nodos restantes. Se existirem múltiplos trajetos ligando dois nodos i e j, a destruição de um deles pode significar que a distância entre eles - d_{ij}, irá crescer o que pode causar o crescimento da distância média entre nodos - L, de toda a rede. Em casos mais severos, quando só existe um único trajeto ligando i e j, a destruição deste caminho leva estes nodos a ficarem desconectados. A conectividade de um rede é robusta (tolerante a falhas) se ela contém um aglomerado gigante que abarca vários nodos, mesmo após a remoção de uma fração destes nodos.

Um exemplo interessante é a antiga ARPANET - rede de comunicação entre computadores, embrião a antecessora da Internet, criada pelo departamento de defesa americano, no final da década de 1960. O objetivo da ARPANET era permitir o suprimento contínuo de informações caso alguma das subredes falhasse. Na atualidade, a Internet cresceu tornando-se uma rede gigantesca passando a exercer um papel determinante em muitos aspectos da vida moderna. Apesar dos vários ataques de que se tem noticia, a Internet continua operacional. À semelhança de outras redes complexas, a “remoção” aleatória de nodos não afeta seu desempenho. Ao contrário, mesmo se 80% dos nodos falharem, isto ainda não causa o colapso da rede. Entretanto, se os nodos mais conectados da rede falharem, a rede inteira falha (Figura 3.11). É provado que tal tolerância a falhas e vulnerabilidade a ataques direcionados são propriedades genéricas das redes livres de escala [Albet et al. (2001), Callaway et al. (2000), Cohen et al. (2000), Cohen et al. (2001)].

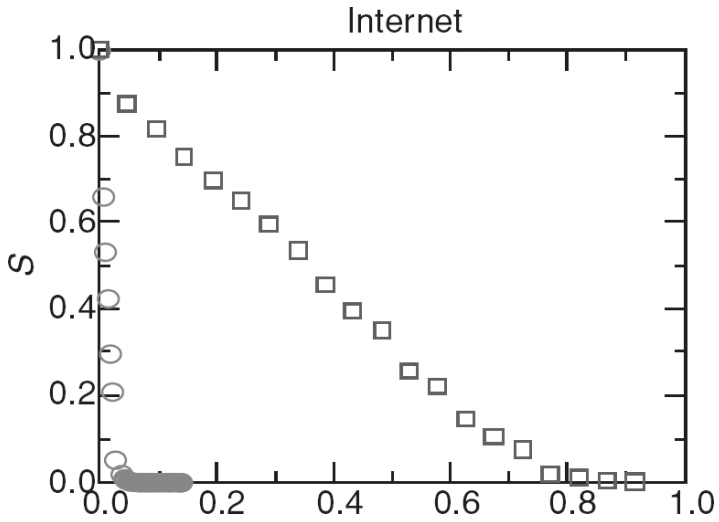


Figura 3.11 *– O tamanho relativo S do maior algomerado na Internet, quando a fração f dos domínios é removida [Albet et al. (2001)]. Os quadrados azuis mostram o efeito quando os nodos são removidos aleatoriamente. Os círculos vermelhos mostram quando os nodos mais conectados são preferencialmente removidos.*

3.3 Métricas Básicas

Os estudos relativos às redes apresentam e se valem de uma gama variada de quantidades e medidas para caracterizar e compreender estas redes complexas. Contudo, dentre essas métricas algumas podem ser ressaltadas: a distribuição do grau de conectividade dos nodos, o coeficiente de aglomeração ou "clustering", número médio de passos entre nodos. Estas métricas são freqüentemente utilizadas no estudo das redes complexas, servindo como parâmetros que permitem comparar estas redes e fazer inferências sobre as mesmas. Originalmente, Watts e Strogatz [Watts e Strogatz (1998)] tentavam em seu trabalho construir um modelo de rede com uma pequena distância média entre nodos, tal como em uma rede randômica, e com um coeficiente de aglomeração relativamente grande, tal como em uma rede regular. Tal modelo evoluiu então para se tornar um novo modelo de rede tal como ele se apresenta hoje. Por outro lado, a descoberta das redes livres de escala baseou-se na observação de que as distribuições de freqüências de contatos das redes reais tendiam a apresentar um decaimento de lei de potência, apesar deste tipo de distribuição já ter sido previamente estudado na física para vários outros sistemas e processos. Nesta seção tais conceitos serão explorados mais detidamente.

3.3.1 Distâncias entre Nodos

Em uma rede, a distância geodésica d_{ij} entre dois nodos quaisquer i e j, é definida como sendo o menor número de arestas que devem ser percorridas para liga-los. Desta forma, o diâmetro D de uma rede é definido como sendo a maior distância dentre todas as distâncias

mínimas existente na rede para ligar dois nodos quaisquer. Já a distância média L da mesma rede é definida como sendo a média de todas as distâncias mínimas existentes na rede para ligar dois nodos. Desta maneira L determina o "tamanho" efetivo da rede, indicando o grau de separação típica entre um par de nodos existentes naquela rede.

Uma abordagem satisfatória para o cálculo de L é faze-la numericamente igual à "média harmônica" das distâncias geodésicas entre todos os pares de nodos:

$$L^{-1} = \frac{1}{\frac{1}{2}n\,(n+1)} \sum_{i \geq j} d_{ij}^{-1}. \tag{3.2}$$

Desta forma a distância de cada um dos nodos até ele mesmo (que é zero) pode ser incluída no cálculo da média. A adoção deste cálculo mostra-se matematicamente conveniente por uma série de razões [Newman (2003c)], mas nem todos os autores o fazem desta forma. Ao mesmo tempo, existem redes onde um dado par de nodos (ou vértices) não apresentam um caminho possível para liga-los. Por uma questão semântica, convencionou-se atribuir valor "infinito" à distância geodésica entre estes pares. Na definição apresentada, a ocorrência de valores infinitos para d_{ij} não contribui para o cálculo.

A titulo de exemplo, em uma rede social, L seria o número médio de pessoas existente no menor "caminho" possível para se ligar duas pessoas nesta rede. Constitui um achado interessante o fato de que nas redes reais o caminho médio mínimo é relativamente pequeno, mesmo nos casos onde as redes apresentam menos conexões que uma típica rede globalmente conectada com igual número de nodos. Esta pequenez típica permite inferir o efeito de "*small-world*", vindo daí o nome de redes "*small-world*".

Os índices relativos às distâncias entre nodos de uma rede exercem uma influência importante na emergência do efeito de "*small-world*" nas redes. Este efeito tem implicações óbvias para a dinâmica dos processos que ocorrem nas redes. Por exemplo, no espalhamento de informações (ou fenômenos similares) através de uma rede, o efeito de "*small-world*" implica que este espalhamento será mais rápido em uma rede que apresenta tal efeito, quando comparado a redes que não apresentam tal efeito. Na média, a percolação em redes que apresentam esta propriedade necessita de seis passos para alcançar qualquer ponto desta. Quando comparado a uma rede que demanda centenas ou milhares de passos para ir de um ponto qualquer a outro, obviamente que este fenômeno ocorrerá muito mais rápido naquela rede.

Isto afeta, por exemplo, o trafego de pacotes na Internet ou em uma rede local de computadores; o número de passos em uma jornada; o tempo de viajem por avião ou trem. Implica também na velocidade de espalhamento de doenças por uma população, de boatos em uma sociedade, de "vírus" em redes de computadores.

Uma rede apresenta o fenômeno de "*small-world*", se para qualquer nodo desta rede, o número de nodos N, presentes dentro de um raio r crescer exponencialmente com r (ou seja $N \propto e^r$), e o valor da distância média entre nodos crescer logaritimicamente (ou a taxas menores) com N (ou seja $L \propto \log(N)$) [Newman (2003c)]. Algumas redes apresentam uma distância média entre vértices que cresce com uma taxa menor que *log n*. Bollobás e Riordan

mostram que redes com escalonamento em lei de potência apresentam valores de l que crescem com uma taxa menor que *log n / log log n*.

3.3.2 Transitividade ou Coeficiente de Aglomeração (*Clustering*)

Dentro de uma turma de alunos de uma universidade é possível que um amigo do amigo do aluno A seja também amigo direto deste aluno. Pondo isto de outra forma, é bem possível que dois amigos do aluno A sejam amigos um do outro. Esta propriedade apresentada pelas redes reais refere-se ao grau de aglomeração destas redes. A métrica adotada para mensurar quantitativamente esta propriedade das redes é denominada *coeficiente de aglomeração* ou "*clustering coefficient*" - C. Entretanto, para entender este conceito deve-se observar o seguinte raciocínio.

Seja o nodo i pertencente a uma rede, o qual possui $k(i)$ vizinhos. É trivial provar que *se o nodo i estiver ligado a todos os seus* $k(i)$ *vizinhos, então existirão, no máximo,* $k(i)\ (k(i)-1)/2$ *ligações (ou arestas) ligando i aos seus* $k(i)$ *vizinhos.* O coeficiente de aglomeração C_i relativo ao nodo i é então definido como sendo a razão entre o número de ligações que realmente existem entre i e seus $k(i)$ vizinhos - E_i, e o número máximo de ligações possível de existir entre estes nodos - $k(i)\ (k(i)-1)/2$. Formalmente, tem-se:

$$C_i = \frac{2E_i}{(k(i)\ (k(i)-1))} \tag{3.3}$$

O coeficiente C de toda a rede é então definido como sendo a média de todos os C_i da rede. É possível observar que $C \leq 1$ sendo que $C = 1$ só ocorre se e somente se a rede estiver globalmente ligada, o que significa que todos os nodos da rede estarão conectados a todos os demais nodos desta rede. Em uma rede completamente randômica constituída de N nodos, $C \sim 1/N$, o que é um valor muito pequeno quando comparado à maioria das redes encontradas no mundo real.

Alternativamente, Newman [Newman (2003c)] propõe que a quantificação de C global (C_N), pode ser feita pela seguinte relação:

$$C_N = \frac{3N_\triangle}{N_3}, \tag{3.4}$$

onde $3N_\triangle$ é o número de triângulos presentes na rede e N_3 é o número de triplas conectadas. Uma "tripla conectada" é um conjunto de três vértices onde cada vértice pode ser acessado pelos outros dois ou seja, os outros dois vértices devem ser adjacentes ao primeiro vértice. Desta forma têm-se

$$3N_\triangle = \sum_{k>j>i} a_{ij}a_{ik}a_{jk}, \tag{3.5}$$

$$N_3 = \sum_{k>j} a_{ij}a_{ik}, \tag{3.6}$$

A título de exemplo, seja o triângulo $B\hat{A}C$. Este triângulo dá origem a três triplas:

$A\hat{C}B \equiv B\hat{C}A$, $C\hat{B}A \equiv A\hat{B}C$,$B\hat{A}C \equiv C\hat{A}B$. Aplicando a relação 3.4 têm-se que $C = 1$ para esta rede.

Com efeito, C mede a fração de triplas que, tendo sua terceira aresta preenchida, completam triângulos. O fator 3 que precede o numerador contabiliza o fato de que cada triângulo contribui com 3 triplas e assegura que C permaneça no intervalo $0 \leqslant C \leqslant 1$.

Por fim, vale ressaltar que que boa parte das redes de grande escala existentes no mundo real apresentam a tendência de se aglomerarem [Wang e Chen (2003)], sendo de que seus coeficientes de aglomeração são usualmente maiores que O(1/N), apesar de se manterem significativamente menores que 1 - estão longe de serem globalmente conectadas. Ao mesmo tempo tal fato indica que as redes que se estabelecem no mundo real também não são completamente randômicas.

3.3.3 Distribuição dos Graus de Conectividade

A mais simples e talvez mais importante característica topológica de um nodo, tomado de forma isolada, é o seu grau de conectividade. O grau $k(i)$ de um nodo i é usualmente definido como sendo o seu número total de conexões. Desta forma, maior o grau de um nodo, mais importante ele será no concerto da rede. A média dos valores de $k(i)$ calculada para todo i é chamada de grau médio da rede sendo denotado na forma $\langle\ k\ \rangle$. Usualmente a distribuição dos graus dos nodos da rede - $v(k)$, é caracterizada pela função de distribuição $p(k)$, a qual é a probabilidade de que um nodo selecionado ao acaso tenha exatamente k ligações. Esta relação pode ser equivalentemente expressa na forma [Newman (2003c)]:

$$p(k) = \frac{\sum v(k)}{\sum v}. \tag{3.7}$$

O gráfico de $p(k) \times k$ para uma dada rede é formado pelo histograma dos graus dos nodos da rede. Uma estrutura regular, por exemplo, tem um único valor de grau de conectividade, visto que todos os nodos da rede apresentam o mesmo número de conexões. Desta forma o gráfico da distribuição dos graus de conectividade, para este caso, irá apresentar um único pico.

No outro extremo, para uma rede completamente randômica, tal como as estudadas por Erdös and Rényi [Erdös e Rényi (1960)],a distribuição dos graus de conectividade tenderá a obedecer uma distribuição de Poisson. Entretanto, os estudos mostraram que as grandes redes reais apresentam uma distribuição dos graus de conectividade que se afasta muito da distribuição de Poisson. Em particular, para algumas redes a distribuição dos graus de conectividade pode ser melhor descrita por uma lei de potência na forma $P(k) \sim k^{-\gamma}$

As distribuições do tipo lei de potência apresentam decaimento mais gradual que aquele típico de uma exponencial, permitindo a existência de nodos com elevado grau de conectividade. Ao mesmo tempo, pelo fato das distribuições de lei de potência serem livres de qualquer característica de escala, as redes com tal comportamento são chamadas de redes livres de escala. As características de "*small-world*" e independência de escala são comuns a muitas redes complexas do mundo real.

3.3.4 Vulnerabilidade da Rede

Quando se analisa a topologia de uma rede, é importante saber quais componentes (vértices e arestas) são cruciais para a estabilidade e funcionamento da rede. Intuitivamente, os vértices críticos da rede são aqueles vértices que apresentam maior grau de conectividade, quando comparados aos demais vértices da rede. Tais vértices são denominados vértices (ou nodos) concentradores, ou simplesmente "*hubs*". Entretanto, existem situações onde estes "*hubs*" não são necessariamente os elementos vitais para a estabilidade ou desempenho do sistema ao qual a rede jaz subjacente. Um exemplo disto são as árvores binárias, onde todos os vértices apresentam o mesmo grau, não havendo "*hub*" neste sistema. Contudo, basta que as arestas dos nodos próximos à raiz sejam desfeitas, para que o impacto sistêmico seja bem mais significativo, que a remoção de arestas próximas às folhas. Fenômenos como este sugerem que as redes podem apresentar propriedades hierárquicas, o que significa que tais elementos apresentam uma relevância (ou hierarquia) maior dentro do concerto sistêmico.

Uma forma de identificar os componentes críticos de uma rede é buscando pelos vértices mais vulneráveis. Se o desempenho de uma rede for associado à sua eficiência global, a vulnerabilidade de um vértice pode ser definida pela perda de eficiência sistêmica perceptível, quando aquele vértice e as arestas a ele associadas forem removidos [Gol'dshtein et al. (2004)]. Assim, a vulnerabiliade de um dado vértice é quantificada pela relação

$$V(i) = \frac{(E - E_i)}{E}$$

onde E é a eficiência global da rede e E_i é a eficiência global após a retirada do vértice i e das arestas associadas. Como sugerido por Gol'dshtein [Gol'dshtein et al. (2004)], a classificação dos vértices de uma rede em função das respectivas vulnerabilidades V_i irá apresentar uma hierarquia dentro da rede, onde o vértice com maior vulnerabilidade aparecerá como sendo o mais importante.

3.4 Redes Complexas Ponderadas

Parâmetros como grau dos vértices, padrão de distribuição estatística dos graus dos vértices dentro de uma rede, centralidade dos vértices, dentre outros, são capazes de prover "insights" úteis para o entendimento das redes em análise. Contudo, as redes não se caracterizam somente pela sua topologia mas também pela dinâmica dos fluxos e das flutuações que nela ocorrem. Em particular, a heterogeneidade na intensidade das conexões pode ser importante para o entendimento dos sistemas, que tais redes representam. De forma análoga, as interações que ocorrem entre os átomos no seio das proteínas representam um papel importante para a definição da topologia e estabilidade das mesmas.

Motivado por esta constatação, serão exploradas nesta seção as propriedades estatísticas das redes complexas cujas arestas têm valores (ou pesos) associados e que podem ser generalizadas para descrever as propriedades das redes em termos de grafos ponderados ("*weighted graphs*")[Yook et al. (2001), Barrat et al. (2004a), Newman (2004), Barrat et al. (2004b), Barrat et al. (2004c), Barthelemy et al. (2005)].

Nesta seção será apresentado um conjunto de métricas que combinam aspectos topológicos das conexões e os pesos a elas assinalados. Estas quantidades provêem uma caracterização geral das propriedades estatísticas das redes ponderadas como centralidade, coesividade local, etc. Contudo, enquanto os estudos que versam sobre a caracterízação das redes não ponderadas baseam suas análises nas métricas estatísticas comuns já citadas, o mesmo não ocorre com a análise das redes ponderadas.

Diferentes autores [Yook et al. (2001), Almaas et al. (2004), Park et al. (2004), Newman (2004), Barrat et al. (2004a), Hu et al. (2005), Wu et al. (2005), Onnela et al. (2005), Barthelemy et al. (2005), Kalna e Higham (2006)] têm apresentando diferentes proposições de generalização das métricas mais significativas para o caso das redes ponderadas, produzindo uma gama variada de possíveis definições para estas métricas. Estudos relativos às redes ponderadas [Barrat et al. (2004a), Li e Chen (2004), Li e Cai (2004), Newman (2004)] têm mostrado que elas podem exibir propriedades complexas adicionais tal como correlações não triviais de pesos que não encontram explicação se forem considerados somente os aspectos topológicos subjacentes. A heterogeneidade na intensidade das conexões aparece como um atributo de importância nos sistemas do mundo real e não deve ser negligenciada.

3.4.1 Conceito de peso em redes ponderadas

Usualmente as propriedades de um grafo podem ser expressas por meio de sua matriz de adjacência $[A]$, cujos elementos assumem valores 1 caso uma aresta conecte um vértice a um outro vértice, ou 0 caso isto não aconteça. De forma similar, grafos ponderados podem ser descritos por uma matriz de adjacências onde cada elemento desta matriz especifica o valor associado à aresta que conecta dois vértices, ou apresenta o valor 0 caso não exista conexão entre estes vértices. No contexto deste trabalho os grafos são considerados como sendo não dirigidos, o que equivale dizer que a matriz de adjacência representativa dos mesmos será simétrica apresentando valores nulos ou positivos.

De uma maneira formal[Barthelemy et al. (2005)], dado o grafo $\mathcal{G}(\mathcal{V}, \mathcal{A}_\nabla)$, as funções $w_\mathcal{V} : \mathcal{V}(\mathcal{G}) \mapsto \mathbb{R}$ e $w_{\mathcal{A}_\nabla} : \mathcal{A}_\nabla(\mathcal{G}) \mapsto \mathbb{R}$ são funções que atribuem, à cada um dos vértices e/ou arestas do grafo $\mathcal{G}$ valores (ou pesos) significativos. No contexto deste trabalho serão considerados como significativos somente os casos onde $w_{\mathcal{A}_\nabla} \geq 0$.

3.4.2 Conectividade em redes ponderadas : *Strength*

Quando os atributos (ou pesos) associados às arestas de uma rede são levados em consideração, em uma análise estatística, deve-se ter em mente que tais atributos devem ser capazes de caracterizar, de forma significativa, tanto a estrutura quanto a organização desta rede. Estes pesos associados às ligações entre pares i e j de vértices - w_{ij}, são representativos da relação entre estes dois vértices. Em redes com estas peculiaridades, a propriedade mais significativa dos vértices não seria o grau deste vértice, mas sim sua força ("*vertex strength*"), definida como [Barrat et al. (2004a)]:

$$s_i = \sum_{j=1}^{N} w_{ij}$$

Esta quantidade expressa a força de um vértice em temos do peso total das conexões que nele chegam. Esta quantidade, segundo Barrat *et al.* [Barrat et al. (2004a)], aparece como uma medida intuitiva da importância ou centralidade de um vértice v_i em uma rede.

Contudo, nem sempre os nodos com maior grau são de fato os mais relevantes [Newman (2004)]. Dependendo do contexto, a estrita avaliação dos vértices somente em função de seus graus perde de vista o fato de que nodos com baixo grau podem ser cruciais para conectar diferentes regiões da rede, ao atuar como pontes. De forma a quantificar a significância deste papel exercido por alguns vértices da rede, a medida de transitividade ("*betweenness*") [Newman (2001b), Goh et al. (2001b), Freeman (1977)], é definida como o número de caminhos mais curtos entre dois pares de vértices quaisquer da rede que passam por um dado vértice v_i. Neste caso, os nodos com maior transitividade participam de um número maior de caminhos curtos dentro da rede, que os vértices ditos "periféricos". Em arranjos estruturais similares às proteínas [Newman (2001b), Newman (2001a), Barthelemy (2004), Barthelemy et al. (2005), Zhao et al. (2005b)], tal classe de vértices responde pela transmissão de sinais pela rede levando a fenômenos de sincronicidade e efeitos não locais. Esta definição de centralidade encontra-se ligado somente a elementos geométricos. Em redes cujas arestas apresentam pesos associados, é necessário considerar a definição alternativa de centralidade construída com base na força s_i do vértice v_i. Tal definição aparece como sendo a mais apropriada para caracterizar a importância do vértice em uma rede ponderada.

Para o caso de uma rede não direcionada e conectada, com N_V vértices, sendo v_i, v_j e v_k três vértices quaisquer. Adota-se a definição de aglomeração ponderada de um vértice tal

forma que [Barrat et al. (2004a), Barthelemy et al. (2005)]

$$c^w(i) \;=\; \frac{1}{s_i(k_i-1)}\sum_{j,k}\frac{(w_{ij}+w_{ik})}{2}a_{ij}a_{ih}a_{jk}.$$

O coeficiente de aglomeração $c^w(i)$ mede o grau de aglomeração (ou coesividade) local e é definido para todo vértice da rede. Esta métrica avalia, para cada tripla existente dentro da rede, os pesos das arestas formadas pelo vértice i com a sua vizinhança. Desta forma não só o número de triângulos formados dentro da rede é contabilizado, mas também o peso relativo das arestas, com relação à força total do vértice. O fator $s_i(k_i-1)$ atua como fator de normalização e garante que $0 \leq c^w(i) \leq 1$. O coeficiente médio da rede

$$C = \frac{1}{N}\sum c^w(i)$$

expressa a densidade média de interações formadas pelos diversos tripletos existentes dentro da rede.

3.5 Análise Espectral de Grafos

Os parâmetros descritivos das redes complexas já citados mostram-se úteis na análise dos dados obtidos acerca das redes observadas no mundo real. Contudo, tais parâmetros descrevem propriedades estruturais, mas informam muito pouco acerca dos atributos sistêmicos das redes como aqueles relacionados aos aspectos dinâmicos como os processos de difusão, arranjo hierárquico dos vértices e emergência de novas propriedades. Relevantes indícios de tais atributos podem, contudo, ser obtidos da aplicação de métodos oriundos da matemática discreta e da teoria dos grafos. Em particular a aplicação da análise espectral de grafos tem sido utilizada na análise de redes complexas [Farkas et al. (2001), Dorogovtsev et al. (2003),Seary e Richards (2003), Dorogovtsev et al. (2004), Rodgers et al. (2005), Zhao et al. (2005a), Paccanaro et al. (2006)], tendo sido aplicada em análses de microarray, em estudos preliminares de estruturas protéicas [Brinda et al. (2002), Krishnadev et al. (2005)], tanto quanto na análise de redes de interação proteínas-proteínas [Farkas et al. (2002), Kamp e Christensen (2005)].

A motivação para o uso da análise espectral de matrizes de grandes dimensões surge na física nuclear durante a década de 1950. Tal aplicação tem como referência o trabalho de Wigner [Wigner (1955)] relacionado com os problemas de mecânica quântica onde os níveis de energia dos *quanta* não são diretamente observáveis, mas podem ser caracterizados pelos valores singulares da matriz de observações. A distribuição espectral empírica -ESD (distribuição empírica dos valores singulares) só é possível por meios computacionais.

Motivados pela forma análoga com que os vários sistemas reais podem ser modelados e analisados, diferentes autores [Farkas et al. (2002), Krishnadev et al. (2005), Rodgers et al.

(2005), Dorogovtsev et al. (2004), Zhao et al. (2005a), Goh et al. (2001a), Farkas et al. (2001), Dorogovtsev et al. (2003), de Aguiar e Bar-Yam (2005), Kamp e Christensen (2005), Seary e Richards (2003), Brinda et al. (2002)] vêm procurando aplicar a análise espectral no estudo de sistemas complexos através da decomposição de suas matrizes representativas. Notadamente, os processos dinâmicos que ocorrem em uma rede, e como os aspectos estruturais das redes determinam tais processos dinâmicos, podem ser inferidos pela observação dos valores e vetores singulares derivados da topologia desta rede.

A análise espectral ("*eigendecomposition*") tem sido aplicada por diferentes pesquisadores, de forma implícita ou explícita desde o final dos anos 1960, quando os computadores se tornaram mais acessíveis na maioria das universidades. Os valores singulares de uma rede estão intimamente relacionados à importantes características tais como a presença de aglomerados coesivos. Os vetores singulares de tais redes têm sido utilizados como um sistema de coordenadas natural para a visualização das redes, podendo ainda prover maneiras de identificação de aglomerados, avaliação do grau de robustez da rede e outras características locais.

Nesta seção são apresentados os aspectos fundamentais da análise espectral de redes, os quais são aplicados ao estudo das proteínas e cujos resultados serão apresentados na seção 5.3.

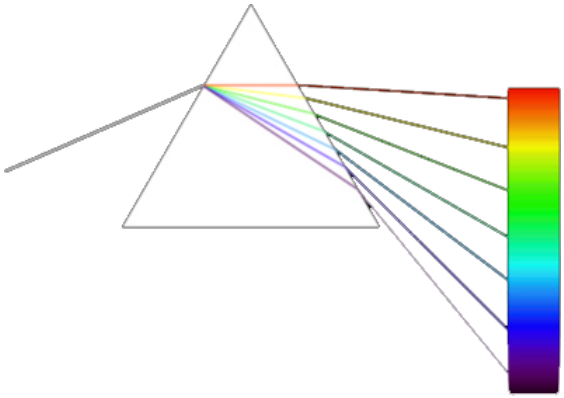
3.5.1 Decomposição de Valores Singulares

Originalmente a palavra *spectrum* (plural *spectra*) de origem latina, era usada como referência a "fantasmas" ou "aparições". Contudo, Isaac Newton, em 1671, agregou a este termo uma nova significância ao reportar seus experimentos relacionados à decomposição da luz solar. Na atualidade este termo ganhou aplicações mais amplas, mas mantém o sentido relativo ao conjunto de componentes singulares no qual alguma coisa pode ser decomposta [Merriam-Webster Online].

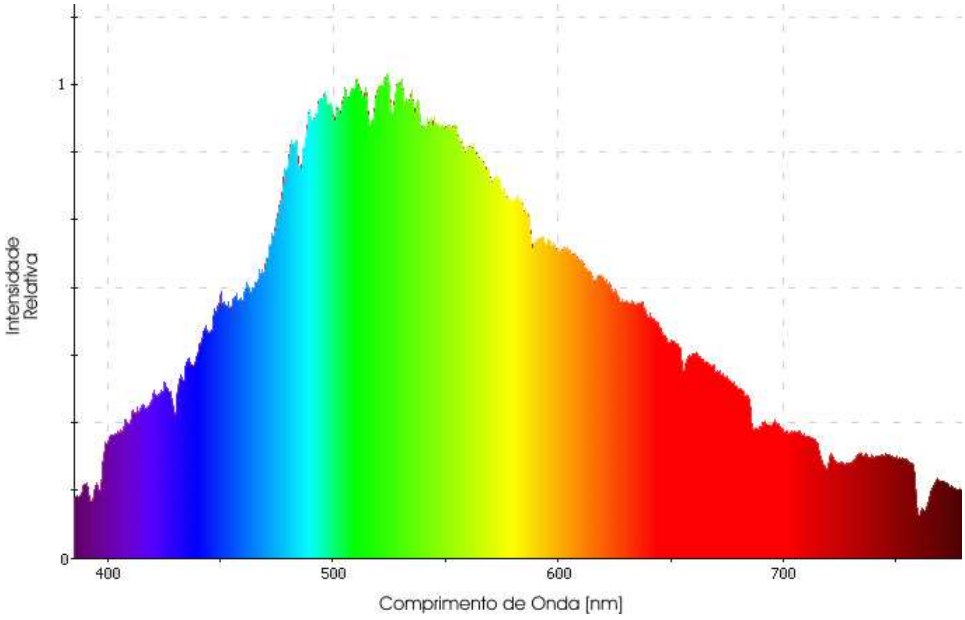
Uma forma simplificada, porém útil, para o entendimento do que vem a ser a análise espectral, é compará-la com a observação do espectro de cores da luz. A figura 3.12 representa a decomposição do espectro de cores da luz solar. A análise do espectro de cores da luz emitida por uma fonte qualquer, permite identificar tanto as propriedades físicas e químicas da fonte da luz, como também dos diferentes meios que esta luz tenha atravessado antes de chegar ao observador. Nesta situação, o estudo dos diferentes comprimentos de onda existentes no espectro e das respectivas intensidades, possibilita fazer inferências acuradas a respeito da fonte da luz e do meio por onde ela propagou.

A extensão dos conceitos relativos a análise espectral, levou a técnicas matemáticas mais elaboradas, voltadas para a identificação dos componentes fundamentais dos modelos representativos de diferentes sistemas.Uma destas técnicas é a Decomposição em Valores Singulares (SVD), o qual surge como um método de fatoração de matrizes retangulares reais ou complexas. A decomposição em valores singulares foi originalmente desenvolvida para atender problemas de geometria diferencial.

Eugenio Beltrami e Camille Jordan desenvolveram em 1873 e 1874 respectivamente, trabalhos independentes onde os valores singulares de sistemas representados por matrizes for-



(a)



(b)

Figura 3.12 – *Espectro de cores fundamentais da luz do Sol. Em (a), o esquema mostra o mecanismo básico de decomposição da luz em suas cores fundamentais. Em (b), o gráfico mostra a distribuição de intensidade relativa, dos diferentes comprimentos de onda fundamentais que compõem a luz do Sol.*

mavam um conjunto de elementos invariantes em condições de substituições ortogonais. A primeira prova formal da decomposição em valores singulares para matrizes retangulares e complexas é atribuída a Eckart e Young em 1936, apresentada como uma generalização da transformação de eixos principais para matrizes Hermitianas [1].

A teoria mais abrangente acerca da existência e das propriedades dos valores singulares foi apresentada em 1910 por Émile Picard, o qual pela primeira vez denominou estes valores invariantes como valores singulares ("*valeurs singulières*").

Métodos práticos para a solução da SVD eram desconhecidos até 1965, quando Golub e Reinsch publicaram seu algoritmo. Em 1970, Golub e Reinsch publicaram uma variação do

[1] Uma matriz inteira ou real é dita Hermitiana se e somente se ela é simétrica.

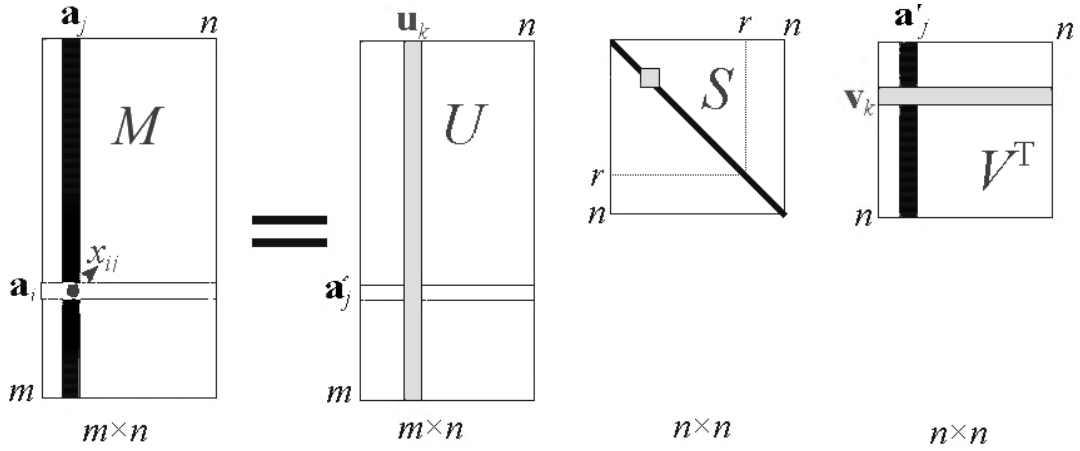


Figura 3.13 *– Esquema da aplicação da SVD em uma matriz [M].*

algoritmo original, o qual é usado até a atualidade.

Seja $[M]$ uma matriz $m \times n$ cujas entradas são reais. Pelos trabalhos já citados, existe uma fatoração na forma

$$M_{m\times n} = U_{m\times n}\ S_{n\times n}\ V^T_{n\times n}$$

onde $U_{m\times n}$ é uma matriz unitária com $U_{ij} \in \mathbb{R}$, onde $S_{m\times n}$ é uma matriz com valores não negativos em sua diagonal principal e zeros na outra diagonal, onde a matriz V^T denota a matriz transposta de $V_{n\times n}$ com $V_{ij} \in \mathbb{R}$. Tal fatoração é denominada de Decomposição em Valores Singulares de $[M]$. Ou seja

$$U^T U = I_{m\times m}$$

$$V^T V = I_{n\times n}$$

As matrizes $[U]$ e $[S]$, na analogia com o espectro de cores da luz do Sol, podem ser vistas, respectivamente, como representativas dos diferentes comprimentos de onda fundamentais (cores) constituíntes desta luz, e da intensidade de cada um destes diferentes comprimentos de onda identificados neste espectro.

Formalmente, a matriz $[S]$ é definada como sendo uma matriz diagonal, apresentando valores somente ao longo de sua diagonal principal. Estes valores são ditos valores singulares ("*eigenvalues*") da matriz $[S]$. Desta forma, $[S]$ pode também ser vista como sendo um vetor onde

$$[\mathbf{S}] = \{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n\}$$

onde

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n \geq 0.$$

As colunas da matriz $[U] = \{u\}$ são definidas como sendo os vetores singulares à esquerda (ou simplesmente vetores singulares) da matriz $[M]$, ao mesmo tempo em que as linhas da matriz $[V] = \{v\}$ são definidas como sendo os vetores singulares à direita da mesma matriz

$[M]$.

Uma propriedade fundamental destas matrizes é aquela onde

$$[\mathbf{M}]v_i \;=\; \lambda_i u_i$$

$$[\mathbf{M}]^T u_i \;=\; \lambda_i v_i$$

3.5.2 Matrizes relacionadas às redes

Como já foi visto em 3.4, um grafo pode ser representado por sua matriz de adjacências. Desta forma, um grafo $\mathcal{G}$ tem sua matriz de adjacências $[A](\mathcal{G})$, na forma canônica, formada da seguinte maneira

$$[A]_{ij} = \begin{cases} 1 & \text{se o vértice } i \text{ está ligado ao vértice } j, \\ 0 & \text{nos demais casos.} \end{cases}$$

No caso dos grafos não serem direcionados, os termos $[A]_{ij}$ e $[A]_{ji}$ apresentam o mesmo valor, o que implica que a matriz $[A](\mathcal{G})$ é simétrica. No caso em que os grafos são ponderados, existe a opção de se adotar uma matriz de adjacências ponderadas, onde

$$[A]_{ij} = \begin{cases} w_{ij} & \text{se o vértice } i \text{ está ligado ao vértice } j, \\ 0 & \text{nos demais casos.} \end{cases}$$

Vale notar que, para a maioria das redes existentes no mundo real, as matrizes de adjacência são esparsas, apresentando a maior parte das suas posições preenchidas com '0's (zeros).

Associada à matriz $[A](\mathcal{G})$ pode-se definir a matriz de distribuição de graus - $[G_r](\mathcal{G})$, que é definida como uma matriz diagonal cujos valores são resultado da soma por linha, dos valores de $[A](\mathcal{G})$.

3.5.3 Análise Espectral da Matriz de Adjacências de um Grafo

Seja o vetor λ o conjunto de valores singulares ("*eigenvalues*") da matriz de adjacências de um grafo $\mathcal{G}$. Dentre os diferentes valores singulares que compõem o vetor λ, o maior deles – λ_1, é denominado *principal valor singular* (ou "*principal eigenvalue*"). Tal como explicado na seção 3.5.1.

A densidade espectral $\rho(\lambda)$, da matriz de adjacências de um grafo $\mathcal{G}$ finito, pode ser escrito como

$$\rho(\lambda_i) = \frac{1}{N} \sum_{j=1}^{N} \delta(\lambda_i - \lambda_j), \tag{3.8}$$

onde λ_j é o j^{o} maior valor singular, da matriz de adjacências de $\mathcal{G}$. Tal soma converge para uma função contínua, quando $N \to \infty$. A densidade espectral de um grafo pode ser

diretamente relacionada às suas propriedades topológicas.

Em um grafo randômico não-correlacionado[2], o principal valor singular λ_1 mostra a densidade de arestas e λ_2 mostra-se relacionado à capacidade deste grafo de conduzir sinais [Farkas et al. (2001)].

3.5.4 Distribuição dos Valores Espectrais

Por definição, o espectro de densidade - $\overline{\rho_\epsilon}(\overline{\lambda})$ dos valores singulares de um grafo é expressa como

$$\overline{\rho_\epsilon}(\lambda_i) = \frac{1}{N}\sum_j \delta_\epsilon(\lambda_i - \lambda_j) \tag{3.9}$$

onde λ_i são os valores singulares da matriz de adjacências $[A]$ e N é o número total de nodos da rede. Uma vez que $[A]$ é simétrica e definida em $\mathbb{R}$, todos os valores singulares são reais.

De forma mais simplificada, a equação 3.9 indica quão diferente o valor λ_i é diferente dos demais λ_j. Desta forma, quanto mais $\overline{\rho_\epsilon}(\lambda_i)$ tender para zero, maior será sua similaridade com relação aos demais valores λ_j. Por outro lado, quanto mais $\overline{\rho_\epsilon}(\lambda_i)$ tender para um, menor será sua similaridade com relação aos demais valores λ_j.

A função $\delta_\epsilon(x)$ é definida como [Farkas et al. (2001), de Aguiar e Bar-Yam (2005), Kamp e Christensen (2005)]:

$$\delta_\epsilon(x) = \begin{cases} 1 & \text{se } x \in [0 \pm \frac{\epsilon}{2}] \\ \\ 0 & \text{se } x \in]0 \pm \frac{\epsilon}{2}[\end{cases}$$

A função $\delta_\epsilon(x)$ é uma função de atenuação que tende para a função δ de Krönecker quando $\epsilon \to 0$.

Enquanto uma rede randômica clássica exibe uma distribuição semicircular para a função de densidade espectral de sua matriz de adjacências [Wigner (1955), Dorogovtsev et al. (2004)], as redes observáveis no mundo real apresentam uma distribuição de espectros bem variada [Farkas et al. (2001), Dorogovtsev et al. (2003), Chung et al. (2003), Goh et al. (2001a)].

A densidade espectral de redes randômicas esparsas, com baixo valor médio de conectividade, apresenta picos nos valores singulares correspondentes a uma forte prevalência de "árvores" finitas, *i.e.*, estruturas organizadas de forma hierarquica, nestas redes [Golinelli (2003), Bauer e Golinelli (2001)]. Wigner [Wigner (1955)] provou que para uma rede randômica com N elementos, representada por uma matriz real simétrica $N \times N$, $\langle [A]_{ij} \rangle = 0$ e $\langle [M]^2_{ij} \rangle = \sigma^2$ para todo $i \neq j$, e com o crescimento de N cada momento de cada $\|[A]_{ij}\|$ mantém-se finito, já que não existe hierarquia estrutural. Então no limite, quando $N \to \infty$, a densidade espectral (a *f.d.p* dos valores singulares) de $[A]/\sqrt{N}$ converge para uma distri-

[2] O termo grafo randômico "não-correlacionado" define um grafo onde: (i) a probabilidade de dois vértices quaisquer estarem conectados é a mesma para todos os vértices do grafo; (ii) estas probabilidades são variáveis independentes.

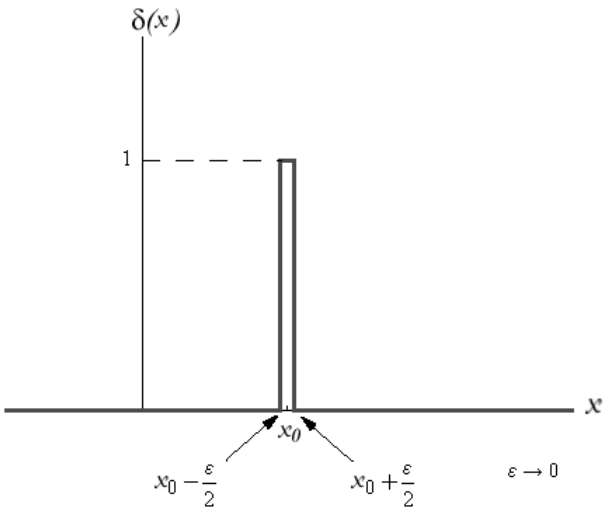


Figura 3.14 – *Função* $\delta_\epsilon(x)$

buição semicircular na forma

$$\rho(\lambda) = \begin{cases} (2\pi\sigma^2)^{-1}\sqrt{4\sigma^2 - \lambda^2} & \text{se } \|\lambda\| < 2\sigma, \\ 0 & \text{caso contrário .} \end{cases} \tag{3.10}$$

Na figura 3.15 são mostradas as distribuições de densidades – $\rho(\lambda)$, apresentadas em [Farkas et al. (2001), Dorogovtsev et al. (2003)], da análise espectral das matrizes de adjacências de cinco modelos de redes:

1. Rede randômico clássico de Erdös e Rényi, com grau médio - $< k >= 10$

   - A distribuição de densidade, obtida da análise no modelo teórico aproxima-se de uma distribuição semi-circular, mostrada como uma linha sólida;
   - A distribuição de densidade obtida de um grafo com 20.000 vértices é similar a uma distribuição semi-circular, mostrada como uma seqüência de círculos [Dorogovtsev et al. (2003)];

2. Rede de Watts e Strogatz, com grau de conectividade - $< k >= 4$

   - A distribuição de densidade, obtida da análise no modelo teórico aproxima-se de uma distribuição em sino, sendo que para um número suficientemente grande de vétices, a distribuição tende a ser semi-circular [Dorogovtsev et al. (2003)];

3. Rede livre de escala ("*scale free*") com $\gamma = 3$ e com menor grau $k_0 = 5$

   - A distribuição de densidade, obtida da análise do modelo teórico é mostrada como uma linha em traço e ponto;

- A distribuição de densidade, obtida da análise do modelo de Barabási-Albert com 70.000 vértices é mostrada como uma distribuição triangular indicada pela seqüência de quadrados [Dorogovtsev et al. (2003)];

4. Rede totalmente conectado com $k = 7$
   - A distribuição de densidade, degenera para uma linha onde $\rho(\lambda) = cte = 1$. Como todos os vértices da rede apresentam o mesmo grau de conectividade, e as arestas apresentam o mesmo peso, os valores de $\rho(\lambda)$ são iguais para todos os vértices;
5. Distribuição de Wigner para um sistema de partículas não conectadas, arranjadas de forma aleatória
   - A distribuição de densidade, para o modelo teórico de Wigner [Wigner (1955)] ajusta-se a uma distribuição semicircular, apresentada na equação 3.10 (linha fina sólida), quando número de vétices tende ao infinito.

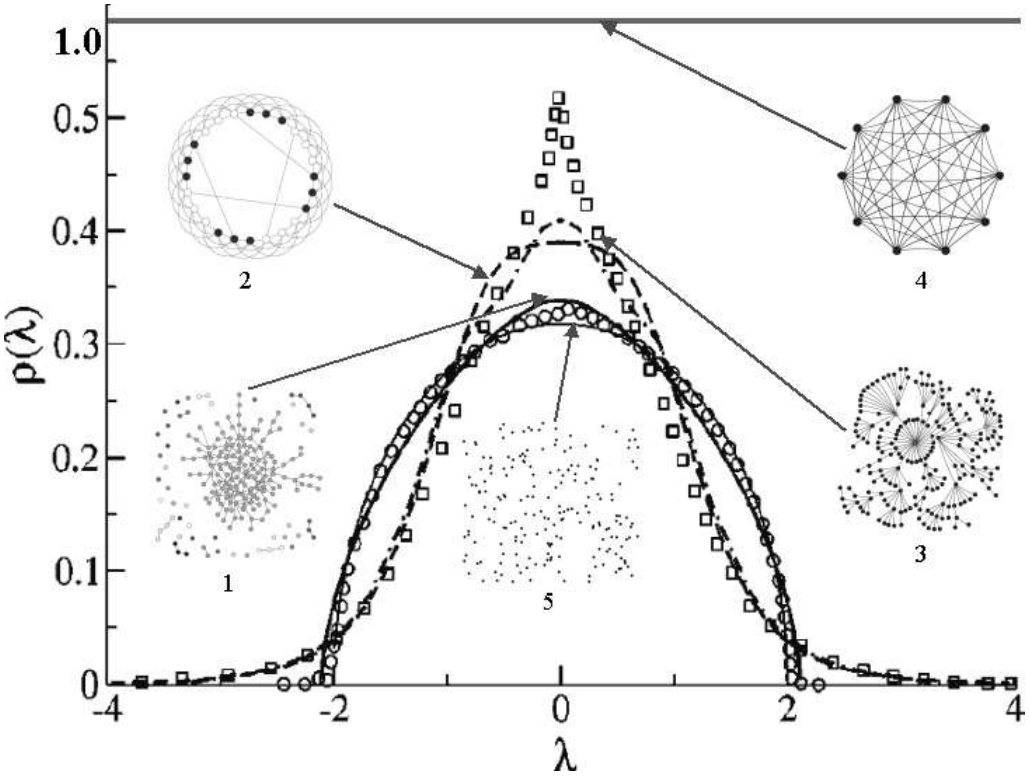


Figura 3.15 *– Distribuição de densidade - $\rho(\lambda)$, de cinco modelos: (1) Rede de Erdös e Rényi, com grau médio - $\langle k \rangle = 10$ – (i) Perfil do modelo contínuo (linha sólida espessa), (ii) Grafo com 20.000 vértices (círculos); (2) Modelo de Watts e Strgatz (*Small World Effect*) com $k = 4$; (3) Modelo de Barabási-Albert (Rede livre de escala) com 70.000 vértices; (4) Modelo de rede totalmente conectada; (5) Modelo teórico de Wigner para sistema totalmente aleatório [Farkas et al. (2001)]*

.

Da comparação dos espectros de densidade destes modelos tem-se a percepção das diferentes características apresentadas por cada um destes modelos clássicos. O modelo randômico

clássico apresenta a distribuição dos graus de conectividade dos seus nodos, seguindo o modelo de Poisson [Dorogovtsev et al. (2003)]. Contudo, o espectro de densidade – $\rho(\lambda)$, dos valores singulares de sua matriz de adjacências apresenta uma parte central elevada que difere da distribuição semi-circular de Wigner.

A distribuição de densidade do espectro dos grafos livres de escala tende para a forma triangular, o que difere fortemente da distribuição semi-circular do caso randômico. O modelo de Barabási-Albert – BA, analisado na figura 3.15 [Farkas et al. (2001)], apresenta a distribuição dos graus de conectividade com decaimento em lei de potência com expoente $\gamma = 3$ [Albert e Barabasi (1999), Barabasi et al. (1999)]. Desta forma, o espectro de um grafo similar ao modelo BA deve apresentar um comportamento similar ao deste [Farkas et al. (2001), Dorogovtsev et al. (2003)]. O espectro de densidades do modelo BA mesmo apresentando forma triangular, estaria mostrando um decaimento à direita, em lei de potência [Dorogovtsev et al. (2003)].

O decaimento em lei de potência do espectro de densidade dos valores singulares de um grafo $\rho(\lambda) \propto \lambda^{-\delta}$, é uma importante característica do espectro de uma rede livre de escala [Dorogovtsev et al. (2003)]. Simulações feitas com o modelo de Barabási-Albert, tendo expoente $\gamma = 3$ [Farkas et al. (2001)], revelaram um espectro de densidade com decaimento em lei de potência com o expoente do valor singular $\delta = 5$ [Dorogovtsev et al. (2003)].

No caso do modelo de Barabási-Albert – BA, a presença de fortes singularidades no espectro de densidades dos valores singulares – $\rho(\lambda)$, mostra a existência de várias partes bem distintas. A porção mais homogênea ("*bulk*") do espectro de densidades – o conjunto de valores singulares $\{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n\}$ – converge para uma função discreta contínua que apresenta formato triangular e com a cauda apresentando decaimento em lei de potência [Farkas et al. (2001)].

Ainda segundo Farkas [Farkas et al. (2001)], o decaimento em lei de potência do espectro de densidades seria causado por vetores singulares associados aos vértices de maior conectividade – "*hubs*" – que são característicos das redes do modelo BA. Desta forma, o espectro do modelo BA converge para uma forma triangular próximo ao centro da distribuição, enquanto as arestas representantes da porção homogênea da rede decaem mais devagar. O primeiro valor singular destaca-se do resto do espectro, e mostra um taxa de crescimento anômala. Os valores singulares com maiores magnitudes pertencem aos vetores singulares localizados nos vértices com vários vizinhos.

Segundo Farkas, a relação:

$$R = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_N}$$

comporta, sistemas grandes, de tal forma que em redes randômicas R converge para um valor constante, enquanto que para redes livres de escala, R decai seguindo uma função em lei de potência. Por outro lado, para redes "*small-world*" apresentam valores menores de R. Neste caso, λ_1 não se destaca do resto do espectro, o que é uma conseqüência da estrutura quase periódica da rede. Em suma, o índice R seria apropriado para distinguir três diferentes tipos de estruturas: (i) – redes periódicas ou quase-periódicas – "*small-world*"; (ii) – redes não

correlacionadas e não periódicas; (*iii*) – redes fortemente correlacionadas e não periódicas – livres de escala.

Da análise dos fundamentos da análise espectral dos valores singulares de uma rede, vale ressaltar as seguintes propriedades:

1. Uma distribuição com perfil em semi círculo é um indicativo de uma estrutura sem hierarquia, o que é mais acentuado em uma rede randômica;

2. As singularidades observáveis em um espectro indicam a presença de vértices de maior conectividade – "*hubs*";

3. O espectro de uma rede livre de escala apresenta três elementos distinguíveis: O primeiro valor singular – λ_1; O grupo de valores singulares centrais – com formato triangular; O grupo de valores que representa a parte mais homogênea da rede - cauda.

Segundo Farkas [Farkas et al. (2001)], uma distribuição de espectro que apresenta ausência de máximos locais, *i.e.* a ausência de singularidades, mostra a ausência de uma estrutura hierarquicamente ordenada. Disto conclui-se que – do ponto de vista do espectro de uma rede – o grande número de triângulos que é uma característica básica das redes "*small-world*", é uma propriedade que se preserva mais que a regularidade ou a periodicidade, mesmo se o nível de aleatoriedade aumente. Este indício, segundo Farkas [Farkas et al. (2001)], encontra respaldo nos resultados apresentados por outros pesquisadores[3], onde o grande número de pequenos ciclos parece ser a propriedade fundamental das redes "*small-world*". Como uma aplicação, o elevado número de pequenos ciclos resulta em um processo típico de difusão nesta classe de redes.

Considerando individualmente os valores singulares de uma rede, o maior valor singular relativo à matriz de adjacências desta rede - λ_1, tem sido utilizado por diferentes autores como um fator quantitativo que representa muitas propriedades importantes para o estudo de uma variedade de processos dinâmicos da rede [Restrepo et al. (2006a)]. Dentre estas propriedades, podem ser citadas:

(1) Em um sistema heterogêneo, cuja dinâmica mostra-se caótica e/ou periódica, onde os elementos são interligados por uma rede de conexões, a força de acoplamento crítica que induz a emergência de coerência é proporcional a $1/\lambda_1$ [Restrepo et al. (2006b)];

(2) Em uma comunidade, a probabilidade crítica de contágio de uma doença, que pode se tornar epidêmica, escala conforme $1/\lambda_1$ [Wang et al. (2003)];

(3) Em uma rede, a estabilidade dinâmica da rede escala conforme λ_1 da respectiva matriz Jacobiana [Brede e Sinha (2005)];

Diferentes autores [Newman (2002), Newman (2003b), Newman (2003a), Brede e Sinha (2005)], têm mostrado uma correlação positiva entre $\langle\lambda_1\rangle$ e o caráter associativo apresentado

[3]Referências não apresentadas.

pela rede[4]. Isto significa que numa rede com caráter associativo mais pronunciado, menor será a densidade crítica de arestas entre os vértices para que surja o primeiro componente gigante. Numa rede com características associativas (tal como nos sistemas sociais), os vértices de maior grau tendem a estabelecer, preferencialmente, relações com outros vértices similares, tendendo para a formação do que, em epidemiologia, é denominado de grupo-núcleo ("*core group*").

Em um grupo-núcleo, a densidade de arestas é alta quando comparada com a rede como um todo, uma vez que os vértices neste grupo têm grau maior que o da média dos vértices da rede. Por outro lado, nas redes com caráter dissociativo, esta "transição" de fase só ocorre quando a densidade de arestas do grafo é maior [Newman (2002), Newman (2003a)]. A presença de um grupo núcleo em uma rede, faculta uma rápida percolação de informação para toda a rede, em uma densidade mais baixa que aquela que seria necessária para a emergência de tal fenômeno em outras redes. Ao mesmo tempo, como este componente gigante da rede é restrito, a densidade de arestas em outros pontos da rede, fora do grupo núcleo, é baixa e desta forma, o componente gigante não se estende até estas regiões. Então, o componente gigante da rede fica confinado ao núcleo da rede e não pode crescer tanto como ocorre em uma rede dissociativa.

Ao mesmo tempo, se for tomado como referência um grupo de redes, com o mesmo grau médio de distribuição de arestas por vértice – $\langle k_i \rangle$, com diferentes índices de associabilidade, existe uma acentuada variação na resiliência[5] da rede face a sua associabilidade. Quanto maior o caráter associativo da rede, maior será o número de vértices, do grupo central, a ser removidos para produzir a desintegração do componente gigante. A título de exemplo [Newman (2003a)], considera-se duas redes $\mathcal{G}_1$ e $\mathcal{G}_2$, ambas com $\approx 10^7$ vértices, com $\langle k_i \rangle = 5000$ apresentando índices de associação $r_1 = 0,2$ e $r_2 = -0,2$ (onde $-1 \leq r \leq 1$). Considerando que o componente gigante de $\mathcal{G}_1$ abarque 30% dos vértices, enquanto o componente gigante de $\mathcal{G}_2$ abarca 70% dos vértices, ainda assim é necessário remover dez vezes mais vértices do grupo-núcleo de $\mathcal{G}_1$, em relação a $\mathcal{G}_2$, para iniciar a desintegração deste núcleo, mesmo tendo $\mathcal{G}_2$ um núcleo duas vezes maior.

Este fenômeno seria típico de uma rede com elevado grau de associabilidade, onde no grupo-núcleo vértices com alto grau de conectividade encontram-se fortemente interligados. Tal núcleo proveria robustez à rede ao concentrar todos os vértices mais sensíveis da rede, em uma porção da mesma. Mesmo que a remoção destes vértices continue sendo a melhor forma de destruir a conectividade da rede, tal estratégia mostra-se ineficiente visto que a remoção de um destes vértices em nada afeta a capacidade percolação da rede [Newman (2002), Newman (2003a)]. Em uma rede com caráter dissociativo, por outro lado, tal estratégia é a mais efetiva, já que estando espalhados pela rede, estes "*hubs*" só estabelecem relações mútuas se

[4] O caráter associativo "*assortativity*" de um vértice, em uma rede, refere-se à sua tendência prefetencial de estar ligado a outros vértices que compartilham algum atributo em comum. Quando um vértice liga-se a outros vértices similares, de algum modo, a ele, diz-se que esta relação é associativa. Caso contrário, diz-se que ela é dissociativa.

[5] Resiliência é um conceito que vem da física e significa a capacidade de um objeto recuperar-se, de se moldar novamente depois de ter sido comprimido, expandido ou dobrado, voltando ao seu estado original.

estas forem intermediadas por outros vértices de menor conectividade.

Quanto ao aspecto de estabilidade da rede, um elevado valor de λ_1 necessariamente implica em uma forte tendência à instabilidade da rede [Brede e Sinha (2005)]. Isto implica em dizer que, nas redes com alta associatividade, as perturbações ambientais surtem pouco efeito na tendência de percolação de informação através da rede [Newman (2003a), Brede e Sinha (2005)]. Contudo, tal tendência à instabilidade não implica em dizer que a rede tende a se degenerar, o que contradiz o que já foi dito nos paragrafos anteriores. Tal instabilidade refere-se à tendência do sistema de mudar de uma conformação estável para uma outra das possíveis conformações alcançáveis dentro do seu espaço de fase[6]. Desta forma, um alto valor de λ_1 informa que o sistema é capaz de mudar de um estado para outro com muita facilidade. A seu turno, as redes com tendência dissociativa mostram-se mais resistentes aos efeitos das flutuações dinâmicas que podem induzir uma mudança de estado [Newman (2002)].

Mais à frente, na seção 5.3, serão apresentados os resultados derivados da aplicação deste estudo na análise das redes formadas pelas interações não-covalentes entre os átomos das proteínas estudadas. Pode-se entretanto especular que a necessidade dos sistemas vivos de evoluir de forma gradativa e com reduzida intensidade de flutuações (homeostase), é o que faz com que prevaleçam as redes dissociativas nas diversas dimensões do mundo biológico.

[6]Sumariamente, o espaço de fases de um sistema refere-se ao conjunto de estados ou conformações que o sistema pode apresentar.

Capítulo 4

Materiais e Métodos

Este capítulo descreve os métodos e os dados utilizados nos estudos citados neste trabalho. Inicialmente são apresentados os conjuntos de proteínas utilizadas nas análises. São também revistos os conceitos apresentados anteriormente no capítulo 3 referente às análises estatística e espectral das redes. Entretanto, a ênfase deste capítulo está na descrição dos métodos desenvolvidos para lidar com os problemas de oclusão entre átomos.

4.1 Proteínas em estudo

Estes estudos focam duas famílias de proteínas: as globinas e as serinoproteases.

Globinas são proteínas globulares que apresentam a habilidade de capturar oxigênio e suprir as células do organismo com este oxigênio [Lehninger et al. (2007)]. O padrão de enovelamento das globinas foi inicialmente caracterizado pela ocorrência de oito α-hélices, designadas de A a H [Perutz (1970), Lesk e Chotia (1980)]. Lesk e Chothia [Lesk e Chotia (1980)],mostraram que somente dois resíduos são invariantes em 700 globinas de diferentes organismos: a histidina proximal na hélice F (His F8) e uma fenilalanina no arco CD (Phe CD1). Globinas ocorrem nos três reinos da vida [Vinogradov et al. (2006)], sendo classificadas em globinas de domínio simples e globinas quiméricas. Este último grupo comporta exemplares como as flavohemoglobinas que apresentam um domínio de ligação de FAD em sua porção C-terminal, e globinas cujo domínio C-terminal é variável e, onde se acoplam sensores reguladores de genes. Um estudo apresentado em [Vinogradov et al. (2006)], mostra que 25% dos *archaea* apresentam globinas, sendo que esta proporção cresce para 65% dos genomas bacterianos. A presença e ocorrência de globinas mostram-se positivamente correlacionadas ao tamanho do genoma. Globinas parecem estar ausentes em *Chlamydia* , *Lactobacillales*, *Mollicutes*, *Rickettsiales*, e *Spirochaetes*. Globinas de domínio único ocorrem em metazoários, enquanto flavohemoglobinas são encontradas em fungos, diplomonadidas e micetozoas. Contudo, mais de 90% dos eucariontes têm globinas: enquanto o nematóide *Caenorhabditis* apresenta 33 diferentes globinas, nenhuma globina ocorre em parasitas eucariontes unicelulares como *Encephalitozoon*, *Entamoeba*, *Plasmodium* e *Trypanosoma*. Uma vez que as globinas em organismos outros que não os animais, atuam como enzimas ou senso-

res [Vinogradov et al. (2006)], isto sugere que a evolução da função de transporte de oxigênio pelas globinas acompanhou a emergência dos animais multicelulares.

As serinoproteases, ou serino-endopeptidases, constituem uma classe de peptidases (enzimas que quebram ligações peptídicas em proteínas), que se caracterizam pela presença de um resíduo serina no sítio ativo da enzima. serinoproteases são agrupadas em grupos que partilham homologia estrutural, podendo ser sub-agrupadas em famílias que partilham alta homologia de seqüência. Os principais grupos incluem as quimotripsinas, as subtilisinas, as α/β hidrolases, e peptidases de sinal. As serinoproteases participam de um vasto leque de funções nos organismos, incluindo a coagulação sanguínea, resposta imunidade, processos inflamatórios, bem como contribui com as enzimas digestivas em eucariotos e procariotos. O principal elemento do mecanismo catalítico das quimotripsinas e subtilisinas é a Tríade Catalítica. A tríade está localizada no sítio ativo da enzima, onde ocorre catálise, sendo preservada em todas serinoproteases. A tríade é uma estrutura coordenada composta por três aminoácidos: histidina (His57), serina (Ser195) (daí o nome "serinoproteases") e ácido aspártico (Asp102). Localizados muito perto uns dos outros, e perto do centro da enzima, estes três resíduos principais desempenham um papel essencial na capacidade catalítica destas proteases.

Estas famílias foram selecionadas devido ao fato de serem famílias que têm sido alvo de freqüentes estudos já publicados ao mesmo tempo em que são objeto de estudos também em nosso grupo de pesquisas. Membros representativos de ambos grupos foram selecionados, a partir do *Protein Data Bank* - PDB [Berman et al. (2000)], de forma garantir uma amostra representativa das características estruturais das famílias selecionadas.

Os códigos PDB das proteínas selecionadas como representativas das globinas são: 1A6G 1A9W 1ASH 1B0B 1BIN 1BZP 1C40 1CG5 1D8U 1DLW 1DLY 1DWT 1ECD 1EMY 1F5O 1FAW 1FDH 1FHJ 1G09 1GCV 1GDJ 1H97 1HBR 1HDS 1HLB 1HLM 1I3D 1IDR 1IT2 1ITH 1JF3 1JF4 1KR7 1LA6 1LHS 1MBS 1MWD 1MYT 1NGK 1NS9 1NXF 1OJ6 1OUT 1Q1F 1QPW 1RQ3 1RTX 1S5Y 1SI4 1SPG 1TU9 1UC3 1UVX 1V4W 1V5H 1VHB 1WMU 1XQ5 2FAL 2MM1.

Os códigos PDB das proteínas selecionadas como representativas das serinoproteases são: 1ANE 1AQ7 1ARB 1BEF 1BIO 1BRA 1BRU 1C5L 1C5Y 1CA8 1CGH 1DLE 1DUA 1DY9 1EAX 1ELT 1EQ9 1ETR 1FON 1FUJ 1G2L 1GVK 1GVZ 1H8D 1H8I 1HPG 1HXE 1HYL 1K2I 1KLI 1LTO 1MBM 1NN6 1NPM 1NTP 1OP0 1OP2 1OWE 1P3C 1PPF 1PPZ 1PQ7 1QNJ 1QY6 1S0R 1S83 1SGP 1SGT 1SQT 1SSX 1TAW 1TE0 1TON 1TRN 1VR1 1VZQ 1WCZ 1WYK 1Y8T 2AIQ 2SFA 2SGA 2TBS 3RP2 5PTP 7LPR.

Estes critérios asseguram que este estudo pode ser generalizado para todas as proteínas destas famílias. Este trabalho não leva em conta outras famílias de enovelamento. O conjunto de átomos estudados é composto por todos os átomos (incluindo os átomos dos grupos ligados a estas proteínas, como Fe, hidrogênio, água, etc.) presentes nos arquivos do *Protein Data Bank* - PDB. O raciocínio subjacente a estes critérios considera que não somente os átomos dos resíduos, mas também outros átomos, como exposto previamente, podem exercer um papel relevante no concerto sistêmico das estruturas das proteínas.

Para tanto, os arquivos PDB foram inicialmente tratados para incorporar os átomos de hidrogênio que compõem as proteínas mas que não se acham presentes nos arquivos PDB. Para este fim foi utilizado o programa REDUCE [Worda et al. (1999)] com a finalidade de agregar os átomos de hidrogênio às estruturas moleculares em estudo. Ao mesmo tempo em que os átomos de hidrogênio são agregados, o REDUCE promove uma otimização da disposição espacial das cadeias laterais dos resíduos constituintes das proteínas. Como será explicado nas próximas seções, todas as "proteínas" foram "solvatadas" e passaram por um processo de minimização e de dinâmica molecular por $\approx$ 300 picosegundos, com o uso do NAMD [Philips et al. (2005)].

Foram estudadas, para cada proteína, todas as interações par a par entre os átomos de forma a verificar se aquelas interações poderiam, ou não, existir no interior da proteína. Com este objetivo foram adotados os critérios apresentados por Sobolev [Sobolev et al. (1999)]. O número das arestas apresenta uma distribuição tal que induz o entendimento onde o uso de grafos gerados desta forma pode muito bem mimetizar a realidade estrutural das proteínas.

Estudos propedêuticos relativos a estas estruturas foram conduzidos objetivando determinar se o arranjo estrutural destas proteínas assemelha se mais a uma rede do tipo "*small-world*" ou "*scale-free*". O conceito de "*neighborhood*" é também introduzido de forma a estipular, em termos quantitativos, o que deve ser considerado uma interação curta ou longa.

Vale ressaltar que, com exceção dos softwares de visualização, de plotagem de grafos, de cálculo da SVD e aqueles relacionados à solvatação de proteínas, todos os softwares e bancos de dados utilizados neste trabalho foram codificados e implementados pelo autor.

4.2 Adoção do modelo de redes na análise da estrutura das proteínas

Considere a estrutura primária de uma proteína $\mathcal{P} = \{A_1, \ldots, A_{(i-1)}, A_i, A_{(i+1)}, \ldots, A_{N_A}, \}$, onde a seqüência $\{A_i \mid i = 0 \ldots N_A\}$ é a seqüência dos resíduos de aminoácidos constituintes da proteína e N_A é o número total de átomos desta proteína. Os átomos pertencentes a um mesmo resíduo R_i interagem com os átomos dos resíduos adjacentes tal como se segue:

A figura 4.1(a) representa a situação onde átomos do resíduo R_i interagem com átomos de outros resíduos pertencentes à mesma proteína, começando com os resíduos contíguos e indo até o resíduo mais distante. Nesta situação define-se a adjacência ou"vizinhança" como tendo valor 0 (zero). A figura 4.1(b) exemplifica a situação onde átomos de R_i interagem com átomos de outros resíduos presentes na proteína, com exceção dos átomos presentes nos resíduos contíguos, compreendendo assim o intervalo de $(i-1)^{o}$ a $(i+1)^{o}$ resíduos. Esta situação é denominada como "vizinhança" 1.

O conceito de "vizinhança", aqui apresentado, será útil quando da análise das redes interações não covalentes identificadas no conjunto das proteínas estudadas. Estas redes de interações irão gerar grafos que representam como os átomos dos resíduos presentes em

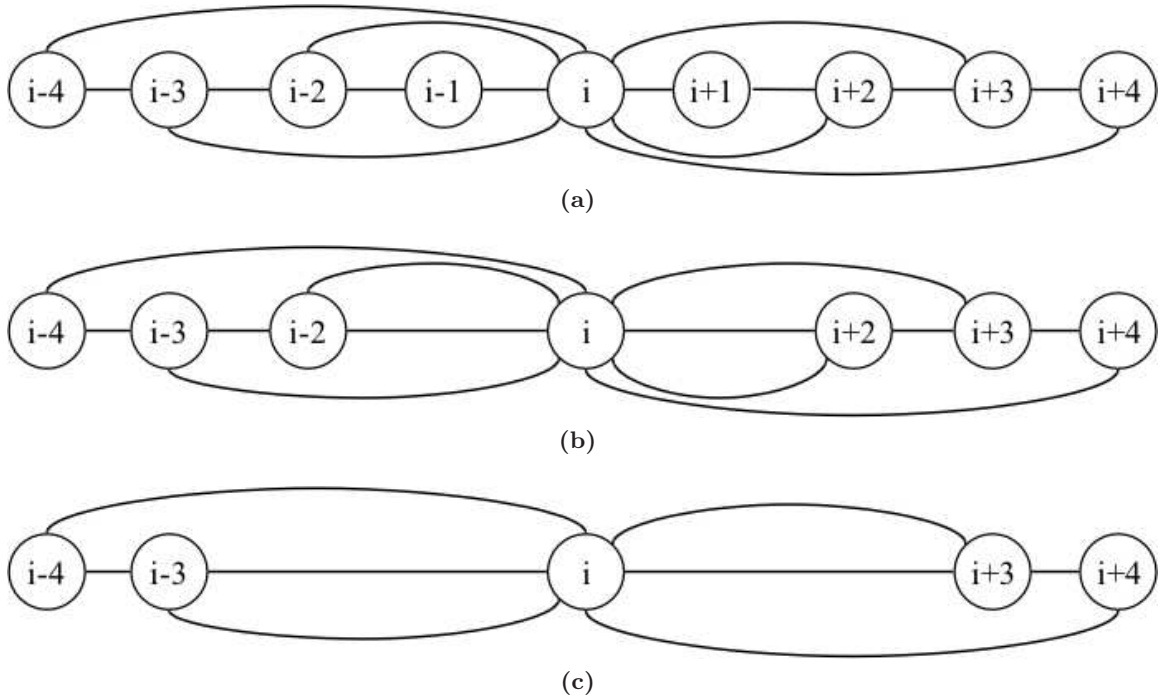


Figura 4.1 *– Interações entre resíduos, definem níveis diferentes de vizinhanças: (a) vizinhança com* k = 1*; (b) vizinhança com* k = 2*; (c) vizinhança com* k = 3*. Vizinhanças com* $k \geq 4$ *são consideradas como "não locais", dentro do escopo deste trabalho.*

uma proteína se relacionam com os átomos dos demais resíduos desta proteina, dos grupos prostéticos e do solvente onde a proteína se encontra imersa. Estes limites irão determinar a classificação das interações como sendo de curto ou de longo alcance.

A figura 4.2 [Atilgan et al. (2004)] exemplifica uma interação interna de longa distância dentro da rede de uma proteína. O diagrama em *ribbon* representa o *backbone* da cadeia peptídica. Os carbonos α dos resíduos de aminoácidos são mostrado os como esferas. A ilustração mostra os contatos de um resíduo os quais são denotados por círculos pontilhados no *core* da proteína, indo além de cinco graus de separação. O código de cor em graus é o seguinte: primeiro (vermelho), segundo (laranja), terceiro (amarelo), quarto (verde) e quinto (azul). Esquematicamente são mostradas as variações no número de ligações entre dois nodos, pela variação da espessura das linhas. Este tipo de padrão é o que se espera encontrar quando forem explorados os diferente grupos de interação quando variados o uso níveis de adjacências em estudo.

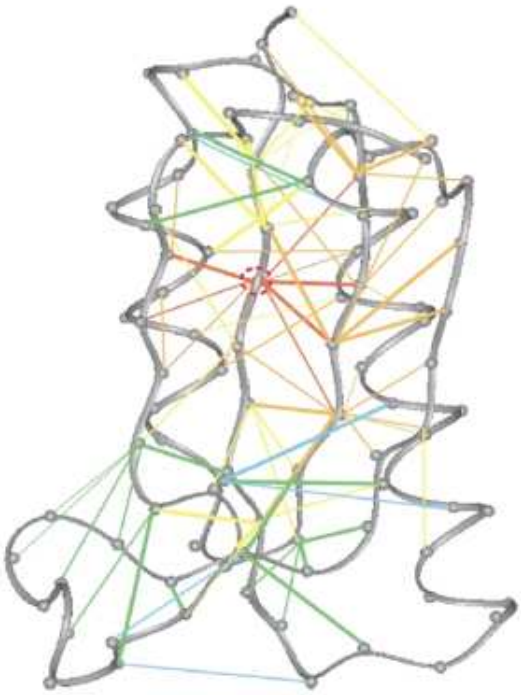


Figura 4.2 – *Interações de longa distância na estrutura de uma proteína*

4.3 Construção da Representação Gráfica da Estrutura das Proteínas

A Representação Gráfica da Estrutura das Proteínas - RGEP, é construída a partir das coordenadas tridimensionais dos átomos, obtidas dos arquivos PDB. Cada proteína neste conjunto é representada como um grafo constituído por um conjunto de vértices e arestas. Cada átomo da proteína é representado por um vértice. As interações estabelecidas por estes vértices são representadas pelas arestas destes grafos. À cada aresta está associado um valor que representa a energia de interação não covalente entre os átomos que a definem.

A energia de interação não covalente entre dois átomos A_i e A_j é dada por:

$$E_{ij} = E_{ij}^{C} + E_{ij}^{LJ} \tag{4.1}$$

onde E_{ij} é a energia total associada à interação entre os átomos $A_i, A_j \,|(i \neq j)$ tal que A_i e A_j não estejam no mesmo resíduo de aminoácido. A energia total associada à cada interação atômica tem dois componentes: a energia derivada do potencial de Coulomb (E_{ij}^{C}):

$$E_{ij}^{C} = cte\ \frac{q_i\ q_j}{\epsilon_{ij}\ r(ij)} \tag{4.2}$$

em $[kcal/mol]$ onde:

cte - Constante de proporcionalidade
q_i, q_j - Carga em elétron-*charges*
ϵ_{ij} - Constante dielétrica aparente do meio
$r(ij)$ - Distância em Angstroms

e a energia associada ao potencial de Lennard-Jones (E_{ij}^{LJ}):

$$E_{ij}^{LJ} = \epsilon_{ij} \left[\frac{m}{(n-m)} \left(\frac{r_{eqm}(ij)}{r(ij)}\right)^n - \frac{n}{(n-m)} \left(\frac{r_{eqm}(ij)}{r(ij)}\right)^m\right] \tag{4.3}$$

onde:

ϵ_{ij} - Constante dielétrica aparente do meio entre i e j
n - Coeficiente Alto (usualmente 12)
m - Coeficiente Baixo (usualmente 6)
$r_{eqm}(ij)$ - Distância parâmetro para uma interação entre i e j
$r(ij)$ - Distância entre i e j

como descrito em AMBER98 [Fox (1998)].

Inicialmente seja a proteína P formada por uma seqüência de N_R resíduos $\{R_1, R_2, R_3, \ldots, R_{N_R}\}$. Seja um resíduo R qualquer formado por uma seqüência de N_A átomos $\{A_1, A_2, A_3, \ldots, A_{NA}\}$. O número total de átomos em uma proteína N_{AP} é calculado como:

$$N_{AP} = \sum_{j=1}^{N_R} N_{A_j} \tag{4.4}$$

Uma proteína P pode ser vista como sendo formada por um conjunto de N_{AP} átomos $\{A_1, A_2, A_3, \ldots, A_{N_{AP}}\}$. O algoritmo 1(página 65), é usado para identificar os pares de átomos (A_i , A_j) que estabelecem uma interação.

Neste procedimento nenhum valor arbitrário de corte para a distância entre o par (A_i , A_j) é necessário, já que o critério mais relevante para estabelecer uma interação é a presença ou não de um terceiro átomo A_k que possa esconder, totalmente ou parcialmente, A_j de A_i. Então, se A_j é "visível" por A_i existe uma interação e a energia de interação é calculada.

Por razões práticas limitou-se a distância de corte em 10 Å, de forma evitar um desnecessário esforço computacional, já que átomos apartados além deste limite são tidos como incapazes de interagir [Desjarlais e Handel (1995), Dominy e Brooks (1999)]. Por um lado, o potencial de Lennard-Jones é o usualmente calculado deste limite para baixo. Ao mesmo tempo, outros valores de corte além deste limite foram estudados, mas nenhuma diferença apreciável foi observada quando os valores gerados foram comparados com os obtidos com os parâmetros correntes tal como apresentado na figura 5.4 (dados não apresentados).

Neste estudo, para os cálculos focados nos átomos presentes nas proteínas, a constante dielétrica foi fixada em $\epsilon_{ij} = 4$, como adotado em Schutz e Warshel [Schutz e Warshel (2001)].

Algoritmo 1: Identificação dos pares de átomos (A_i , A_j) que estabelecem uma interação não covalente

Entrada: Coordenadas dos átomos presentes na proteína em análise

Saída: Tabela dos átomos presentes na proteína em estudo, que estabelecem interações não covalente

início

para cada $A_i \in P$ **faça**

para cada $A_j \in P | A_j \neq A_i \wedge ((A_i, A_j))$ *não pertencente a um mesmo resíduo* R **faça**

Avalia a existência e a natureza do contato entre (A_i, A_j)

para cada $A_k \in P | (A_k \neq A_i \wedge A_k \neq A_j)$ **faça**

// A_f é a área exposta de A_j

$A_f = \bigcup$ área de A_j que não está oclusa de A_i por A_k

// E_{ij} é a energia de interação não covalente entre A_i e A_j

$E_{ij} = \left(E^C(A_i, A_j) + E^{LJ}(A_i, A_j)\right)(A_f/A_j)$

fim

4.4 Cálculo de Oclusão entre Átomos

4.4.1 Introdução

Os princípios associados ao mundo das proteínas, que resultam em um balanço de estabilidade e flexibilidade, enquanto mantém suas funcionalidades, é um mecanismo ainda não perfeitamente entendido. O mecanismo chave para a termoestabilização de uma proteína, parece ser a otimização das interações entre os átomos dentro desta proteína. Os princípios de estabilidade e enovelamento de proteínas têm sido estudados através de uma variedade de métodos de análise aplicados à um grande número de diferentes estruturas de proteínas.

Alguns estudos teóricos relativos às estruturas das proteínas, e outros métodos empíricos, vêm sendo usados para entender a estabilidade das proteínas. Desde então, diferentes tipos de experimentos têm sido feitos para entender a estabilidade das proteínas, e se existe algum

resíduo específico que possa ser identificado como tendo um papel relevante neste fenômeno [Onuchic (2004), Fersht e Daggett (2002)].

Redes representando estruturas de proteínas têm sido modeladas a partir de diferentes conceitos sobre a natureza dos nodos e arestas [Vendruscolo et al. (2002); Bagler e Sinha (2005)]. Estes estudos prévios focaram no entendimento das propriedades das redes tais como distâncias entre vértices e coeficientes de aglomeração entre outras propriedades. De forma similar, neste trabalho as estruturas das proteínas são modelada como redes e um grafo representativo da estrutura da proteína-(RGPS), é construído definindo-se os átomos constituintes dos resíduos de aminoácidos como sendo vértices e as interações não-covalentes existentes entre eles como sendo as arestas.

A modelagem da proteína como grafo tem sido utilizada na identificação de aglomerados de átomos que podem estabilizar a estrutura das proteínas. Um aspecto importante de tais grafos é a definição das arestas baseada nas energias de interação entre os átomos em uma proteína. De forma a lidar com a dependência dos valores de restrição ("$cutoffvalue$") das distâncias de interação entre átomos usadas na construção destes grafos [Vendruscolo et al. (2002);Greene e Higman (2003),Atilgan et al. (2004), Bagler e Sinha (2005)], um método de análise baseado no estudo da oclusão entre os átomos é apresentado. Tal abordagem leva a redes subjacentes às proteínas que podem ser diferentes daquelas previamente relatadas em outros trabalhos.

Uma outra característica importante que deve ser identificada é o fato de que os átomos mais conectados parecem ser determinantes para a estabilidade das proteínas. Tais elementos constituem aquilo que se denomina como vértices concentradores ou simplesmente "*hubs*". Um importante, mas não exclusivo aspecto destes elementos é que eles, em muitas redes do mundo real, são conhecidos por serem os pontos frágeis das redes que se mostram susceptíveis a ataques (ou mutações) dirigidos a estes [Albert e Barabasi (2002)]. Desta forma é possível esperar que as mutações específicas nos resíduos "*hubs*" possam levar a uma alteração da estrutura da proteína que, por sua vez, podem prejudicar os aspectos dinâmicos e de adequação ambiental desta proteína.

Este estudo foca no entendimento dos princípios estruturais das proteínas considerando-as como redes complexas de interações não-covalentes. Aqui foi adotada uma nova abordagem onde as estruturas das proteínas podem ser modeladas com base nas energias das interações entre os átomos e considerando a oclusão espacial dos átomos entre si como tendo um importante papel nas características destas redes. Neste trabalho também é demonstrada a necessidade da solvatação das proteínas antes das análises das interações atômicas de forma evitar resultados não realísticos, o que ocorre quando nenhum valor limite é utilizado para restringir a distância entre os átomos quando as interações são calculadas. Em especial, os resíduos expostos ao solvente tendem a apresentar um excessivo número de interações com os outros átomos expostos na superfície da proteína quando não solvatada. A solvatação da proteína mimetiza de forma mais realística o escopo onde este tipo de interação acontece.

4.4.2 A Necessidade de Solvatação

Nas rodadas iniciais de análises, as proteínas foram modeladas como redes utilizando os dados dos átomos diretamente dos arquivos PDB. Estes dados foram utilizados para a modelagem das redes. Entretanto, os resultados iniciais mostraram algumas características irreais quando nenhum valor de corte foi utilizado para limitar a distância entre os átomos de forma a avaliar as interações entre eles. Quando um valor arbitrário de corte para as distâncias foi utilizado o número de arestas ligando os "*hubs*" tende a decair. Entretanto fica a pergunta: qual deveria ser o valor de corte mais apropriado a ser aplicado? Diferentes estratégias têm sido usadas visando lidar com este problema [Greene e Higman (2003), Amitai et al. (2004), Atilgan et al. (2004), Rao e Caflisch (2004), Bagler e Sinha (2005), Brinda e Vishveshwara (2005), Kundu (2005), del Sol e O'Meara (2005), Kundu (2005), Aftabuddin e Kundu (2006), Alves e Martinez (2006), Higman e Greene (2006), del Sol et al. (2006b), del Sol et al. (2006a), Atilgan et al. (2007), Jiao et al. (2007)]. Nas rodadas iniciais de testes as moléculas eram consideradas como estando no vácuo, ao mesmo tempo em que nenhum valor de corte era adotado. Os átomos dos resíduos expostos na superfície das proteínas tenderam a apresentar um excessivo número de interações. Estas ligações se formavam principalmente com os átomos dos demais resíduos expostos na superfície destas proteínas.

Objetivando a solução deste problema uma solvatação prévia das proteínas foi realizada numa segunda rodada de análises das interações atômicas, de forma evitar resultados não realísticos. A solvatação das proteínas mimetiza de forma mais realista o escopo onde estas interações ocorrem. O aplicativo utilizado para solvatar as proteína neste estudo, foi o SOLVATE do NAMD utilizando os parâmetros padrão [Philips et al. (2005)].

Os resultados observados após a solvatação da proteína, mostraram uma drástica redução no número de ligações entre os átomos, sem nenhum tipo de intervenção mesmo na ausência de qualquer valor de corte para a das distâncias interatômicas. Conseqüentemente, o comportamento das redes resultantes tornou-se mais aceitável, mostrando "*hubs*" sem um número excessivo de ligações, o que evita resultados não realísticos, como será mostrado na seção 5.1.

4.4.3 O Problema da Oclusão

Fisicamente é não realístico considerar as interações não-covalentes entre diferentes átomos de uma proteína sem levar em conta a interveniência existente entre estes mesmos átomos dentro da proteína. Um dado par de átomos A_i e A_j mesmo estando espacialmente próximos dentro de uma proteína, podem não interagir devido à presença de outros átomos em suas imediações que poderiam estar obstruindo tal interação. Para entender melhor este problema considera-se inicialmente cada átomo em estudo como sendo uma partícula eletricamente carregada com carga positiva, tal como mostrado na figura 4.3.

Nesta figura as setas indicam o sentido das linhas de força relativas ao campo elétrico emanado por esta carga. Um segundo átomo carregado presente nas vizinhanças, agora com carga negativa, iria estabelecer uma interação de natureza eletrostática entre estas cargas, tal como representada na figura 4.4(a).

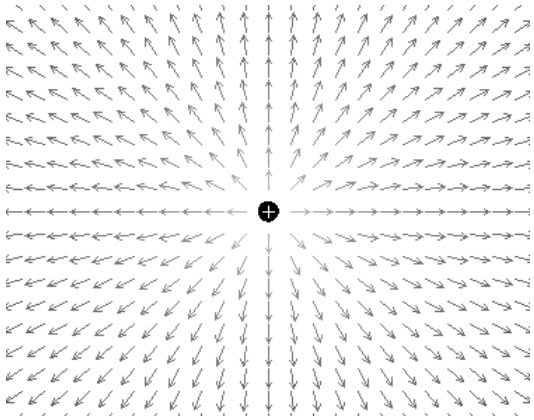


Figura 4.3 – *Um átomo qualquer considerado como uma partícula carregada.*

Contudo, quando uma terceira carga A_k, entra nas imediações deste par de cargas, ocorre um rearranjo na distribuição das linhas de força desta interação. A figura 4.4(b) mostra como as linhas de força se rearranjam na presença de uma terceira carga. É possível observar ainda na figura 4.4(b) que praticamente metade das linhas de força relativas à interação entre A_i e A_j mantêm-se inalterada, enquanto a outra metade foi redirecionada para a interação entre A_i e A_k.

Na figura 4.4(c) uma quarta carga é inserida nas imediações do par de átomos A_i e A_j. Mesmo havendo alteração das linhas de força, é possível perceber que continuam existindo linhas de força ligando o par de átomos original. Comportamento similar pode ser observado nas figuras 4.4(d) e (e). Na figura 4.4(f) a interposição de uma carga diretamente entre A_i e A_j impede completamente a interação entre este par de cargas.

Tal raciocínio pode ser extendido para múltiplas cargas, caindo no problema clássico de interação de N-cargas. O problema do cálculo da energia de interação entre duas cargas, sob a influência de outras cargas, parece ser um problema em aberto, visto não ter sido identificado nenhum trabalho relativo a este problema, em especial no contexto das interações desta natureza no interior de macromoléculas. A solução deste problema não é uma tarefa trivial, sendo que uma primeira solução para este problema foi apresentado por Veloso *et al.* [Veloso et al. (2007)].

Devido a importância tecnológica deste problema, principalmente na área de semicondutores, vários estudos têm sido feitos ao longo dos anos, com diferentes graus de sucesso. Potenciais como os de Stillinger-Weber [Stillinger e Weber (1985)] e Tersoff [Tersoff (1988)] são os mais frequentemente citados, notadamente nas áreas de semicondutores e nanotubos de carbono. Alguns trabalhos foram publicados citando a possível aplicação destes potenciais ao caso das proteínas [Barkema e Mousseau (2001), Gujrati (2007)]. Entretanto, não há registros da aplicação destes na especificação energética das interações atômicas e na identificação da malha de interações da estrutura terciária das proteínas. Isto se deve ao fato destes potenciais terem sido concebidos visando arranjos atômicos de silício e arseniato de gálio no estado sólido (o que não é o caso das proteinas), ao mesmo tempo em que eles se

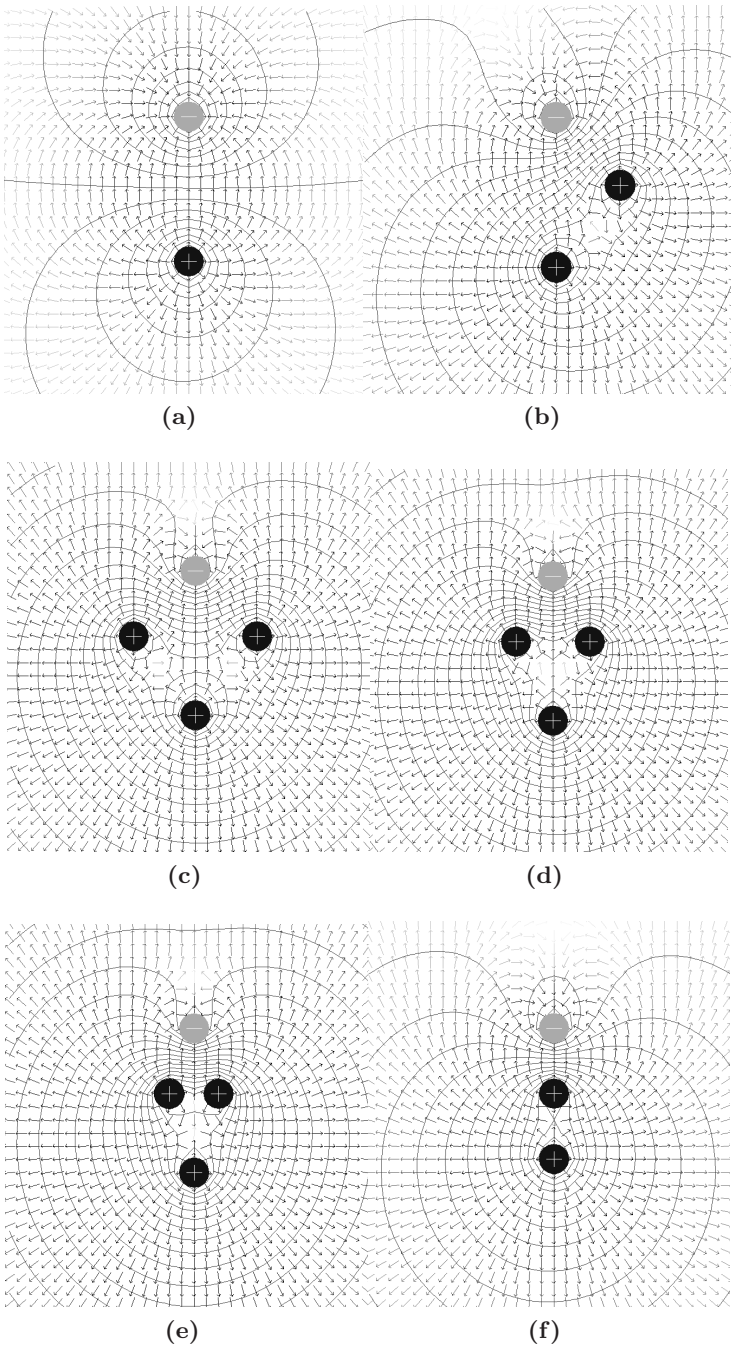


Figura 4.4 *– Linhas de força de duas cargas de sinais contrários em interação próxima.*

prestam para o cálculo dos potenciais elétricos entre cargas, mas não para tratar efeitos de atenuação de uma carga sobre a interação de outras duas.

Para contornar estas limitações, a heurística adotada neste trabalho para lidar com este problema basea-se no seguinte raciocínio. Voltando à figura 4.4(b) é possível observar que praticamente metade das linhas de força relativas à interação entre A_i e A_j mantêm-se inalterada, enquanto a outra metade foi redirecionada para a interação entre A_i e A_k. Desta

forma, como metade do campo elétrico original encontra-se inalterado, é possível provar que a energia de interação entre A_i e A_j, em presença de A_k, é metade da energia de interação original. Ao mesmo tempo, nas figuras 4.4(d) e (e) observa-se que na presença de um número suficientemente de cargas, as linhas de força que fluem de A_i para A_j tendem a se manter, sugerindo ser possivel associar a energia de interação entre ao átomos como sendo proporcional à área "visível" entre os átomos que estão interagindo. Tal premissa torna possível apresentar uma solução para este problema, mesmo não sendo uma solução rigorosa.

Contudo, a aproximação apresentada por esta abordagem permite estimar de forma razoável a energia de interação entre os átomos neste estudo. Ademais, é possível percerber que nestas condições é possível lidar com o problema de blindagem entre as cargas, como sendo proporcional à área mutuamente "visível" entre os átomos.

Em outras palavras, a energia de interação entre os átomos seria proporcional ao efeito de oclusão devida à interveniência existente entre os átomos. Esta heurística permite endereçar este problema dado oportunidade para o desenvolvido de um algoritmo que simplifica o entendimento das imediações dos átomos que estão interagindo ao mesmo tempo que permite desconsiderar as interações não realísticas de longo alcance que não poderiam ocorrer entre dois átomos quando um terceiro (ou mais) átomos ocluem um do outro. O cálculo da oclusão entre os átomos implica na determinação da distância euclidiana e os ângulos entre eles.

Seja $T = \{ A_i,\ A_j,\ A_k \} \mid i \neq j \neq k$ um conjunto de três átomos tal como mostrado na figura 4.5. Seja V_{ij} o vetor ligando (A_i, A_j) e seja V_{ik} o vetor ligando (A_i, A_k). Seja θ o angulo formado por $A_j - A_i - A_k$. Se $\cos(\theta) > 0$ e a perpendicular de A_k até V_{ij} é menor que a soma dos raios de van der Waals de A_k e A_j então A_j é considerada como sendo oclusa de A_i por A_k.

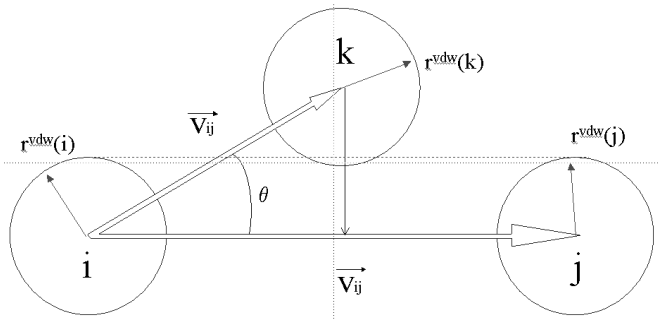


Figura 4.5 – *Esquema mostrando a oclusão entre átomos*

Desta forma torna-se importante mensurar como a energia de interação entre A_i e A_j decresce à medida que o átomo A_j torna-se ofuscado de A_i por A_k. Uma abordagem razoável seria considerar a energia de interação entre os átomos como sendo proporcional a área projetada de A_j "visível" por A_i. Esta abordagem é análoga àquela de um "eclipse lunar", onde a Terra lança a sua sombra sobre a Lua. Geometricamente, este raciocínio pode ser ilustrado tal como apresentado na figura 4.6.

Para melhor explicar este conceito, considere as projeções de dois círculos de raios r_k e

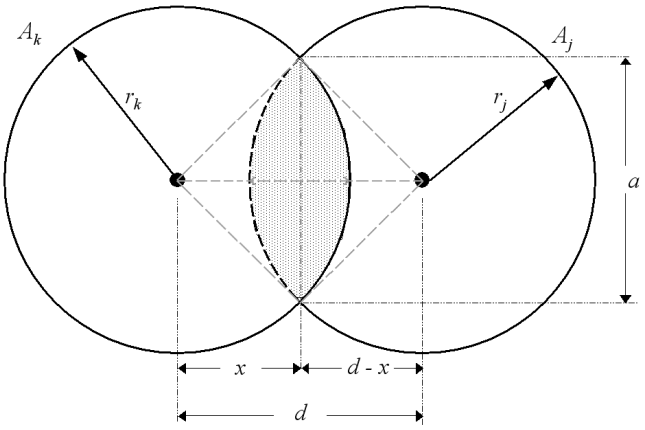


Figura 4.6 – *Esquema mostrando a oclusão do átomo A_j pelo átomo A_k como " visto" por A_i.*

r_j e centrados em $(0,0)$ e $(d,0)$ que se interceptam em uma região na forma de uma lente assimétrica. As equações dos dois círculos são:

$$\begin{aligned} x^2 \ + \ y^2 \ &= \ r_k^2 \\ (x \ - \ d)^2 \ + \ y^2 \ &= \ r_j^2 \end{aligned}$$

De forma a encontrar a área de uma lente assimétrica na qual os círculos se interceptam, pode se adotar uma expressão para o segmento circular de raio $R^{'}$ e altura triangular $d^{'}$ onde:

$$A(R^{'}, d^{'}) \ = \ R^{'2} \ \arccos\left(\frac{d^{'}}{R^{'}}\right) \ - \ d^{'}\sqrt{R^{'2} \ - \ d^{'2}}$$

.

Como existem duas "lentes" na interseção das esferas, a área total da interseção será encontrada realizando se este cálculo duas vezes. Desta forma, a altura dos dois segmentos triangulares será:

$$d_1 \ = \ x \ = \frac{d^2 \ - \ r_j^2 \ + \ r_k^2}{2d}$$

$$d_2 \ = \ d_1 \ - \ x \ = \frac{d^2 \ + \ r_j^2 \ - \ r_k^2}{2d}$$

A área total das "lentes" pode então ser expressa como

$$A \ = \ A(\ r_k, d_1) \ + \ A(\ r_j, d_2)$$

$$= r_j^2 \arccos\left(\frac{d^2 + r_j^2 - r_k^2}{2dr_j}\right) + r_k^2 \arccos\left(\frac{d^2 + r_k^2 - r_j^2}{2dr_k}\right) - \frac{1}{2} \times [(-d + r_j + r_k)(d + r_j - r_k)(d - r_j + r_k)(d + r_j + r_k)]^{\frac{1}{2}} .$$

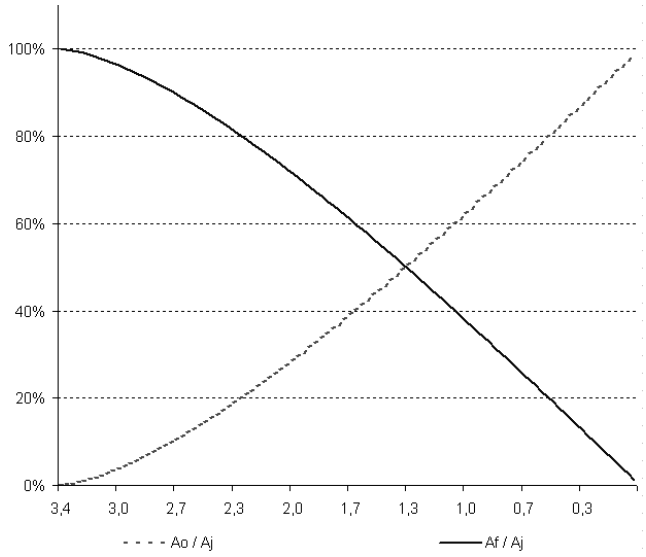


Figura 4.7 – *Gráfico comparando as áreas expostas e oclusas como funções da distância entre os centros dos átomos A_k e A_j. O termo A_o é a área oclusa de A_j. O termo A_f é a área livre de A_j. a curva contínua mostra o percentual do total da área de A_j que está exposta a A_i como função da distância entre os centros de A_k e A_j.*

O caso limite para esta expressão pode ser verificado variando-se $d = 0$ até $d = r_j + r_k$. como pode ser visto na figura 4.7, a área exposta de A_j (A_f), decai quase linearmente à medida que A_k intercepta o espaço entre A_i e A_j. No escopo deste trabalho, esta função será usada como um fator de atenuação para a energia de interação, de forma a reproduzir as interferências causadas pela presença de um terceiro átomo em uma interação entre pares de átomos.

Entretanto, esta abordagem não permite tratar de forma adequada os casos onde uma área do átomo j já computada como sendo ofuscada por um átomo k_1 está sendo também ofuscada por um outro átomo k_2. Esta situação pode ser melhor entendida pela análise do caso apresentado no figura 4.8

No esquema mostrado na figura 4.8, a fração de área da superfície de i ofuscada pelos átomos k_1 e k_2 pode ser calculada pela relação

$$(k_1 \cap k_2) \cap j = (k_1 \cap j) \cup (k_2 \cap j) - ((k_1 \cap j) \cap (k_2 \cap j)).$$

Isto permite perceber que a área do átomo j ofuscada pela presença de outros átomos não pode ser calculada pela simples soma das áreas de sombreamento, mas devem ser também

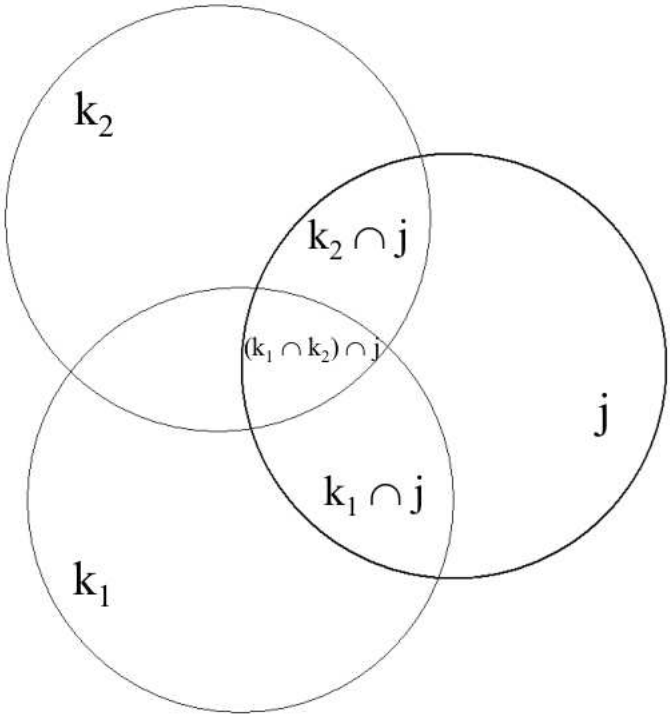


Figura 4.8 *– Situação onde existe a sobreposição das projeções de dois átomos k_1 e k_2 sobre a área visível do átomo j.*

computadas as áreas onde ocorrem as sobreposições de sombreamentos.

Para lidar com este problema, desenvolvemos uma solução onde a área do átomo j é mapeada sobre uma matriz e as áreas sombreadas são marcadas também nesta matriz. A figura 4.9 apresenta uma visão esquemática desta solução.

Com esta solução, caso haja superposição das áreas oclusas, a marcação ocorrerá sobre uma célula já marcada, o que passa a ser inócuo. Com as áreas oclusas mapeadas sobre esta matriz o cálculo da área visível reduz-se a um exercício de contar o número de células marcadas como livres e o número de células marcadas como oclusas.

Contudo, para que este processo de marcação das áreas oclusas possa ser feito de forma confiável, translações e rotações devem ser feitas sobre cada uma das várias tríades A_i,A_j e A_k para que todas elas possam ser analisadas do mesmo ponto de vista. Este procedimento ocorre observando as seguintes etapas:

- Translação do sistema A_i,A_j e A_k tal que o átomo A_j esteja no centro do sistema de referência ou seja $A_j = (0, 0, 0)$;
- Rotação do sistema A_i,A_j e A_k em torno do eixo Z tal que $A_i = (x_i, 0, z_i)$;
- Rotação do sistema A_i,A_j e A_k em torno do eixo Y tal que $A_i = (x_i, 0, 0)$;

Com esta série de transformações o átomo A_i é posto na posição do observador, que passa a ver o posicionamento relativo de A_k e A_j, ficando fácil identificar a área de sombreamento.

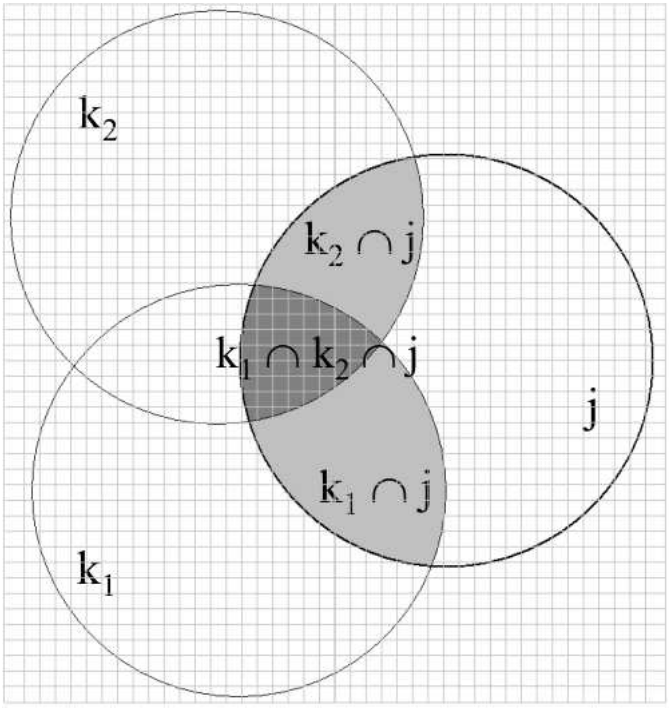


Figura 4.9 – *Projeções de dois átomos k_1 e k_2 sobre átomo j mapeadas sobre uma matriz.*

As seções seguintes explicam estas etapas de cálculo de forma mais detalhada.

Translação do sistema A_i, A_j e A_k : Sejam os átomos A_i, A_j e A_k definidos pelas suas coordenadas

$$A_i = (x_i, y_i, z_i)\ A_j = (x_j, y_j, z_j)\ A_k = (x_k, y_k, z_k)$$

Para que o átomo A_j esteja no centro do sistema de referência, ou seja $A_j = (0, 0, 0)$, é necessário que este sistema de referência se desloque sobre o vetor $\overrightarrow{d} = (-x_j, -y_j, -z_j)$ de forma que a operação $A_j + \overrightarrow{d} = (0, 0, 0)$. De forma similar, o mesmo deslocamento é aplicado para os átomos A_i e A_k.

Com este deslocamento as novas coordenadas da tríade A_i,A_j e A_k serão

$$\begin{aligned} A_i &= (x_i - x_j, y_i - y_j, z_i - z_j) \\ A_j &= (x_j - x_j, y_j - y_j, z_j - z_j) \\ A_k &= (x_k - x_j, y_k - y_j, z_k - z_j) \end{aligned}$$

ou seja

$$\begin{aligned} A_i &= (\quad x_i - x_j \quad y_i - y_j \quad z_i - z_j \quad) \\ A_j &= (\quad 0 \quad 0 \quad 0 \quad) \\ A_k &= (\quad x_k - x_j \quad y_k - y_j \quad z_k - z_j \quad) \end{aligned}$$

Rotação do sistema A_i, A_j e A_k em torno do eixo Z tal que $A_i = (x_i, 0, z_i)$: Para um dado conjunto de pontos, as novas coordenadas resultantes da rotação deste mesmo conjunto em torno do eixo Z podem ser calculadas pela equação matricial

$$A_1 = A_0 \cdot R_z(\gamma)$$

onde $R_z(\gamma)$ é

$$\begin{bmatrix} cos\gamma & -sen\gamma & 0 \\ sen\gamma & cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Assim, os novos valores das coordenadas dos pontos A_i, A_j e A_k são obtidos pelas seguintes relações:

$$\begin{aligned} x_2(i) &= x_1(i)cos\gamma - y_1(i)sen\gamma \\ y_2(i) &= x_1(i)sen\gamma + y_1(i)cos\gamma \\ z_2(i) &= z_1(i). \end{aligned}$$

Como $y_2(i) = 0$,tem-se:

$$\begin{aligned} y_2(i) &= x_1(i)\ sen\gamma + y_1(i)\ cos\gamma = 0 \\ x_1(i)\ sen\gamma &= -y_1(i)\ cos\gamma \\ \frac{sen\gamma}{cos\gamma} &= -\frac{y_1(i)}{x_1(i)} \\ tan\gamma &= -\frac{y_1(i)}{x_1(i)} \\ \gamma &= arctan(-\frac{y_1(i)}{x_1(i)}) \end{aligned}$$

Rotação do sistema A_i,A_j e A_k em torno do eixo Y tal que $z(i) = 0$ ou $A_i = (x_i, 0, 0)$: Para um dado conjunto de pontos, as novas coordenadas resultantes da rotação deste mesmo conjunto em torno do eixo Z podem ser calculadas pela equação matricial

$$A_1 = A_0 \cdot R_y(\beta)$$

onde $R_y(\beta)$ é

$$\begin{bmatrix} cos\beta & 0 & sen\beta \\ 0 & 1 & 0 \\ -sen\beta & 0 & cos\beta \end{bmatrix}$$

Assim, os novos valores das coordenadas dos pontos A_i, A_j e A_k são obtidos pelas seguintes relações:

$$x_3(i) = x_2(i)cos\beta + z_2(i)sen\beta$$
$$y_3(i) = y_2(i)$$
$$z_3(i) = -x_2(i)sen\beta + z_2(i)cos\beta.$$

Como $z_3(i) = 0$,tem-se:

$$z_3(i) = -x_2(i)\ sen\beta + z_2(i)\ cos\beta = 0$$
$$x_2(i)\ sen\beta = z_2(i)\ cos\beta$$
$$\frac{sen\beta}{cos\beta} = \frac{z_2(i)}{x_2(i)}$$
$$tan\beta = \frac{z_2(i)}{x_2(i)}$$
$$\beta = arctan(\frac{z_2(i)}{x_2(i)})$$

4.4.3.1 Identificação das áreas de sombreamento

Uma vez que o sistema formado pelos pontos A_i, A_j e A_k esteja posicionado adequadamente para análise, é possível identificar três situações quanto à posição relativa entre os átomos A_j e A_k, tal como apresentadas na figura 4.10:

- A_j e A_k são concêntricos;
- A_j e A_k estão sobrepostos, mas não são concêntricos;
- A_j e A_k estão parcialmente sobrepostos.

No primeiro caso, é trivial perceber que $d(A_j, A_k) = 0$. Nesta situação, dada à pouca diferença entre os diâmetros dos diferentes átomos existentes em uma proteína (1,3Å $\leq$ r $\leq$ 1,8Å), assume-se que o átomo A_j estará totalmente ofuscado.

No segundo caso, tem-se $0 \leq d(A_j, A_k) \leq (r(A_k) - r(A_j))$. Entretanto, os raios dos átomos tratados no escopo deste trabalho variam dentro de uma faixa estreita. Ao mesmo tempo, considera-se que os átomos de hidrogênio (r = 1,3Å) não são considerados como capazes de gerar oclusão (devido à sua baixa densidade eletrônica). Com isto, a faixa dentro da qual os raios dos átomos variam dentro do escopo deste trabalho fica limitada entre 1,3Å $\leq$ r $\leq$ 1,8Å. Com isto, o valor máximo de distância entre os centros dos átomos é de 0,3Å ou seja 0,17% r_{max}. Nestas condições, a área visível de A_j é desprezível sendo, para fins práticos, considerada como sendo nula.

O terceiro caso, é o mais complexo onde aos átomos A_j e A_k apresentam sobreposição parcial onde

$$|r_{A_k} - r_{A_j}| < d(A_j, A_k) < |r_{A_k} + r_{A_j}|$$

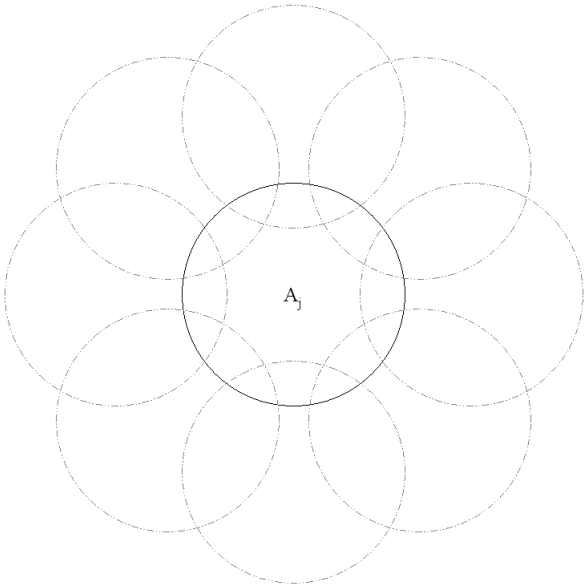


Figura 4.10 *– Situações possíveis quanto à posição relativa entre os átomos A_j e A_k.*

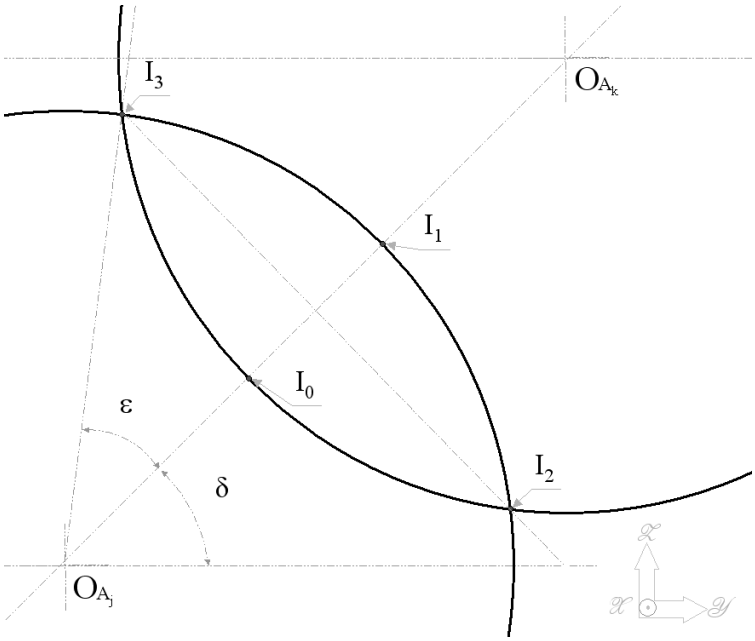


Figura 4.11 *– Situação de sobreposição parcial entre os átomos A_j e A_k.*

Na figura 4.11 a lente formada pela sobreposição dos átomos A_j e A_k é identificada por quatro pontos: I_0, I_1, I_2, I_3. A identificação das coordenadas do ponto I_0 é feita conforme o raciocínio que se segue. Seja $\overline{jk}$ o segmento de reta que une os pontos O_{A_j} e O_{A_k}. A inclinação do segmento $\overline{jk}$ é obtida pela relação

$$\tan\delta = \frac{z(k) - z(j)}{y(k) - y(j)}.$$

Considerando que o ponto O_{A_j} está no centro do sistema de referência, a relação precendente é reduzida para

$$\tan\delta = \frac{z(k)}{y(k)}$$

ou seja

$$\delta = \arctan\left(\frac{z(k)}{y(k)}\right).$$

Seja $r(k)$ o raio do átomo A_k, e $d(O_{A_k}, I_0)$ a distância de O_{A_k} à I_0. Da figura 4.11 tem-se

$$\begin{aligned} r(k) &= d(O_{A_k}, I_0) \quad \therefore \\ -I_0 &= r(k) - O_{A_k} \quad \therefore \\ I_0 &= O_{A_k} - r(k). \end{aligned}$$

Decompondo as componentes de I_0, tem-se

$$\begin{cases} y(I_0) = y(O_{A_k}) - r(k)cos\delta \\ z(I_0) = z(O_{A_k}) - r(k)sen\delta \end{cases}$$

A dedução das relações que identificam as coordenadas de I_1 é trivial. Os valores das coordenadas de I_1 são dados pelas relações:

$$\begin{cases} y(I_1) = r(j)cos\delta \\ z(I_1) = r(j)sen\delta \end{cases}$$

As coordenadas de I_3 são dadas pelas relações:

$$\begin{cases} y(I_3) = r(j)cos(\delta + \epsilon) \\ z(I_3) = r(j)sen(\delta + \epsilon) \end{cases}$$

onde

$$\epsilon = \left[\frac{\left[\frac{I_1 - I_0}{2} + I_0\right]}{r(j)}\right]$$

Como o ponto I_2 é simétrico a I_3 em relação ao segmento $\overline{jk}$, suas coordenadas são dadas pelas relações:

$$\begin{cases} y(I_2) = r(j)cos(\delta - \epsilon) \\ z(I_2) = r(j)sen(\delta - \epsilon) \end{cases}$$

Identificados estes pontos, os pontos de demarcam a área retangular que contêm a lente que limita a área ofuscada (figura 4.12), são calculados pelas relações:

$$y_{up} = max(y(I_0), y(I_1), y(I_2), y(I_3))$$
$$y_{lw} = min(y(I_0), y(I_1), y(I_2), y(I_3))$$

$$z_{up} = max(z(I_0), z(I_1), z(I_2), z(I_3))$$
$$z_{lw} = min(z(I_0), z(I_1), z(I_2), z(I_3))$$

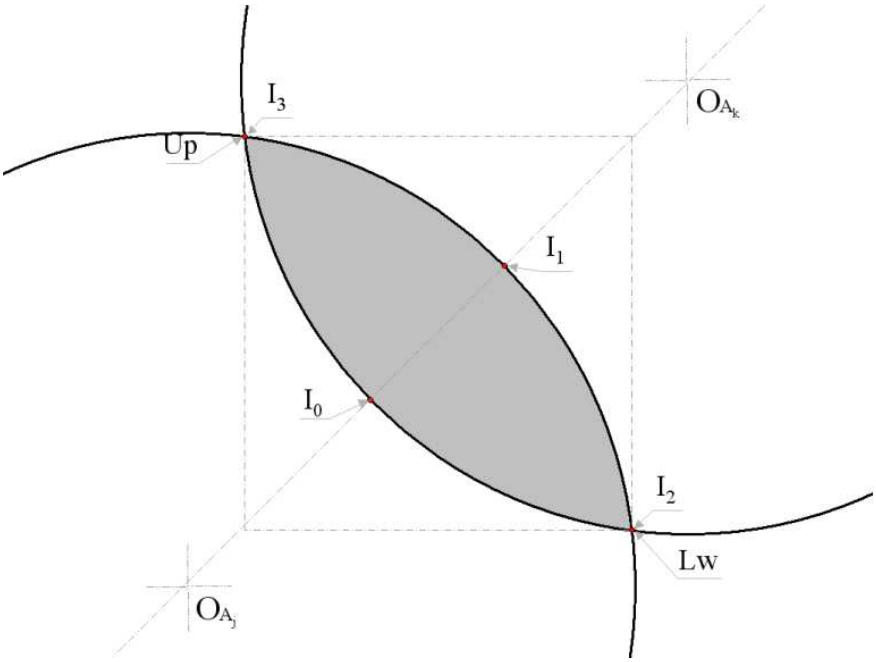


Figura 4.12 – *Área retangular que contém a área de A_j ofuscada por A_k.*

O algoritmo 2 mostra o procedimento pelo qual a área sobre A_j ofuscada por A_k é mapeada sobre a matriz. Com este artifício, garante-se que os pontos sobre a área visível de A_j que já tenham sido previamente marcados como oclusos (ou ofuscados) previamente, não sejam computados mais de uma vez dando assim uma dimensão mais próxima da realidade da natureza das interações entre os átomos no interior das proteínas.

Algoritmo 2: Mapeamento de pontos oclusos

Entrada: Matriz de Pontos,
Raios dos átomos A_j e A_k,
Pontos limítrofes da lente de sobreposição

Saída: Matriz de Pontos,
Fração de área livre

início
  $y_{up} = max(y(I_0), y(I_1), y(I_2), y(I_3))$
  $y_{lw} = min(y(I_0), y(I_1), y(I_2), y(I_3))$
  $z_{up} = max(z(I_0), z(I_1), z(I_2), z(I_3))$
  $z_{lw} = min(z(I_0), z(I_1), z(I_2), z(I_3))$

  para $y = y_{lw}$ **até** y_{up} **faça**
    para $z = z_{lw}$ **até** z_{up} **faça**
      se $(y^2 + z^2 \leq r_j^2) \wedge ((y - y_k)^2 + (z - z_k)^2 \leq r_k^2\ )$ **então**
        Matriz [y , z] = O
fim

4.4.3.2 Remoção de interações sem significância física

Visando trabalhar de uma forma mais realística, considera-se que todos os valores de energia de interação não covalente que estejam abaixo de um limiar serão descartados.

Este limiar é definido como

$$E = k_b T$$

onde:

k_b	$\rightarrow$ constante de Boltzmann	$= 1{,}3806505 \times 10^{-23}$	$[joule/kelvin]$
T	$\rightarrow$ temperatura do sistema	$= 298$	$[kelvin]$.

Fazendo a análise dimensional e as devidas conversões, tem-se

$$E = 5,92 \times 10^{-1} \left[\frac{kcal}{mol}\right] \ \therefore\ E = 0,6 \left[\frac{kcal}{mol}\right]$$

Assim, todas as interações com valores abaixo deste limiar serão consideradas como incapazes de se manterem estáveis.

4.4.4 Análise das RGPSs

As redes foram analisadas observando vários aspectos descritivos relativos às redes complexas conforme descrito na literatura. Para cada RGPS, o número de vértices N(k) com k arestas (links) foi avaliado. O valor N(k) para todas as proteínas presentes no conjunto de dados foi calculado e a relação N(k) versus k foi plotada.

De forma com algumas propriedades das RGPS elas foram representadas como matriz de adjacências A^E, onde:

- $A^E_{ij} = E_{ij}$,
  se $i \neq j$ e i e j não estão oclusas uma da outra;
- $A^E_{ij} = 0$,
  se $i \neq j$ e i e j estão oclusas uma da outra;
- $A^E_{ij} = 0$, se $i = j$

A matriz de adjacência A^E é então analisada para identificar aglomerados distintos e nodos (átomos) formadores de aglomerados na RGPS. O maior aglomerado é então identificado e seu tamanho é determinado para toda a RGPS. O valor normalizado do tamanho do maior aglomerado (em relação ao número total de resíduos na proteína) é plotado para todas as proteínas do conjunto estudado. Especificamente, os átomos com maior número de contatos($NAP \geq 10$), e os resíduos onde estes encontram-se localizados, são identificados como sendo "*hubs*" nas estruturas das proteínas.

4.4.5 Perfil de distribuição das arestas dos átomos e resíduos de aminoácidos

Para um dado átomo A_i, o número total de contatos estabelecidos por ele, pode ser definido como:

$$N_{CA}(i) = \sum_{j=1}^{NAP} (A_i, A_j) \;\; | \;\; \{A_i, A_j\} \in P \wedge i \neq j. \tag{4.5}$$

Para um dado resíduo R_m, o número do total de contatos para este resíduo no $N_{CR}(m)$ é calculado como:

$$N_{CR}(m) = \sum_i \sum_j (A_i, A_j) \tag{4.6}$$

onde P é a proteína estudada, e:

$$i \neq j \wedge A_i \in R_m \wedge A_j \ni R_m \wedge \{A_i, A_j\} \in P.$$

Estes cálculos são feitos para todas as proteínas do conjunto em estudo. Os valores são obtidos utilizando todas as proteínas e a distribuição de frequência é plotada.

4.5 Análise do Ganho de Informação

Lockless e Ranganathan [Lockless e Ranganathan (1999)] apresentam uma forma de avaliar os padrões de interações entre resíduos que se mantém conservados durante o processo de evolução. Segundo estes, a evolução do enovelamento de uma proteína é o resultado de uma mutagênese aleatória de larga escala, onde a necessidade da manutenção da função da proteína aparece como o principal fator de restrição seletiva destas mutações. O argumento fundamental destes autores é baseado em duas hipóteses que derivam de observações empíricas sobre a evolução das seqüências de resíduos. Desta forma, a falta de restrições evolucionárias em uma posição da estrutura primária poderia causar a distribuição dos aminoácidos observados naquela posição da seqüência, quando em um alinhamento estrutural múltiplo. A observação desta distribuição de freqüências em todas as proteínas de interesse, poderia representar quantitativamente a conservação de um dado resíduo naquela posição da seqüência. Da mesma forma, o acoplamento funcional de duas posições, mesmo distantes na seqüência primária ou terciária, poderia agir como uma restrição da metagênese evolucionaria, o que poderia ser percebido pelo acoplamento estatístico da distribuição de freqüências dos aminoácidos.

Assim, Lockless e Ranganathan [Lockless e Ranganathan (1999)] conduzem o desenvolvimento do método com base em duas premissas:

1. Que a conservação de um resíduo em uma dada posição em um alinhamento estrutural múltiplo é definida como o desvio global das freqüências dos resíduos de aminoácidos naquela posição, com relação ao seus valores médios;

2. O acoplamento estatístico de duas posições, i e j, é definido como o grau no qual a freqüência de cada resíduo de aminoácido, na posição i muda em resposta à uma perturbação de freqüência em uma outra posição j. Esta definição de acoplamento não requer que a conservação geral na posição i seja alterada a cada perturbação em j, mas somente que a população de resíduos seja rearranjada.

Desta forma, um vetor de 20 distribuições binomiais de probabilidades de ocorrências, relativas à cada um dos 20 aminoácidos, seria necessário para identificar estes padrões de conservação. Esta distribuição representaria a distribuição de ocorrências dos aminoácidos independente da sua posição na seqüência.

Na mecânica estatística clássica, estando um sistema em equilíbrio, a temperatura T que o mesmo apresenta é proporcional à velocidade média de transição de estados das suas partículas. Nesta condição define-se a unidade fundamental de energia k_bT, onde k_b é a constante de Boltzmann.

Para Lockless e Ranganathan, cada uma das várias posições de uma seqüência de resíduos, em um alinhamento estrutural múltiplo- AEM, poderiam ser vistas como sistemas mecânicos estatísticos isolados que podem variar entre discretos estados possíveis em um espaço de estados representado pelas freqüências de ocorrência dos diferentes resíduos naquela posição. Desta forma, estes autores definem uma "temperatura" T^* a qual estaria relacionada à média dos estados estatisticamente possíveis de serem apresentados por cada uma destas diferentes

posições presentes no AEM. Porém, neste conceito a unidade de energia kT^* não é necessariamente relacionada aos sistemas mecânicos clássicos.

Assim, para um conjunto de proteínas evolutivamente relacionadas e amostradas em um AEM, onde as freqüências de ocorrência dos resíduos em cada uma das posições são mutuamente independentes, a probabilidade de um resíduo x em uma posição i relativa a uma outra posição j, está relacionada ao ganho de informação[1], relacionado às posições i e j para um certo resíduo x ($\Delta G^x_{i\to j}$), dado por uma distribuição similar à de Boltzmann:

$$\frac{p^x_i}{p^x_j} = e^{\frac{\Delta G^x_{i\to j}}{kT^*}} \tag{4.7}$$

onde $kT^* = 1$ arbitrariamente definida pelos autores.

A probabilidade de ocorrência de um resíduo x na posição i (P^x_i) é dada pela função de densidade binomial

$$P(x) = \frac{N!}{n_x!(N-n_x)!} p_x^{nx}(1-p_x)^{N-n_x}$$

onde N é o número total de seqüências em estudo, n_x é o número de seqüências com o resíduo x, e p_x é a freqüência média do resíduo x em todas as proteínas em estudo.

Desta forma, fica definido o parâmetro de conservação empírico (ΔG^{stat}) para a posição i

$$\Delta G^{stat}_i = kT^* \sqrt{\sum_i \left(\ln \frac{p^x_i}{p^x_{AEM}} \right)^2} \tag{4.8}$$

o qual informa o ganho (ou conservação) de informação representado pela ocorrência do resíduo x na posição i.

Segundo Lockless e Ranganathan, a medida do acoplamento funcional de duas posições i e j, é calculada para uma dada posição i considerando duas condições

1. O valor de ΔG^{Stat}_i;
2. Um subconjunto selecionado a partir do AEM, para o qual um resíduo qualquer y é mantido constante na posição j. Para este conjunto calcula-se $\Delta G^{Stat}_{i|(y,j)}$

A magnitude da diferença entre estes dois parâmetros - ΔG^{Stat}_i e $\Delta G^{Stat}_{i|(y,j)}$ mostra o grau de acoplamento estatístico ($\Delta\Delta G^{Stat}_{i,j}$) entre as posições i e j.

Os resultados da aplicação destes métodos são apresentados na página 146, onde as propriedades dos diferentes resíduos presentes nas globinas serão analisadas.

[1] Lockless e Ranganathan dão a este termo o nome de "*statistical free energy*", mas neste trabalho será usado o termo "ganho de informação".

Capítulo 5

Resultados e Discussão

Objetivando a avaliar a propriedade das hipóteses apresentadas neste trabalho, estudos focados em atributos estruturais de um conjunto selecionado de proteínas foram conduzidos conforme os métodos explicados previamente no capítulo 4. Os resultados obtidos destas análises são apresentados e discutidos neste capítulo.

5.1 Comparação entre os resultados obtidos com e sem o uso do método de oclusão entre átomos

Nesta primeira seção são apresentados e discutidos os resultados relativos ao processo de descobrimento das interações não-covalentes que surgem do processo de enovelamento das proteínas.

No gráfico apresentado na figura 5.1 são mostradas duas seqüências de dados. A seqüência (**b**) mostra a distribuição de freqüências média para as distâncias entre átomos sem a solvatação das proteínas e sem a aplicação dos critérios de oclusão. Percebe-se que o número médio de ligações entre os átomos, nesta condição, tende a crescer à medida que o raio do espaço de pesquisa (a partir de um dado átomo) cresce. Apesar de parecer óbvia, tal constatação mostra a existência de uma grande quantidade de interações plausíveis cuja distância entre os átomos varia de 1,00 Å até 10,00 Å. Tal constatação mostra que a simples adoção de uma distância limítrofe arbitrária pode estar deixando de contemplar interações relevantes, cuja distância esteja além deste limite arbitrado.

Ainda na figura 5.1, a seqüência (**a**) mostra a distribuição de freqüências média para as distâncias entre átomos agora com a solvatação das proteínas e com a aplicação dos critérios de oclusão. Neste caso observa-se que o perfil da distribuição do número médio de ligações entre os átomos, não apresenta os picos de freqüência observados na faixa de 1Å a 1,25Å observados na seqüência (**b**).

As discrepâncias encontradas nas condições da seqüência (**b**), estão associadas aos efeitos de "empacotamento geométrico" dos átomos no interior das proteínas, fenômeno estudado na física do estado sólido [Silbert et al. (2002), Aste et al. (2004), Aste et al. (2006), Lochmann et al. (2006), Aste e Senden (2007)], mas que pouco interfere nas interações "não locais"

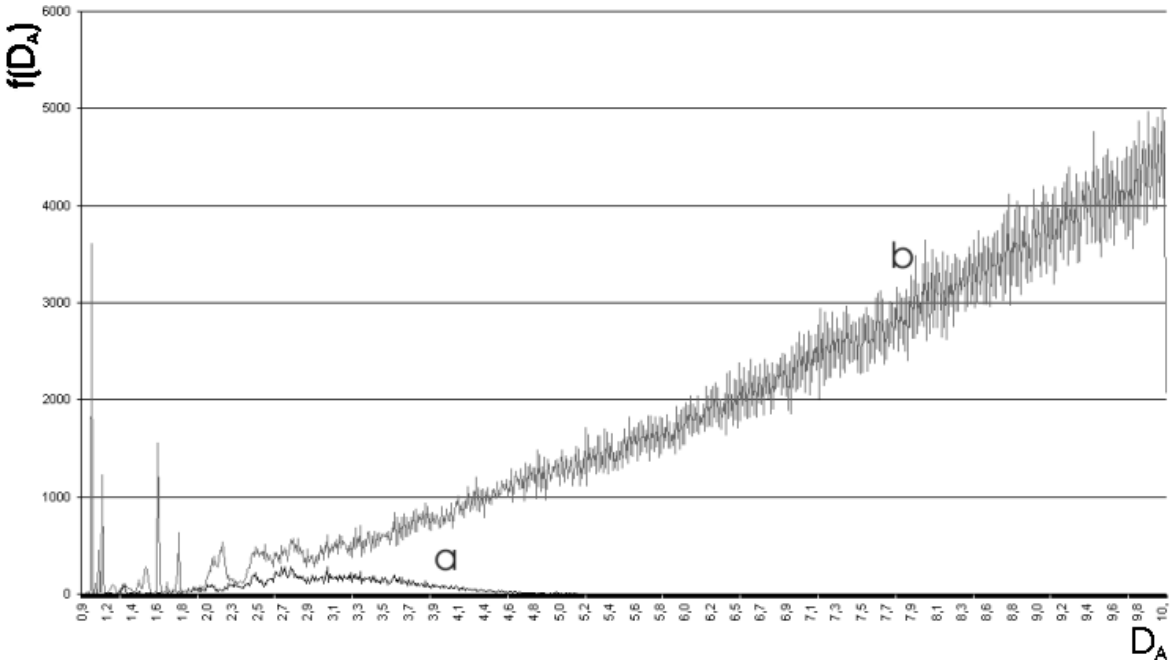


Figura 5.1 *– Distribuição média das freqüências das distâncias entre átomos ($f(D_A)$) para ligações oclusas e não oclusas para globinas e serinoproteases. A seqüência (**a**) mostra a distribuição de freqüências média para as distâncias entre átomos das proteínas na condição de solvatação e com a adoção dos critérios de oclusão. A seqüência (**b**) mostra a distribuição de freqüências média para as distâncias entre átomos das proteínas sem solvatação e sem critério de oclusão.*

estudadas neste trabalho. Tal ponto é melhor explicado na seção 5.2.1. Ao mesmo tempo, existe a contribuição das interações de curta distância entre átomos que estão localizados na superfície das proteínas. Estas interações não são observáveis nas condições da seqüência (**a**) já que nestas condições as proteínas estão solvatadas. A presença das moléculas de água ao redor das proteínas impede a interação entre os átomos da superfície, que não devem existir no caso real. Desta forma a solvatação das proteínas evita a emergência de interações irreais que aparecem como artefatos derivados de uma modelagem errônea dos sistemas moleculares em estudo.

Observa-se, ainda, o completo descolamento das seqüências (**a**) e (**b**) a partir do valor de 2,7Å. Tal descolamento mostra o efeito da aplicação dos critérios de oclusão (que mimetizam melhor os fenômenos reais), onde a eleição de interações atômicas são submetidas aos critérios já definidos para serem consideradas como efetivas ou não. À medida que os valores de distância entre os átomos cresce, as freqüências observadas nas seqüências (**a**) e (**b**) tornam-se cada vez mais discrepantes, deixando claro que enquanto o número de interações consideradas efetivas pelos critérios de oclusão tende a diminuir, o número de interações consideradas efetivas na ausência destes critérios tende a aumentar.

Na figura 5.2, as seqüências (**a**) e (**b**) foram suavizadas com o intuito de eliminar os valores abruptos e permitir uma melhor comparação de ambas seqüências. Nesta escala e após este tratamento, observa-se que ambas seqüências apresentam boa adesão na faixa de

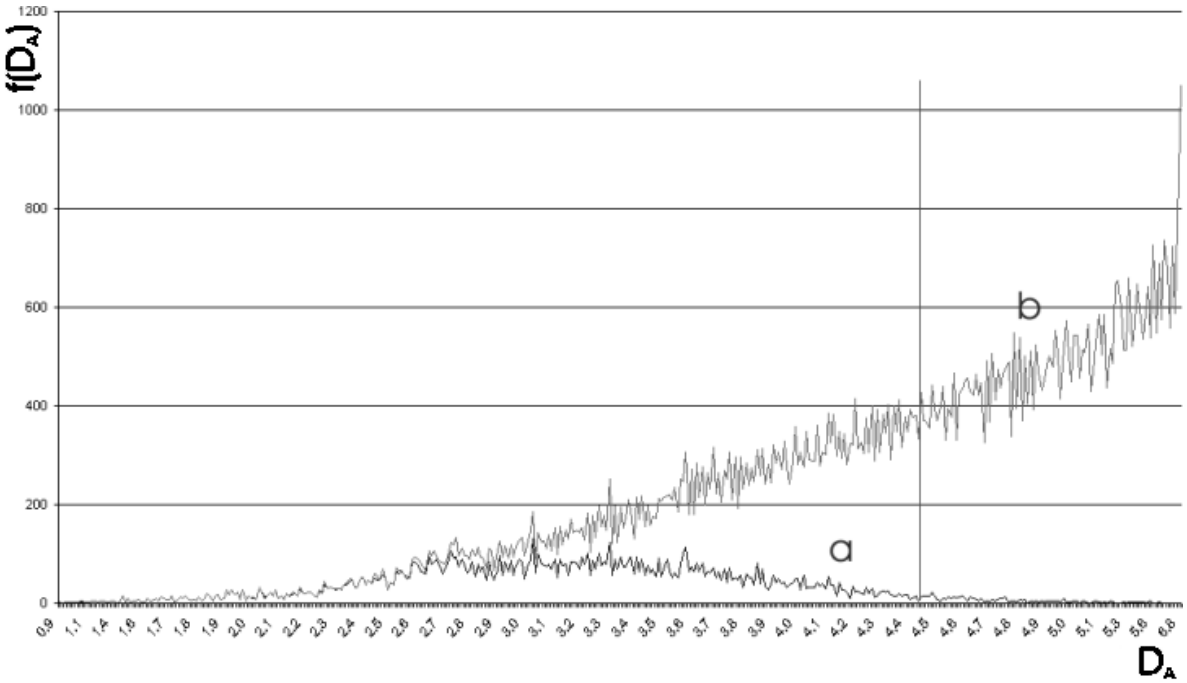


Figura 5.2 *– Distribuição média das freqüências das distâncias entre átomos ($f(D_A)$), nas globinas e serinoproteases, para ligações oclusas (**a**) e não oclusas (**b**).*

1,0Å até 2,6Å. O descolamento a partir do valor de 2,7Å ainda é observado. Novamente, a partir do valor de 4,5Å (que é adotado em vários trabalhos [Greene e Higman (2003), Amitai et al. (2004), Atilgan et al. (2004), Rao e Caflisch (2004), Bagler e Sinha (2005), Brinda e Vishveshwara (2005), Kundu (2005), del Sol e O'Meara (2005), Kundu (2005), Aftabuddin e Kundu (2006), Alves e Martinez (2006), Higman e Greene (2006), del Sol et al. (2006b), del Sol et al. (2006a), Atilgan et al. (2007), Jiao et al. (2007)] como valor limite para distâncias entre átomos), observa-se a existência de interações passíveis de análise tanto para os dados observados com a adoção de critérios arbitrários, quanto com a aplicação dos critérios de oclusão.

Tal constatação mostra que os critérios de oclusão entre átomos, dentro de uma proteína, agem como um filtro natural para a eleição de interações com chances de serem efetivas, muito melhor do que a adoção arbitrária de qualquer valor limítrofe para seleção de interações efetivas ou não.

Ressalta-se que o método de critérios de oclusão apóia se em critérios físicos bem definidos. Com base somente nestes critérios é que a existência das interações atômicas é estimada, não existindo aqui nenhum parâmetro determinado de forma arbitrária por quem realiza a análise.

A figura 5.3 mostra novamente distribuição de freqüências média para as distâncias entre átomos com solvatação das proteínas e com a aplicação dos critérios de oclusão (seqüência (**a**)). Esta observação mais detalhada mostra que esta distribuição de freqüências assemelha-se a uma distribuição normal. Tal constatação leva a crer, inicialmente, que o padrão da rede formada pelas interações não-covalentes no interior de uma proteína seria o mesmo de um

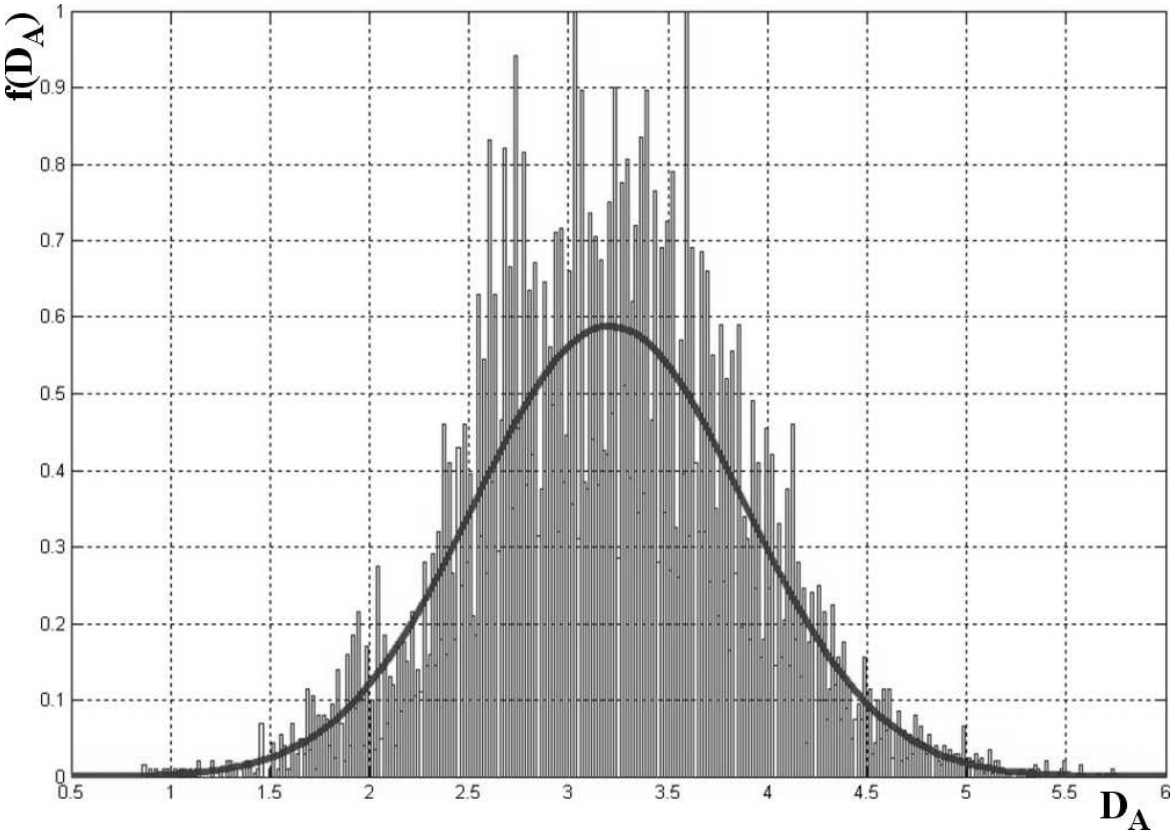


Figura 5.3 *– Distribuição média das freqüências normalizadas, das distâncias entre átomos ($f(D_A)$) para ligações oclusas para globinas e serinoproteases. Curva de tendência Normal foi criada com o uso do MATLAB.*

grafo randômico nos moldes do modelo de Erdös e Rényi [Erdös e Rényi (1959)].

5.2 Análise Estatística das Estruturas das Proteínas

Nesta seção são apresentados e discutidos os resultados das análises estatísticas descritas na seção 3.3, para as proteínas do grupo das globinas e das serinoproteases . Estes resultados permitem identificar vários aspectos estruturais relativos às proteínas estudadas, interpretando-os à luz dos conceitos apresentados no capítulo 3. Tais resultados irão permitir identificar em que medida as estruturas das proteínas aderem aos modelos teóricos apresentados na literatura corrente.

5.2.1 Perfis de Distribuição das Interações Atômicas

Na análise estatística da rede de interações atômicas não-covalentes das proteínas estudadas, estas interações foram inicialmente estudadas sob a condição onde as proteínas não estavam solvatadas e utilizando uma distância limite para a vizinhança de busca de 10 Å. A adoção desta distância limite é feita com o intuito de evitar a demanda de esforço computacional além do necessário para este estudo. Ao mesmo tempo, estipular este limite para a vizinhança de busca acaba não interferindo no cálculo das energias, já que o uso dos critérios de oclusão mostrou não existir interações plausíveis com comprimento superior a 8Å(figura 5.1).

Para essa situação o número de interações por átomo (N_{CA}) e a freqüência dessas interações ($f(N_{CA})$) foram plotados, como mostrado nas figuras 5.4 e 5.5, para todas as proteínas em estudo tanto para as globinas quanto para as serinoproteases respectivamente.

Nestas figuras, todas as interações não-covalentes entre os átomos das proteínas, identificadas são utilizadas no cômputo das distribuições estatísticas apresentadas. Cada proteína estudada é representada, nestas figuras, por um conjunto de símbolos e cores.

Como pode ser visto nas figuras 5.4 e 5.5 os resultados revelam que os padrões das distribuições dos valores das ligações de $f(N_{CA})$ sugerem distribuições de Poisson, similares aos resultados preliminares apresentados na literatura [Greene e Higman (2003), Atilgan et al. (2007)].

Uma rede com esta distribuição de contatos apresenta uma topologia randômica conforme apresentado na seção 3.1.2. Tal topologia não permite inferir nenhuma hierarquia entre os nodos. Indo mais além, os padrões apresentados pelas distribuições mostradas nas figuras 5.4 e 5.5 (as mesmas adotadas em todos os estudos similares a este [Brinda et al. (2002), Vendruscolo et al. (2002), Greene e Higman (2003), Atilgan et al. (2004), Higman e Greene (2006), Atilgan et al. (2007), Ghosh et al. (2007)]), seriam recorrentemente obtidos quaisquer que fossem os valores de distâncias das interações (1,0Å $\leq$ d $\leq$ 10,0Å), mantida a desconsideração das interferências estéricas entre os átomos.

Se de fato a distribuição de interações não-covalentes no interior das proteínas fosse esta, isto equivaleria a dizer que as redes de ligações não-covalentes que estabilizam as proteínas seriam insensíveis à natureza dos resíduos que estabelecem estas interações. Ao mesmo tempo, deve-se ter em mente que as ligações não-covalentes atribuem estabilidade e especificidade à estrutura terciária das proteínas. Novamente, as distribuições das figuras 5.4 e 5.5 estariam

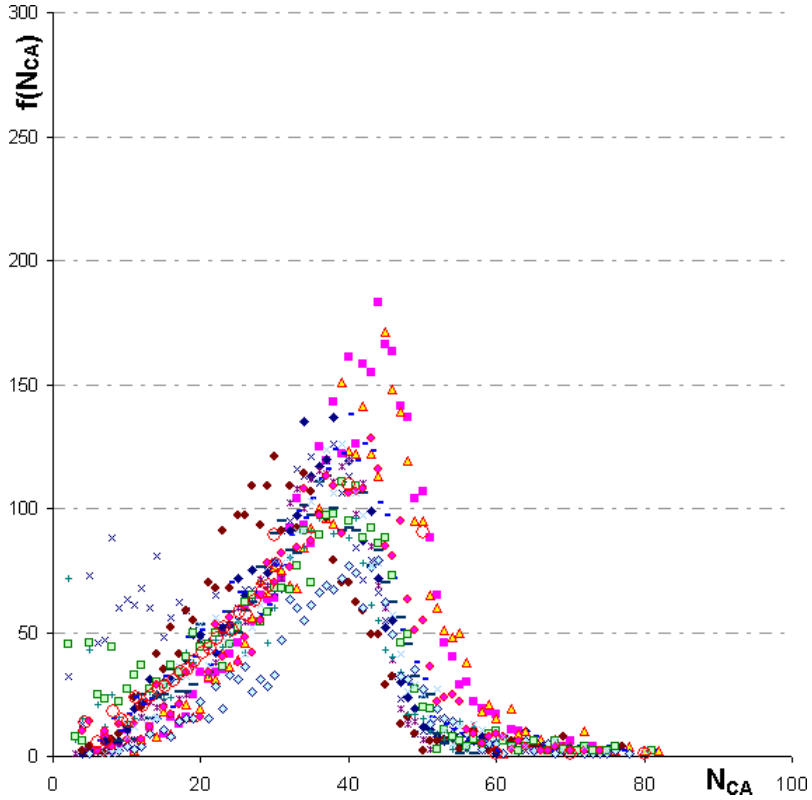


Figura 5.4 *– Distribuição, para as globinas, das freqüências dos números de contatos por átomo ($f(N_{CA})$) e o número de contatos por átomo (N_{CA}) para proteínas não solvatadas sem critério de oclusão. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes.*

informando que as topologias similares apresentadas pelas proteínas de uma mesma família seriam definidas por um arranjo de ligações que, em conjunto, mostra-se inespecífico para qualquer família topológica estudada.

Todavia, tal raciocínio não encontra respaldo na realidade observável. Ao contrário, uma constatação intrigante no estudo das proteínas é que proteínas com alto grau de dissimilaridade no nível das estruturas primárias, podem ser topologicamente bem similares. Ao mesmo tempo, a proteína enovelada apresenta uma hierarquia estrutural, com a formação de padrões secundários e terciários.

Esta constatação permite inferir a existência de uma rede de interações comum à todas as proteínas de uma mesma família. Não obstante a existência de estudos propedêuticos realizados neste sentido [Vendruscolo et al. (2001), Vendruscolo et al. (2002), Greene e Higman (2003), Bastolla et al. (2005), Krishnadev et al. (2005), Higman e Greene (2006), Atilgan et al. (2007)], ainda não foi possível definir tal padrão característico para qualquer família de proteínas. Contudo, tal padrão subjacente deve existir e, portanto, deve ser passível de ser observado.

Tal como exposto nas seções 4.4.3 e 4.4.3.2, o arranjo físico dos átomos no interior das proteínas permite inferir que as interações entre os átomos não se dá de forma inespecífica,

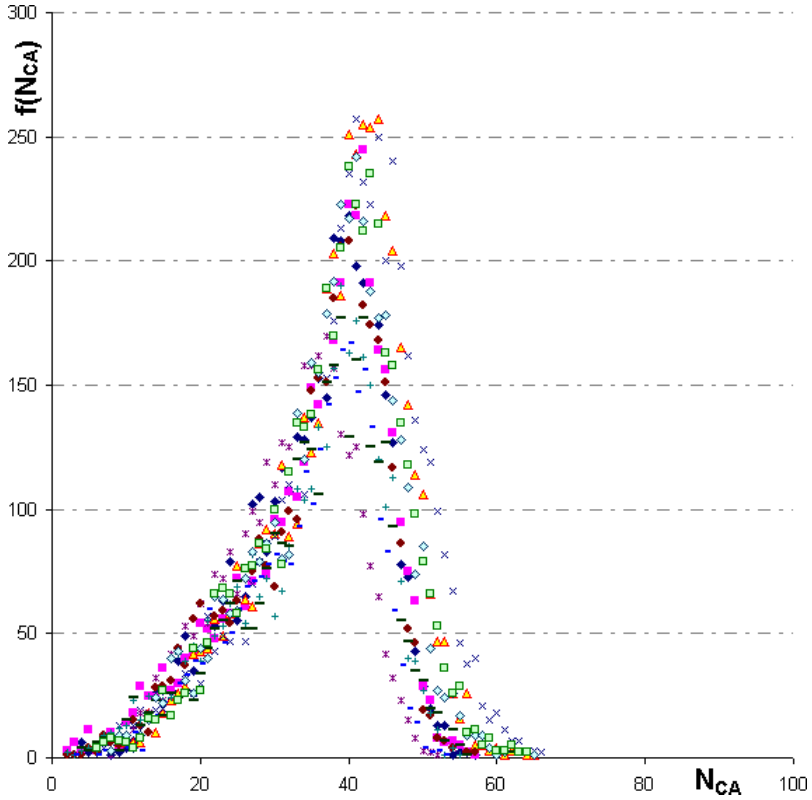


Figura 5.5 *– Distribuição, para as serinoproteases, das freqüências dos números de contatos por átomo ($f(N_{CA})$) e o número de contatos por átomo (N_{CA}) para proteínas não solvatadas sem critério de oclusão. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes.*

existindo limitações de ordem física que restringem as interações atômicas factíveis. Dizendo de outra forma, as restrições às interações atômicas são devidas aos impedimentos estéricos que os átomos impingem uns sobre os outros. Esta é a premissa sobre a qual se apóia o método descrito na seção 4.4.3.

Aplicando este método de análise (onde as interações são analisadas tendo em conta solvatação das proteínas e utilizando o critério de oclusão) às proteínas em estudo, obtém-se um novo conjunto de interações plausíveis. Tais interações são observadas considerando, por enquanto, o valor do comprimento euclidiano de cada uma e a quais átomos cada interação encontra-se vinculada. Para esta situação obtém-se a distribuição de número de contatos por átomo (N_{CA}) contra a freqüência do número de contatos por átomo ($f(N_{CA})$), apresentada nas figuras 5.6 e 5.7.

Com relação às figuras 5.4 e 5.5, estas novas distribuições de N_{CA} e $f(N_{CA})$ apresentam um novo padrão, mostrado na figura 5.6 para todas as globinas em estudo e na figura 5.7 para todas as serinoproteases em estudo. Estes novos padrões mostram-se mais regulares, sem a necessidade de nenhuma intervenção arbitrária. Nestas figuras, todas as interações não-covalentes entre os átomos das proteínas, que atendem aos critérios apresentados nas seções 4.4.3 e 4.4.3.2, são utilizadas no cômputo das distribuições estatísticas apresentadas.

Cada proteína estudada é representada, nestas figuras, por um conjunto de símbolos e cores.

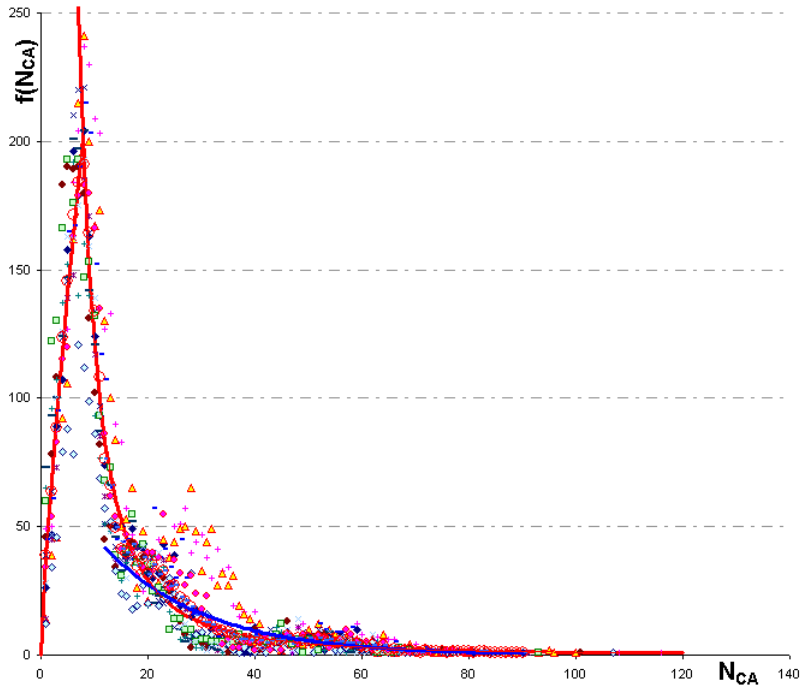


Figura 5.6 – *Distribuições de $f(N_{CA})$ e N_{CA} para as globinas solvatadas e analisadas com os critérios de oclusão. A distribuição de densidade de freqüências apresenta dois regimes distintos. Entre as abscissas $1 \leq N_{CA} \leq 8$, a distribuição de $f(N_{CA})$ apresenta crescimento médio em lei de potência – $f(N_{CA}) = \alpha N_{CA}^{\gamma} \mid \alpha = 38,85, \gamma = 0,80$ $(R^2 = 0,99)$. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. Entre as abscissas $8 \leq N_{CA} \leq 120$, o decaimento médio das curvas pode se ajustar tão bem a uma distribuição de lei de potência (curva em vermelho), quanto a uma distribuição exponencial (curva azul).*

Em ambas distribuições observa-se, nas figuras 5.6 e 5.7, que os gráficos apresentam duas regiões com comportamentos bem distintos: (1) – O segmento entre as abscissas $1 \leq N_{CA} \leq 8$ para as globinas e $1 \leq N_{CA} \leq 10$ para as serinoproteases; (2) – O segmento entre as abscissas $8 \leq N_{CA} \leq 120$ para as globinas e $10 \leq N_{CA} \leq 120$ para as serinoproteases. O segmento (1) em ambas famílias de proteínas mostra um crescimento médio seguindo uma lei de potência – $f(N_{CA}) = \alpha N_{CA}^{\gamma}$ com $\alpha = 38,85; \gamma = 0,80$ $(R^2 = 0,99)$ para o caso das globinas, e $\alpha = 64,36; \gamma = 0,64$ $(R^2 = 0,99)$ para o caso das serinoproteases.

O comportamento deste segmento, nas duas famílias de proteínas, revela o fenômeno de empacotamento dos átomos ("*atomic packing*"), que ocorre no núcleo hidrofóbico das mesmas. Tal fenômeno é análogo ao observado na física do estado sólido, o qual tem sido estudado na literatura sobre materiais granulados ("*granular matter*") [Silbert et al. (2002), Aste et al. (2004), Aste et al. (2006), Lochmann et al. (2006), Aste e Senden (2007)]. Este comportamento traz informações interessantes sobre a estabilidade do empacotamento e da coesão do núcleo da proteína. O crescimento em lei de potência do número de contatos, observado neste segmento da curva, é similar ao padrão observado em [Lochmann et al. (2006)], para a distribuição freqüência do número de contatos entre átomos, para um sistema

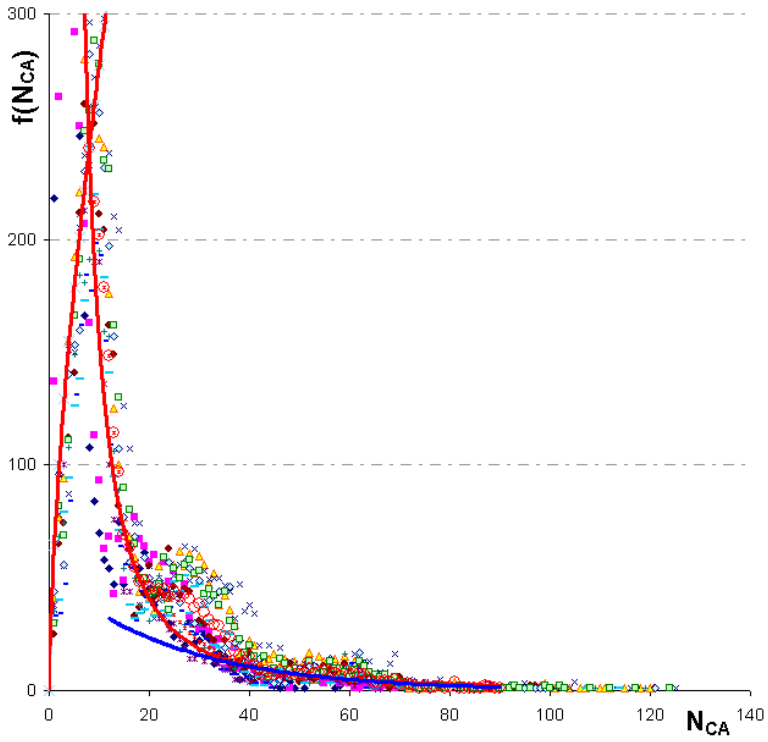


Figura 5.7 – *Distribuições de $f(N_{CA})$ e N_{CA} para as serinoproteases solvatadas e analisadas com os critérios de oclusão. A distribuição de densidade de freqüências apresenta dois regimes distintos. Entre as abscissas $1 \leq N_{CA} \leq 8$, a distribuição de $f(N_{CA})$ apresenta crescimento médio em lei de potência – $f(N_{CA}) = \alpha N_{CA}^{\gamma} \mid \alpha = 64,36, \gamma = 0,64$ $(R^2 = 0,99)$. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. Entre as abscissas $8 \leq N_{CA} \leq 120$, o decaimento médio das curvas pode se ajustar tão bem a uma distribuição de lei de potência (curva em vermelho), quanto a uma distribuição exponencial (curva azul).*

similar a um cristal amorfo.

Em seus trabalhos, Aste [Aste et al. (2004)] apresenta resultados de simulações de dinâmica molecular, onde sistemas de esferas rígidas idênticas, comportam-se como gases, quando a densidade do sistema – $\rho \simeq 0,49$. O comportamento similar aos fluídos é observado quando tem-se $\rho \simeq 0,55$, apresentando transição para a fase amorfa ("*glassy*") quando $\rho \simeq 0,56$ e, atingindo $\rho \simeq 0,645$ não é mais possível induzir o aumento de densidade do sistema. Nenhuma evidência empírica ou simulada sugere que algo especial deva ocorrer com a geometria do empacotamento quando o valor da densidade do sistema encontra-se na faixa $0,56 \leq \rho \leq 0,645$, que justifique este limiar do processo de densificação [Aste et al. (2004)]. Ainda em Aster, existem outros arranjos geométricos onde a densidade pode chegar a $\rho \sim 0,74$, como no caso no arranjo cúbico com faces centradas ("*fcc*"), ou no arranjo "*hexagonal closed-packed*". Aster ainda chama a atenção para o fato onde a relação entre o raio da maior esfera do sistema – R_{Max}, e o menor raio do sistema– R_{Min} ($r = R_{Max}/R_{Min}$), encontra-se na faixa $1,4 \leq r \leq 1,7$. Nesta situação particular, o sistema com arranjo desordenado *surpreendente-*

mente[1] apresenta uma eficiência de empacotamento média maior que a eficiência apresentada pelo arranjo ("*fcc*"), já que fora desta faixa a eficiência do arranjo desordenado é, no máximo, similar a apresentada pelo arranjo ("*fcc*"). Vale ressaltar que para as proteínas estudadas, o valor de $r = 1,5$, o que permite acreditar que o nível de compactação apresentado pelo núcleo hidrofóbico das proteínas seria o máximo alcançável por qualquer sistema de partículas nas mesmas condições.

Com base nos dados apresentados em [Aste e Senden (2007)], e reproduzidos na figura 5.8, tem-se que para o número de contatos (ou coordenação) mais freqüente apresentado pelas globinas ($N_c = 8$), a densidade do núcleo hidrofóbico desta família seria $\rho = 0,652$, valor pouco acima o limiar máximo observado em simulações.

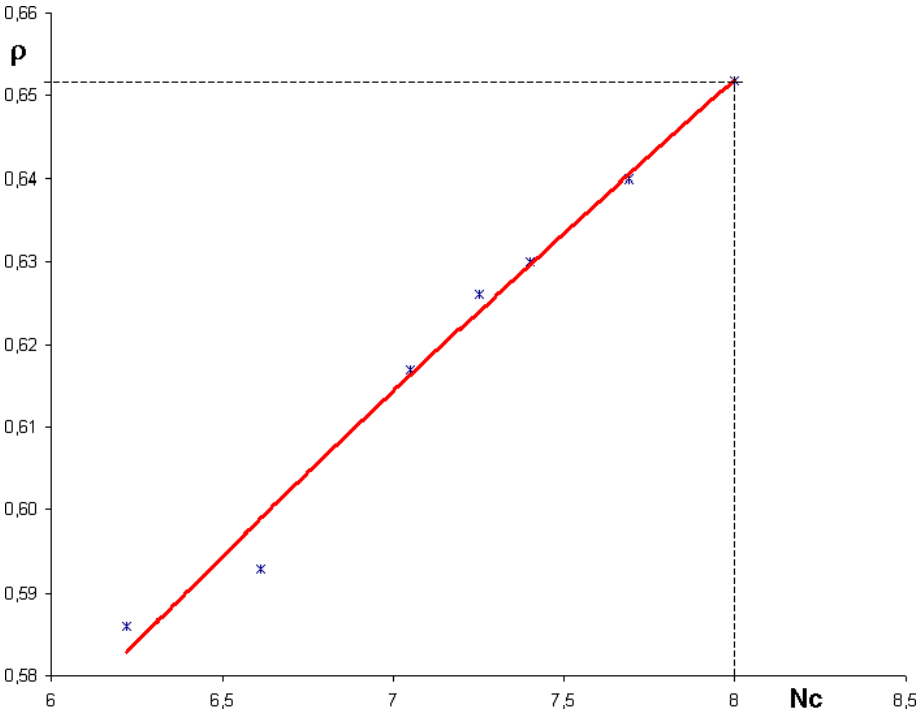


Figura 5.8 – *Número de coordenação médio - N_c, versus densidade do arranjo de partículas - ρ. As marcas no gráfico, correspondem aos arranjos de partículas estudados e apresentados por Aste e Senden [Aste e Senden (2007)]. Observa-se que para o número de coordenação mais freqüente apresentado pelas globinas ($N_c = 8$), a densidade do núcleo hidrofóbico desta família seria $\rho = 0,652$.*

O segmento (2) em ambas famílias de proteínas mostra um decaimento médio que pode se ajustar tão bem a uma distribuição de lei de potência (curva em vermelho), quanto a uma distribuição exponencial (curva azul). Os parâmetros representativos destas duas distribuições para as duas famílias de proteínas são apresentados na tabela 5.1.

Tais padrões indicam a existência de uma estrutura hierarquizada no arranjo dos vértices (átomos) e arestas (interações) das redes representativas das proteínas em estudo. Constata-se a existência de um pequeno número de vértices com grande número de ligações, ao mesmo

[1] Ênfase do próprio autor.

	Distribuição Exponencial $f(N_{CA}) = \alpha e^{-\gamma N_{CA}}$	Distribuição por Lei de Potência $f(N_{CA}) = \alpha N_{CA}^{-\gamma}$
globinas	$\alpha = 216{,}18$ $\gamma = 0{,}09$ $R^2 = 0{,}95$	$\alpha = 13721$ $\gamma = 2{,}06$ $R^2 = 0{,}97$
Serino-Proteases	$\alpha = 259{,}48$ $\gamma = 0{,}07$ $R^2 = 0{,}93$	$\alpha = 13100$ $\gamma = 1{,}93$ $R^2 = 0{,}94$

Tabela 5.1 *– Parâmetros representativos distribuições $f(N_{CA}) \times N_{CA}$ por lei de potência e exponencial com maior similaridade em relação às médias das distribuições das globinas e serinoproteases (N=120). Nas figuras 5.6 e 5.7, as distribuições por lei de potência estão em vermelho e as distribuições exponenciais estão em azul.*

tempo em que se observa a existência de um grande número de vértices com um pequeno número de ligações. Tais achados são consistentes com os apresentados em outros trabalhos [Vendruscolo et al. (2001), Vendruscolo et al. (2002), Greene e Higman (2003), Bastolla et al. (2005), Krishnadev et al. (2005), Higman e Greene (2006), Atilgan et al. (2007)].

Neste caso, aspectos geométricos de cada interação tornam-se insignificantes quando comparados ao ganho de informação, sobre cada uma destas interações, proporcionado pela energia potencial associada à estas interações. O aspecto mais relevante desta abordagem é que um peso pode ser dado à cada ligação das redes estudadas. Desta forma os pesos das arestas da rede representativa das proteínas em análise, apresentam uma significância física expressa como a energia das interações não-covalentes entre os átomos.

Para melhor entender a correlação entre o número de ligações por átomo (N_{CA}) e a energia potencial média por átomo (E_{MA}), tais valores são plotados para todas as proteínas (globinas e serinoproteases) em estudo. Estes resultados são apresentados na figura 5.9, onde pode ser observada a existência de uma forte correlação entre o número interações e a energia das interações não-covalentes relacionadas à cada átomo na proteína. Desta forma, a energia associada à cada interação não covalente pode ser vista como o peso das arestas quando a estrutura das proteínas é modelada como rede, implicando que se entre dois átomos A_{t1} e A_{t2} existe uma aresta $A_r(A_{t1}, A_{t2})$, então o peso desta aresta $w(A_r(A_{t1}, A_{t2}))$, é a energia potencial não covalente E_{ij} existente entre estes átomos:

$$\exists A_r(A_{t1}, A_{t2}) : w(A_r(A_{t1}, A_{t2})) = E_{ij}.$$

Considera-se agora as interações identificadas pela aplicação do método dos critérios de oclusão (apresentados na seção 4.4.3), ponderadas por suas respectivas energias potenciais. Em outras palavras, tais interações são observadas, agora, considerando o valor da energia potencial de cada uma e a quais átomos cada interação encontra-se vinculada. Para esta nova situação de análise, obtém-se a distribuição da energia potencial dos contatos por átomo (E_A) em $kcal/mol$, contra a freqüência dos valores de energia potencial por átomo ($f(E_A)$), para cada uma das famílias de proteínas estudadas. Tais distribuições são apresentadas nas

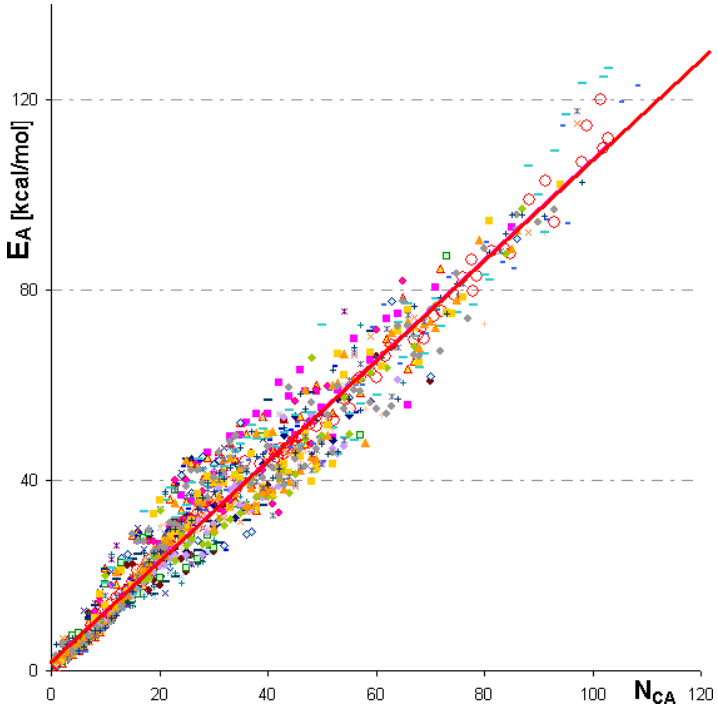


Figura 5.9 – *Relacionamento entre a média das energias de interação e o número de interações por átomo, para todas as proteínas em estudo. A linha central mostra uma correlação linear existente entre essas duas variáveis.*

figuras 5.10 e 5.11.

Para as distribuições das figuras 5.10 e 5.11, o decaimento ajusta-se muito melhor a uma distribuição de lei de potência que a uma distribuição exponencial. Os parâmetros representativos destas duas distribuições para as duas famílias de proteínas são apresentados na tabela 5.2.

Estes padrões indicam, que a adoção dos níveis de energia potencial como atributo ponderal das interações não-covalentes entre os átomos, para a análise das redes de interações, internas às proteínas, ressalta ainda mais a existência de uma estrutura hierarquizada dos vértices (átomos) e arestas (interações) das redes que representam as proteínas em estudo. Novamente, percebe-se a existência de um pequeno número de átomos fortemente conectados às suas vizinhanças e existência de um grande número de átomos fracamente ligados às suas respectivas vizinhanças. Comparando os valores apresentados nas tabelas 5.1 e 5.2, constata-se que a hierarquia dos vértices das redes subjacentes às proteínas formadas pelas interações não-covalentes entre os átomos que as formam, ajusta-se melhor a um padrão em lei de potência. Tal abordagem não havia sido feita anteriormente, para qualquer tipo de proteína, em qualquer dos trabalhos precedentes.

Tal comportamento é consistente com o comportamento das redes livres de escala [Albert e Barabasi (1999), Barabasi et al. (1999)], indicando que a rede de interações não-covalentes entre os átomos de uma proteína segue o mesmo padrão. A análise estatística mostra que as

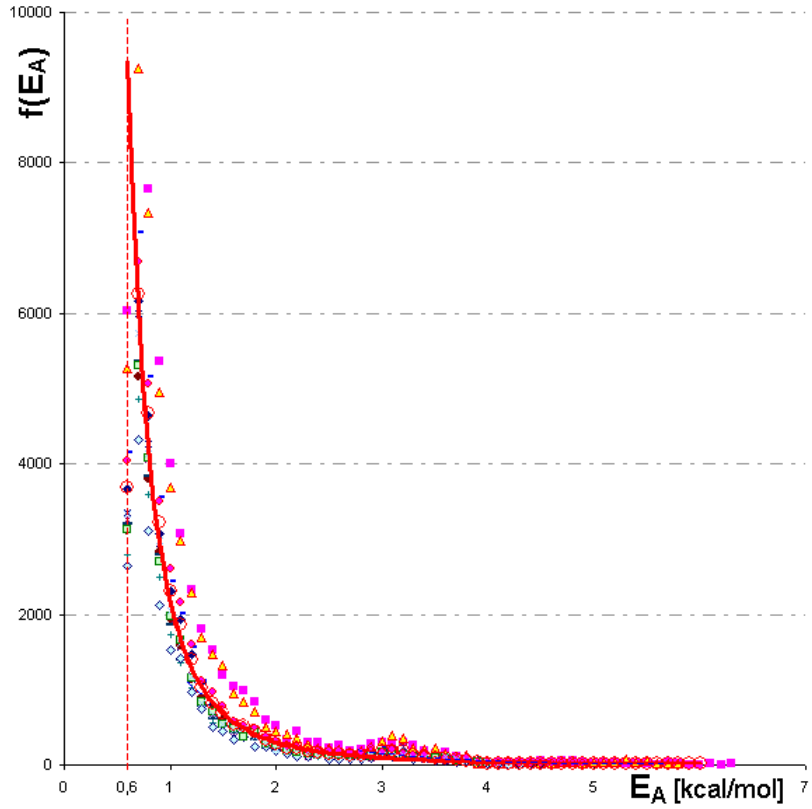


Figura 5.10 – *Distribuições, para as globinas, da freqüência de níveis de energia por átomo $f(E_A)$, e níveis de energia por átomo E_A (em kcal/mol) para o a proteínas solvatadas com oclusão. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média (em vermelho) segue uma curva de potência.*

curvas de tendência média para ambas distribuições seguem preferencialmente o padrão

$$f(E_A) \propto E_A^{-\gamma} \tag{5.1}$$

Para as distribuições analisadas, as curvas de tendência média podem ser expressas como: $f(E_A) \propto E_A^{-2,85}$ com $R^2 = 0,96$ para as globinas, e $f(E_A) \propto E_A^{-2,64}$ com $R^2 = 0,97$ para as serinoproteases (tabela 5.2). Os valores de γ neste caso, quando comparados com os mesmos valores apresentados na tabela 5.1, mostram-se bem mais próximos dos valores dos expoentes identificados para as demais redes representativas dos diversos fenômenos estudados em outros trabalhos [Albert et al. (1999), Faloutsos et al. (1999), Barabasi et al. (2000), Jeong et al. (2001), Liljeros et al. (2001), Ravasz et al. (2002), Vazquez e Weigt (2003)], onde para todos os casos estudados, tem-se $2 \leq \gamma \leq 3$. Em Newman [Newman (2003c),pgs. 23-24] existe a prova formal (a qual foge ao escopo deste trabalho) que apresenta as implicações para as redes, da variação do expoente γ. Aqui interessa ressaltar que para valores de $\gamma < 2$, o componente gigante da rede passa a ocupar toda a rede. Nesta condição a rede seria totalmente conectada. Por outro lado, para $\gamma > 3,479$ a rede se torna fragmentada e perde sua integridade. À medida em que o valor de γ tende para o valor limite de $3,479$,

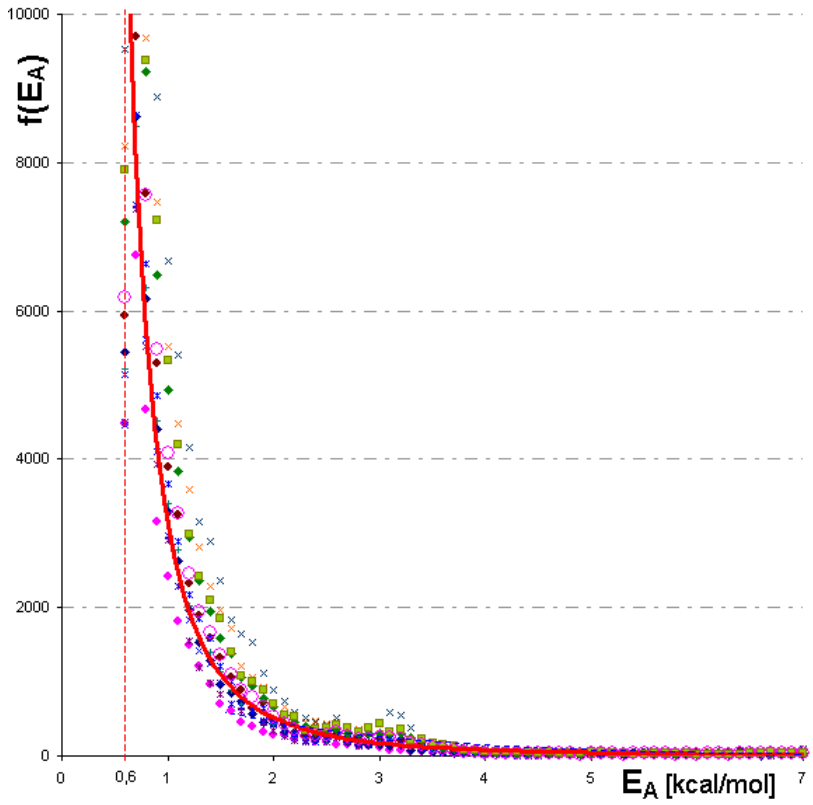


Figura 5.11 – *Distribuições, para as serinoproteases, da freqüência de níveis de energia por átomo $f(E_A)$, e níveis de energia por átomo E_A (em kcal/mol) para o a proteínas solvatadas com oclusão. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média (em vermelho) segue uma curva de potência.*

o decaimento da distribuição de $f(E_A)$ torna-se cada vez mais alongado. Nas redes com este perfil, os nodos mais conectados tendem a ser desproporcionalmente sujeitos a disparar efeitos de escala global, caso sejam afetados.

Diante dos resultados obtidos é possível propor que a rede de interações não-covalentes subjacente à estrutura 3D das proteínas apresenta uma hierarquia de nodos, onde aqueles mais energeticamente conectados exercem um papel importante para a estabilidade desta rede. Ao mesmo tempo, estas redes apresentam um componente gigante que garante a percolação, por toda a malha, dos impulsos percebidos por qualquer um dos nodos da rede. Em especial, os nodos mais conectados são os principais difusores destes impulsos por toda a estrutura das proteínas.

A constatação da existência desta hierarquia entre os átomos na estrutura das proteínas permite iniciar uma discussão acerca da resilência da rede de interações não-covalentes. Uma propriedade interessante das redes em lei de potência é a robustez que as mesmas apresentam à remoção de seus vértices. Tal propriedade tem sido tema de vários estudos encontrados na literatura [Callaway et al. (2000), Holme et al. (2002), Watts (2002), Newman (2003c), Vazquez e Moreno (2003)]. A maioria das redes observadas no mundo real devem suas propriedades às suas conectividades, ou seja, à existência de caminhos que ligam pares de

	Distribuição Exponencial $f(N_{CA}) = \alpha e^{-\gamma N_{CA}}$	Distribuição por Lei de Potência $f(N_{CA}) = \alpha N_{CA}^{-\gamma}$
globinas	$\alpha = 3724{,}40$ $\gamma = 1{,}06$ $R^2 = 0{,}82$	$\alpha = 2178$ $\gamma = 2{,}85$ $R^2 = 0{,}96$
Serino-proteases	$\alpha = 1431{,}10$ $\gamma = 0{,}57$ $R^2 = 0{,}83$	$\alpha = 3153{,}1$ $\gamma = 2{,}64$ $R^2 = 0{,}97$

Tabela 5.2 *– Parâmetros representativos distribuições $f(E_A) \times E_A$ por lei de potência e exponencial com maior similaridade em relação às médias das distribuições das globinas e serinoproteases (N=120). Nas figuras 5.10 e 5.11, as linha de tendência por lei de potência estão em vermelho.*

vértices não adjacentes. Se os vértices de uma rede forem eliminados, o comprimento típico destas irá aumentar e à medida que os nodos forem retirados alguns pares de nodos serão totalmente desconectados e a comunicação entre os mesmos será impossível. As redes variam na sua resilência à tal processo de remoção de vértices.

Albert *et al.* [Albert e Barabasi (2000)] estudaram os efeitos da remoção aleatória e direcionada dos vértices de diferentes redes. Neste estudo, a distância média entre os vértices era calculada em função do número de vértices retirados. Dois processos de remoção eram adotados. Em um caso os vértices eram escolhidos de forma aleatória e eliminados. No outro caso, o vértice com maior conectividade, em cada interação, era escolhido e eliminado. Destes experimentos foi possível observar, para as redes estudadas, que a distância média entre os vértices não era significativamente afetada no processo de remoção aleatória. Em outras palavras, as redes estudadas mostravam-se resilentes a este tipo de remoção. Isto é bastante intuitivo, já que a maioria dos vértices, nestas redes, apresenta baixa conectividade. A remoção de um destes nodos pouco afeta a comunicação dos demais. Por outro lado, quando a remoção era direcionada aos nodos com maior grau, percebeu-se que o efeito era devastador para a integridade das redes. Neste caso, à medida que os nodos eram eliminados, a distância média entre nodos, de todas as redes, crescia vertiginosamente. Tipicamente, a retirada de um pequeno percentual do total de nodos das redes era suficiente para destruir toda a integridade estrutural das mesmas.

De forma análoga, seria de esperar que mutações, em uma proteína, que alterem significativamente a malha formada pelos resíduos com maior número de interações não-covalentes, gerariam estruturas instáveis ou mesmo desnaturadas. Contudo, mais à frente, será mostrado que a evolução das proteínas levou ao desenvolvimento de outros artifícios que atenuam os impactos destas possíveis mutações.

5.2.2 Transitividade e distâncias entre os átomos: Índices para as globinas e serinoproteases

A distribuição do grau de conectividade dos nodos de uma rede está entre os parâmetros

estatísticos mais adotados para descrever os aspectos estruturais das redes complexas. Porém, outros parâmetros apresentam igual importância para caracterização das redes que permitem a inferência de algumas propriedades dinâmicas do sistema. Nesta seção serão analisados os parâmetros "coeficiente de aglomeração" (ou "*clustering*"), e o "número médio de passos entre nodos" para as proteínas em estudo.

Os resultados discutidos nesta seção fazem referência aos índices previamente apresentados seção 3.3.

5.2.2.1 Índices relativos às globinas

A análise da rede formada pelas ligações não-covalentes presentes no interior das globinas, gerou resultados relativos aos índices de aglomeração para cada átomo destas proteínas. O índice de aglomeração médio para cada resíduo de aminoácido constituinte das globinas foi calculado tomando a média dos valores dos respectivos átomos. Os resultados obtidos são apresentados na figura 5.12.

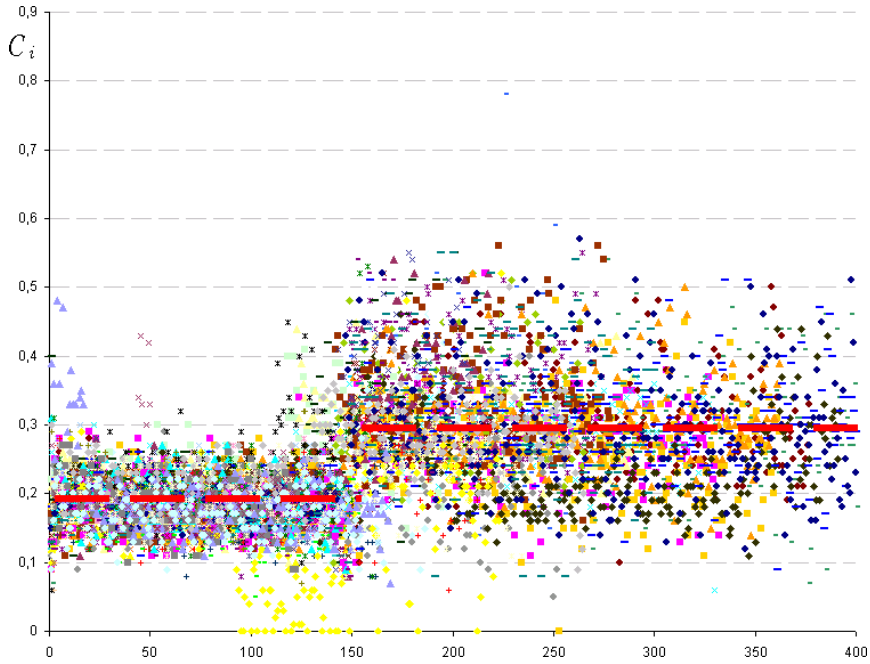


Figura 5.12 *– Distribuições, para as globinas, dos índices de aglomeração por resíduo na estrutura primária. O eixo das abscissas mostra os índices dos resíduos (abscissas $\leq$ 150), e dos demais grupos químicos (abscissas $>$ 150) das proteínas. O eixo das ordenadas mostra os índices de aglomeração apresentados pelas proteínas. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média está em vermelho. O aglomerado de pontos à esquerda mostra o comportamento dos resíduos das proteínas, o mais à direita mostra o comportamento das moléculas de água estruturais.*

Nesta figura cada uma das globinas estudadas está representada por um grupo distinto de símbolos e cores. O valor que corresponde à cada um dos resíduos de aminoácido - i, de cada proteína é marcado no eixo das abscissas. De forma análoga, os valores referentes aos índices de aglomeração são marcados no eixo das ordenadas - C_i.

É possível perceber, neste gráfico, duas regiões distintas para a distribuição dos valores dos índices de aglomeração. Mais à esquerda do gráfico estão representados os valores para os índices de aglomeração dos resíduos que formam a estrutura primária das proteínas. Mais à direita, neste mesmo gráfico, estão representados os valores para os índices de aglomeração das moléculas que não fazem parte da estrutura primária das proteínas, tais como moléculas de água, grupos prostéticos, etc. As linhas horizontais em vermelho, representam o índice de aglomeração médio para todas as proteínas analisadas, segmentado conforme as regiões já descritas do gráfico.

A observação dos índices de aglomeração relativos aos resíduos das proteínas mostra que eles encontram-se concentrados em uma estreita faixa de valores entre 0,1 e 0,3. De fato, o coeficiente médio de aglomeração para as globinas estudadas ficou $\langle C \rangle = 0,22$, com desvio padrão em $\sigma = 0,06$. A tabela 5.3 mostra coeficiente de aglomeração para as globinas estudadas.

PDBID	$\langle C \rangle$	σ	PDBID	$\langle C \rangle$	σ
1A6G	0,24	0,07	1ASH	0,19	0,03
1B0B	0,25	0,06	1BIN	0,26	0,10
1BZP	0,25	0,08	1C40	0,18	0,03
1CG5	0,19	0,04	1D8U	0,20	0,06
1DLW	0,24	0,06	1DLY	0,25	0,07
1DWT	0,21	0,04	1ECD	0,20	0,03
1FAW	0,17	0,02	1FDH	0,17	0,03
1FHJ	0,24	0,10	1G09	0,17	0,03
1GCV	0,25	0,09	1GDJ	0,23	0,07
1H97	0,25	0,08	1HBR	0,17	0,03
1HDS	0,17	0,03	1HLB	0,20	0,06
1HLM	0,17	0,03	1I3D	0,26	0,09
1IDR	0,25	0,09	1IT2	0,22	0,06
1ITH	0,27	0,10	1JF3	0,24	0,09
1JF4	0,25	0,10	1KR7	0,23	0,06
1LA6	0,23	0,10	1LHS	0,20	0,05
1MBS	0,16	0,04	1MWD	0,26	0,07
1MYT	0,16	0,09	1NS9	0,17	0,03
1NXF	0,19	0,04	1OUT	0,20	0,06
1Q1F	0,20	0,06	1QPW	0,26	0,09
1RTX	0,20	0,04	1S5Y	0,21	0,10
1SPG	0,25	0,10	1TU9	0,25	0,06
1UVX	0,18	0,04	1V5H	0,19	0,06
1VHB	0,19	0,05	2FAL	0,21	0,06
2MM1	0,18	0,04			

Tabela 5.3 *– Coeficiente médio de aglomeração para as globinas em estudo. Média para as globinas é 0,22 ± 0,06*

Quando estes valores são comparados com outros valores apresentados por outros autores

[*op.cit.* Newman (2003c)], é possível perceber que a faixa de valores é similar à de outras redes biológicas encontradas no mundo real e que já foram estudadas em outros trabalhos (tabela 5.4).

Descrição	$\langle C \rangle$
Redes Metabólicas	0,670
Interações entre proteínas	0,071
Rede de alimentar Marinha	0,230
Rede alimentar em água doce	0.200
Rede neural	0,280

Tabela 5.4 – Alguns coeficientes para redes biológicas já estudadas.

Comparativamente, o índice de aglomeração valor médio $\langle C \rangle$, apresentado pelas globinas mostra-se similar ao apresentado por algumas redes biológicas encontradas no mundo real. Tal como apresentado na seção 3.4.2, o coeficiente de aglomeração médio $\langle C \rangle$, é uma métrica indicativa do grau de aglomeração (ou coesividade) da proteína, ao mesmo tempo em que foi também calculada para todo átomo i das proteínas. Esta última métrica mostra, para cada átomo existente nas proteínas, como este está energeticamente vinculado aos demais átomos presentes em sua vizinhança. Quantitativamente, os índices apresentados pelas globinas indicam que as proteínas desta família são bastante coesas, o que é corroborado por outros estudos onde o valor médio de densidade de massa – ρ, apresentado pelas proteínas seria $\rho \approx 0,7$ (*vide* pág 93).

Com relação às distâncias entre átomos nas globinas estudadas, alguns aspectos importantes merecem ser relembrados antes da apresentação dos resultados. Como citado na seção 3.3.1, os índices relativos às distâncias entre nodos de uma rede exercem uma influência importante na emergência do efeito de "*small-world*" nas redes. Este efeito tem implicações óbvias para a dinâmica dos processos que lá ocorrem. No caso particular das proteínas, a propagação dos efeitos locais (como a acomodação de ligantes em sítios ativos), por toda a estrutura da proteína, pode implicar na alteração de toda a sua topologia e de seu comportamento. Desta forma, quanto mais pronunciado for o efeito "*small-world*" na estrutura de uma proteína, mais rápida será a propagação de perturbações por toda sua estrutura e mais rápida será sua resposta a tais perturbações.

Os resultados das análises referentes aos índices de distância entre átomos, nas globinas estudadas, são apresentados na figura 5.13. Seguindo a mesma convenção adotada neste trabalho, cada uma das globinas estudadas está representada por um grupo distinto de símbolos e cores. O valor que corresponde à cada um dos resíduos de aminoácido de cada proteína é marcado no eixo horizontal - i. De forma análoga, os valores referentes aos índices de distância média entre átomos estão marcados no eixo vertical - L_i.

De forma similar ao observado para os índices de aglomeração, percebe-se, neste gráfico, a existência de duas regiões distintas para a distribuição dos valores dos índices de distância.

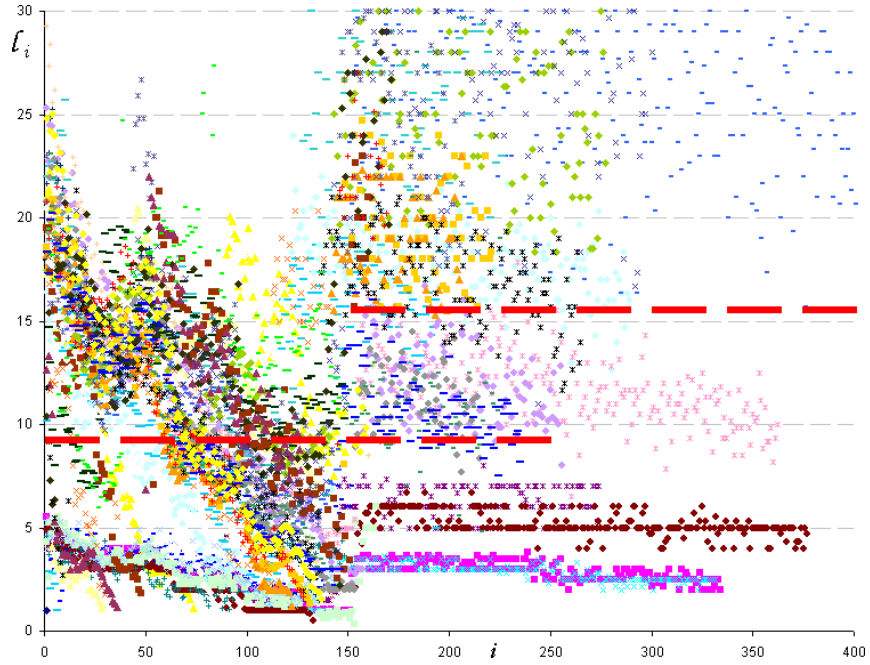


Figura 5.13 *– Distribuições, para as globinas, dos índices de distância média - L_i (em passos) entre átomos, por resíduo presente estrutura primária. O eixo das abscissas mostra os índices dos resíduos (abscissas $\leq$ 150), e dos demais grupos químicos (abscissas $>$ 150) das proteínas. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média está em vermelho. O aglomerado de pontos à esquerda mostra o comportamento dos resíduos das proteínas, o mais à direita mostra o comportamento das moléculas de água estrutural.*

Mais à esquerda do gráfico estão representados os valores para os índices relativos aos resíduos que formam a estrutura primária das proteínas. Mais à direita, estão representados os valores para os índices de aglomeração das moléculas que não fazem parte da estrutura primária das proteínas (água, grupos prostéticos, etc). As linhas horizontais em vermelho, representam o valor médio dos índices para as proteínas analisadas, segmentado conforme as regiões já descritas do gráfico.

A observação dos índices de distância média entre átomos, relativos aos resíduos das globinas, mostra os valores concentrados em uma faixa ampla de valores variando de 2,90 a 20,52. De fato, o coeficiente médio para as globinas estudadas é $\langle L \rangle = 9,16$, com desvio padrão em $\sigma = 4,05$. A tabela 5.5 mostra coeficiente de aglomeração para as globinas estudadas.

Os resultados apresentados na figura 5.13 mostram que, diferente do comportamento dos índices de aglomeração para as globinas, os índices de distância média apresentam uma grande variabilidade. Esta variabilidade indica a existência de uma assimetria entre a capacidade de percolação entre os resíduos que compõem as globinas. Ou seja, a transmissão de sinais entre os resíduos de uma Globina com maior distância média entre átomos, ocorreria de forma assimétrica, existindo, desta forma, regiões da proteína que responderiam mais prontamente às perturbações induzidas em uma região qualquer da estrutura, enquanto ou-

PDBID	$\langle L \rangle$	σ	PDBID	$\langle L \rangle$	σ
1A6G	2,86	0,86	1ASH	9,92	6,12
1B0B	2,77	0,79	1BIN	4,43	1,91
1BZP	3,92	1,82	1C40	3,06	1,14
1CG5	3,36	1,12	1D8U	6,73	2,56
1DLW	11,35	4,76	1DLY	13,61	6,02
1DWT	2,79	1,32	1ECD	9,44	5,91
1FAW	4,74	1,95	1FDH	3,08	1,16
1FHJ	2,37	0,95	1G09	3,33	1,54
1GCV	5,51	2,03	1GDJ	11,20	3,05
1H97	2,56	1,62	1HBR	4,62	1,00
1HDS	8,00	2,51	1HLB	11,28	4,69
1HLM	10,82	6,89	1I3D	20,52	7,88
1IDR	19,16	7,61	1IT2	5,26	1,31
1ITH	19,30	7,97	1JF3	12,34	7,01
1JF4	11,83	6,87	1KR7	10,04	5,21
1LA6	16,33	7,28	1LHS	10,51	4,94
1MBS	10,13	6,70	1MWD	4,17	1,00
1MYT	10,72	5,56	1NS9	5,26	1,18
1NXF	12,56	3,24	1OUT	14,45	5,95
1Q1F	10,58	5,61	1QPW	5,00	0,85
1RTX	9,00	5,83	1S5Y	11,79	1,31
1SPG	19,73	8,96	1TU9	12,95	5,15
1UVX	9,10	5,73	1V5H	11,17	7,18
1VHB	15,09	5,65	2FAL	10,19	4,11
2MM1	10,00	6,47			

Tabela 5.5 – *Coeficiente médio de distância entre átomos para as globinas em estudo. Média dos valores é 9,16 ± 4,05*

tras iriam responder com um retardo maior. Caso os valores deste índice (para os resíduos desta proteína), fossem mais homogêneos, a transmissão dos sinais induzidos por quaisquer perturbações percolariam por toda a estrutura de forma mais homogênea.

Observando com mais acuidade os pontos localizados, na figura 5.13, com abscissas compreendidas entre $1 < i < 150$, podem ser observados dois aglomerados distintos de pontos, um localizado entre as coordenadas $1,5 < L_i < 5,5$ e outro entre as coordenadas $1,5 < L_i < 25,5$.

O aglomerado localizado entre as coordenadas $1,5 < L_i < 5,5$ (vamos chamá-lo de "1-5"), apresenta pontos com índice de distância entre átomos, com uma variabilidade bem menor que aquela apresentada pelo aglomerado localizado entre as coordenadas $1,5 < L_i < 25,5$ (vamos chamá-lo de "1-25"). As globinas representadas nestes diferentes aglomerados apresentam propriedades de percolação com comportamentos bem distintos.

As globinas do aglomerado "1-5" devem apresentar uma percolação de sinais mais eficaz que a apresentada pelas globinas do aglomerado "1-25", já que os índices de transitividade são mais homogêneos, tendo uma transitividade média de 3,3 passos, enquanto a média do

aglomerado “1-25” fica no entorno de 12,5 passos. Em outras palavras, conjectura-se que as globinas em “1-5” devam apresentar a tendência para responder mais rápido às perturbações, quando comparadas às globinas do grupo “1-25”. Uma amostra das globinas que participam de “1-5” é apresentada na tabela 5.6, enquanto das globinas que participam de “1-25” é apresentada na tabela 5.7.

Tabela 5.6 *– globinas do aglomerado“1-5”*

PDBID	Organismo	Taxonomia	Nome	Mutante
1A6G	Physeter catodon	Phylum: Chordata Subphylum: Craniata Class: Mammalia Order: Cetacea Family: Physeteridae	Cachalote	N
1B0B	Lucina pectinata	Phylum: Mollusca Class: Bivalvia Order: Veneroida Family: Lucinidae	Lambreta	N
1BIN	Glycine max	Phylum: Streptophyta Class: Rosids Order: Fabales Family: Fabaceae	Soja	N
1BZP	Physeter catodon	Phylum: Chordata Subphylum: Craniata Class: Mammalia Order: Cetacea Family: Physeteridae	Cachalote	N
1C40	Anser indicus	Phylum: Chordata Subphylum: Craniata Class: Aves Order: Anseriformes Family: Anatidae	Ganso	N
1FDH	Homo Sapiens	Phylum: Chordata Subphylum: Craniata Class: Mammalia Order: Primates Family: Hominidae	Hemoglobina Fetal	N

Tabela 5.7 *– globinas do aglomerado "1-25"*

PDBID	Organismo	Taxonomia	Nome	Mutante
1ASH	Ascaris suum	Phylum: Nematoda Class: Chromadorea Order: Ascaridida Family: Ascarididae		N
1GDJ	Lupinus luteus	Phylum: Streptophyta Class: Rosids Order: Fabales Family: Fabaceae	Tremoço amarelo	N
1HLB	Caudina arenicola	Phylum: Echinodermata Class: Holothuroidea Order: Molpadiida Family: Caudinidae	Pepino do Mar	N
1HLM	Caudina arenicola	Phylum: Echinodermata Class: Holothuroidea Order: Molpadiida Family: Caudinidae	Pepino do Mar	N
1LHS	Caretta caretta	Phylum: Chordata Subphylum: Craniata Class: Reptilia Order: Testudines Family: Cheloniidae	Tartaruga Marinha	S
1MBS	Phoca vitulina	Phylum: Chordata Subphylum: Craniata Class: Mammalia Order: Carnivora Family: Phocidae	Foca	S
1MYT	Thunnus albacares	Phylum: Chordata Subphylum: Craniata Class: Actinopterygii Order: Perciformes Family: Scombridae	Albacora-laje	S
2FAL	Aplysia limacina	Phylum: Mollusca Class: Gastropoda Order: Opisthobranchia Family: Aplysiidae		S

Continua . . .

Tabela 5.7 – *(continuação)*

PDBID	Organismo	Taxonomia	Nome	Mutante
2MM1	Homo Sapiens	Subphylum: Craniata Class: Mammalia Order: Primates Family: Hominidae	Homen	S

Da leitura da tabela 5.6 observa-se que, curiosamente, todas estas proteínas são "nativas"[2]. Ao mesmo tempo, o grupo é formado por proteínas oriundas de organismos taxonomicamente diferentes, havendo aí proteínas de vegetais, aves e mamíferos. Por outro lado, na tabela 5.7, encontram-se proteínas "nativas" e mutantes. Da mesma forma, as proteínas referem-se a organismos taxonomicamente bem distintos. Estes dados, são mostrados buscando evidenciar que estas discrepâncias, a princípio, não devem estar relacionadas aos fatores aí apresentados.

Buscando avaliar outros fatores que poderiam explicar estes achados, uma breve análise das estruturas secundárias e terciárias das globinas dos dois grupos foi conduzida. Contudo, tal análise não revelou nenhum padrão especial que pudesse justificar estes diferentes comportamentos dos índices de distância média entre átomos para estes dois grupos.

Foi possível identificar que, para as globinas do grupo "1-5" (partindo do alinhamento estrutural de seis destas globinas), oitenta e quatro resíduos pertencentes a hélices, alinham-se perfeitamente, sendo necessária a inclusão de dezenove "gaps". Ao seu turno, as globinas do grupo "1-25", foram alinhadas outras seis proteínas, e foram identificados oitenta e um resíduos pertencentes a hélices, perfeitamente alinhados, com a inclusão de 21 "gaps". Isto talvez seja um indicativo de uma maior conservação de padrões estruturais secundários para globinas.

Este ponto em particular merece um estudo mais detalhado e, mesmo que a elucidação deste problema fuja ao escopo deste trabalho, seria particularmente útil que isto venha a ser tema de estudos posteriores.

Com o intuito de complementar a análise destas discrepâncias, os gráficos das figuras 5.14 e 5.15 foram elaborados. Na figura 5.14, são apresentadas em separado, as distribuições das distâncias médias entre os átomos, para as globinas monoméricas. Já na figura 5.15 são apresentadas as distribuições para as globinas multiméricas. Observando os pontos localizados no intervalo de abscissas $[1, 150]$, nota-se que os padrões de distribuição de distâncias entre átomos, para os átomos constituintes dos resíduos das proteínas, não muda significativamente entre os dois grupos. Isto indica que, os fatores estruturais que implicam nestas discrepâncias deste índice, para os dois grupos identificados, não devem estar associados ao fato da proteína ter uma ou múltiplas cadeias.

[2] O termo em inglês é "wild" ("selvagem"), e indica que a proteína encontra-se na forma em que ela é encontrada no organismo de origem.

Entretanto, a observação dos pontos localizados entre as abscissas $150 \leq i \leq 400$, nos dois gráficos, já mostra comportamentos dignos de nota. Estes pontos são representativos das moléculas de água estruturais presentes nas globinas em estudo. É interessante observar que, para as globinas monoméricas, o índice médio de distância entre átomos é menor que aquele apresentado pelas globinas multiméricas.

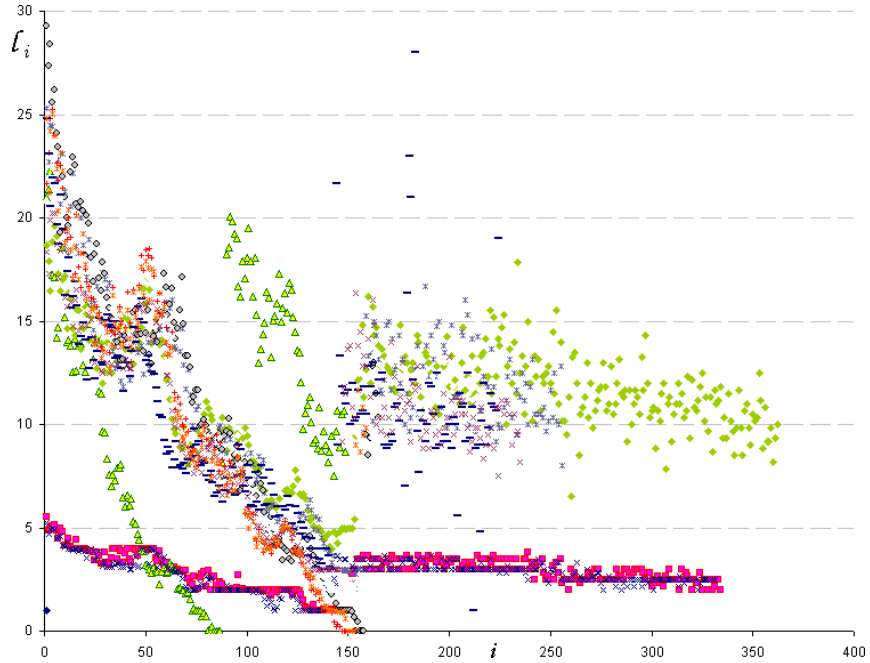


Figura 5.14 *– Distribuições, para as globinas monoméricas, dos índices de distância média - L_i (em passos) entre átomos, por resíduo na estrutura primária. O eixo das abscissas mostra os índices dos resíduos (abscissas $\leq$ 150), e dos demais grupos químicos (abscissas $>$ 150) das proteínas. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média está em vermelho. O aglomerado de pontos à esquerda mostra o comportamento dos resíduos das proteínas, o mais à direita mostra o comportamento das moléculas de água estrutural.*

Outro aspecto interessante de ser analisado, para avaliar em que medida as estruturas das globinas em análise, aderem aos modelos teóricos de redes complexas. Como apresentado na seção 3.3.1, uma rede apresenta propriedades que a definem como apresentando o fenômeno de "*small-world*", se para qualquer nodo desta rede, o número de nodos N, presentes dentro de um raio r crescer exponencialmente com r (ou seja $N \propto e^r$), e o valor da distância média entre nodos crescer logaritimicamente (ou a taxas menores) com N (ou seja $L \propto \log(N)$) [Newman (2003c)]. Desta forma, buscou-se avaliar, para as globinas em estudo, como o número de átomos presentes dentro de uma esfera com raio variável r, evolui à medida que este raio varia. Da mesma forma, avaliou-se como a distância média entre os átomos varia em função deste raio. Os resultados obtidos estão apresentados nos gráficos das figuras 5.16 e 5.17.

Para entender o raciocínio associado aos resultados apresentados nestas figuras, é útil que alguns conceitos sejam melhor explicados. Inicialmente, considere um átomo qualquer i

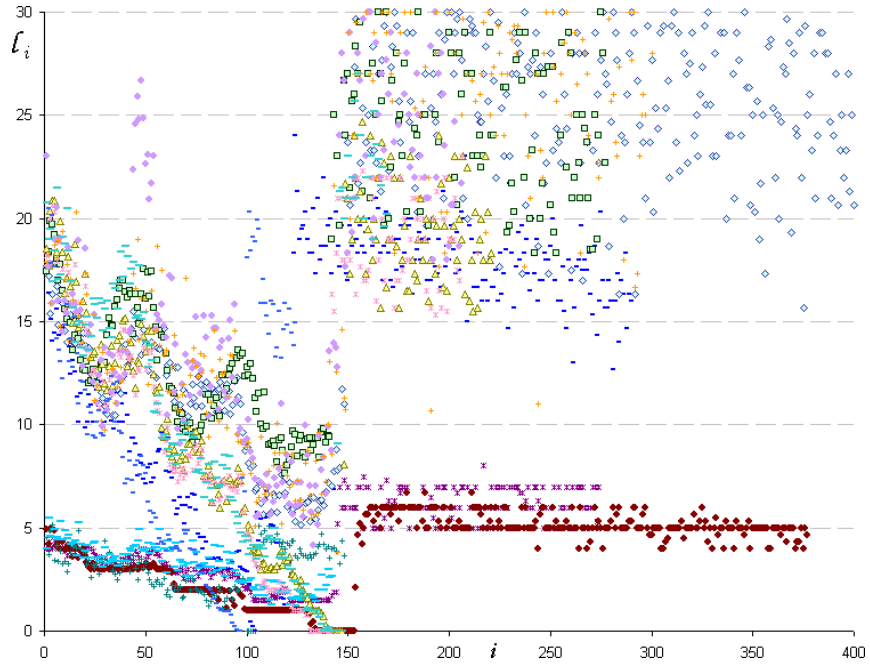


Figura 5.15 – *Distribuições, para as globinas multiméricas, dos índices de distância média - L_i (em passos) entre átomos, por resíduo na estrutura primária. O eixo das abscissas mostra os índices dos resíduos (abscissas $\leq$ 150), e dos demais grupos químicos (abscissas $>$ 150) das proteínas. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média está em vermelho. O aglomerado de pontos à esquerda mostra o comportamento dos resíduos das proteínas, o mais à direita mostra o comportamento das moléculas de água estrutural.*

pertencente a uma proteína. Seja então $N(i,r)$ o número de átomos compreendidos dentro de uma esfera com centro em i e de raio r, e $N(r)$ o número médio de átomos encontrados em todas as esferas com raio r encontradas na análise.

Fazendo r variar de 1Å até 10Å com[3] incrementos de 1Å, calcula-se $N(i,r)$ para cada um dos valores de r. Este cálculo é repetido para cada um dos átomos i, constituintes de cada uma das proteínas estudadas. Para cada um dos valores alcançados por r, foi feita a média de $N(i,r)$ ($N(r)$), para todas as proteínas pertencentes à cada um dos grupos estudados (*i.e* tanto para globinas como para serinoproteases em separado). Estes valores de $N(r)$ para as globinas estão apresentados na figura 5.16. A curva de tendência média apresentada no gráfico 5.16 revela que $N(r)$ varia com r segundo uma exponencial, ou seja: $N(r) \propto \beta e^{\alpha r}$. Este comportamento ajusta-se ao comportamento previsto, para este parâmetro, do o modelo de Watts [Watts e Strogatz (1998)], com $\beta = 0,86$ e $\alpha = 0,41$ e $R^2 = 0,96$.

Por outro lado, para entender o raciocínio relativo ao crescimento da distância média entre nodos, considere novamente um átomo qualquer i pertencente a uma proteína e que o parâmetro $N(i,r)$ está previamente definido. Define-se o parâmetro $L(i,r)$ como sendo a distância média (em passos) do átomo i a todos os $N(i,r)$ átomos compreendidos em uma

[3]Este é o limite identificado na seção 5.1 para as distâncias possíveis entre os átomos para todas as proteínas estudadas.

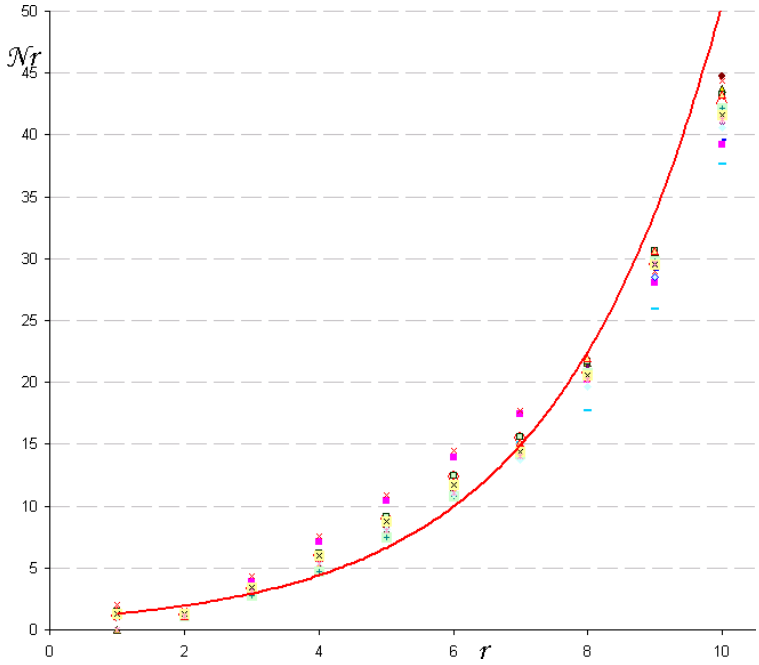


Figura 5.16 – *Distribuições, para as globinas, da evolução do número de átomos $N(r)$, presentes dentro de um raio r em função do crescimento deste raio. Observa-se o crescimento exponencial de $N(r)$. Para as globinas analisadas, $N(r)$ ajusta-se a uma curva exponencial $N(r) = 0,86e^{0,41r}$ com $R^2 = 0,96$.*

esfera com centro em i e de raio r.

Fazendo r variar de 1Å até 10Å com incrementos de 1Å, calcula-se $L(i,r)$ para cada um dos valores de r. Este cálculo é repetido para cada um dos átomos i, constituintes de cada uma das proteínas estudadas. Para cada um dos valores alcançados por r, foi feita a média de $L(i,r)$ $(L(r))$, para todas as proteínas pertencentes a cada um dos grupos estudados. Estes valores de $L(r)$ para as globinas estão apresentados na figura 5.17. Para este parâmetro, a curva de tendência média apresentada no gráfico 5.17 revela que $L(r)$ varia com r segundo uma curva logarítmica (*i.e* $N(r) \propto \beta \ln(r) + \gamma$). Novamente percebe-se que o comportamento deste parâmetro, ajusta-se ao previsto para o modelo de Watts [Watts e Strogatz (1998)], com $\beta = 1,75$ e $\gamma = 0,58$ e $R^2 = 0,98$.

Face estes resultados, é possível afirmar que o comportamento dos índices de aglomeração e distância entre átomos, para o caso das globinas, apresenta um comportamento típico das estruturas que apresentam o fenômeno de "*small-world*". Daí, pode-se inferir que a estrutura das proteínas deve apresentar comportamentos com algum grau de analogia com aqueles estimados para os modelos de redes complexas, desta mesma classe.

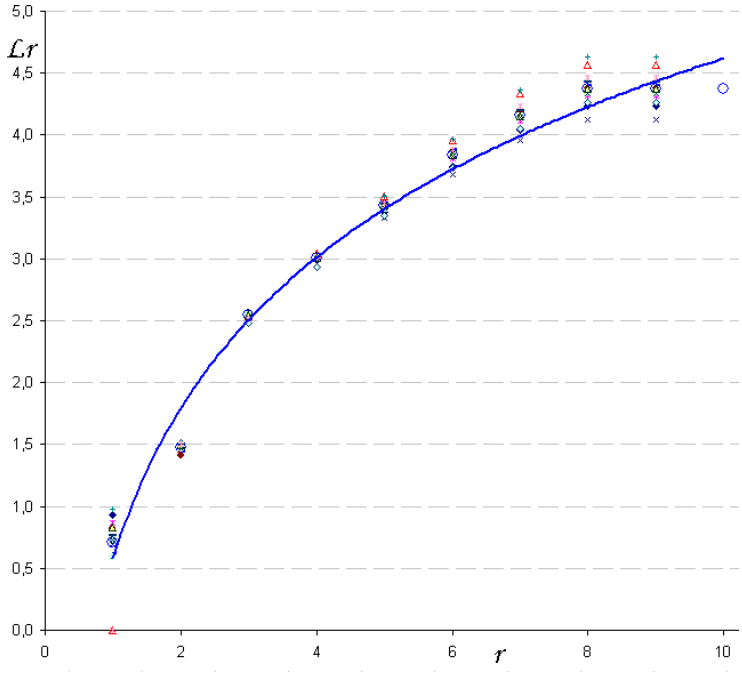


Figura 5.17 *– Distribuições, para as globinas, da evolução valor da distância média (em passos) entre átomos $L(r)$, presentes dentro de um raio r em função do crescimento deste raio. Observa-se o crescimento logarítmico de $L(r)$. Para as globinas analisadas, $L(r)$ ajusta-se a uma curva logarítmica $L(r) = 1,75\ln(r) + 0,58$ com $R^2 = 0,98$.*

5.2.2.2 Índices relativos às serinoproteases

A análise da rede formada pelas ligações não-covalentes presentes no interior das serinoproteases , gerou resultados relativos aos índices de aglomeração para cada átomo destas proteínas. O índice de aglomeração médio para cada resíduo de aminoácido constituinte das serinoproteases foi calculado tomando a média dos valores dos respectivos átomos. Os resultados obtidos são apresentados de forma não normalizada na figura 5.18.

Nesta figura cada uma das serinoproteases estudadas está representada por um grupo distinto de símbolos e cores. O valor que corresponde à cada um dos resíduos de aminoácido de cada proteína é marcado no eixo horizontal. De forma análoga, os valores referentes aos índices de aglomeração estão marcados no eixo vertical - C_i.

De forma similar às globinas, este gráfico apresenta duas regiões distintas para a distribuição dos valores dos índices de aglomeração. Mais à esquerda do gráfico estão representados os valores para os índices de aglomeração dos resíduos que formam a estrutura primária das proteínas. Mais à direita, neste mesmo gráfico, estão representados os valores para os índices de aglomeração das moléculas que não fazem parte da estrutura primária das proteínas, tais como moléculas de água, grupos prostéticos, etc. As linhas horizontais em vermelho, representam o índice de aglomeração médio para todas as proteínas analisadas, segmentado conforme as regiões já descritas do gráfico.

A observação dos índices relativos aos resíduos das serinoproteases mostra que eles encontram-se concentrados em uma estreita faixa de valores entre 0,1 e 0,3, tal como encontrado para

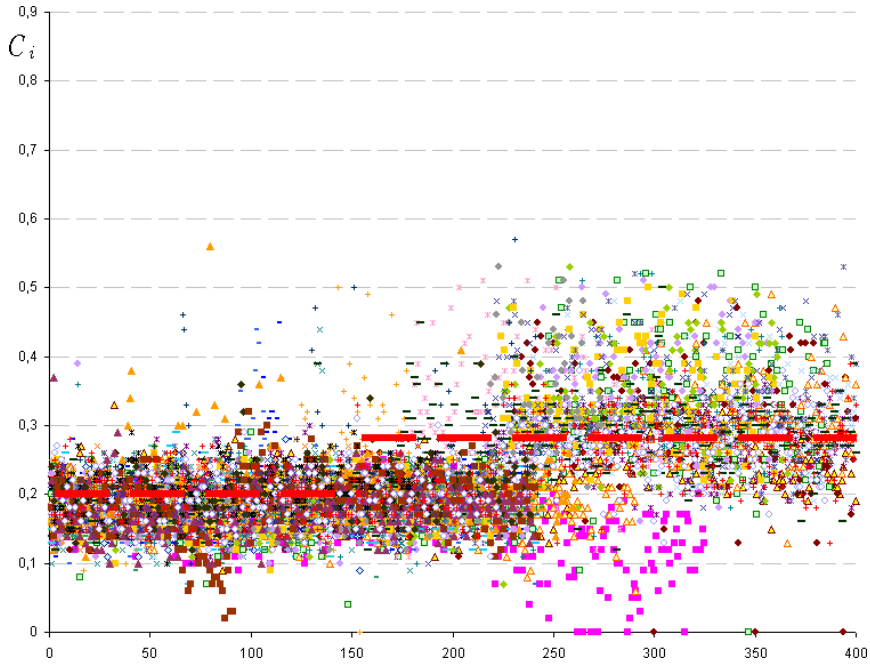


Figura 5.18 – *Distribuições, para as serinoproteases , dos índices de clusterização. O eixo das abscissas mostra os índices dos resíduos (abscissas $\leq$ 250), e dos demais grupos químicos (abscissas $>$ 250) das proteínas. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média está em vermelho. O aglomerado de pontos à esquerda mostra o comportamento dos resíduos das proteínas, o mais à direita mostra o comportamento das moléculas de água estrutural.*

as globinas. Da mesma forma, o coeficiente médio de aglomeração para as serinoproteases estudadas ficou $\langle C \rangle = 0,22$, com desvio padrão em $\sigma = 0,06$. A tabela 5.8 mostra coeficiente de aglomeração para as serinoproteases estudadas.

Tabela 5.8 – *Coeficiente médio de aglomeração médio para as serinoproteases em estudo.*

PDBID	$\langle C \rangle$	σ	PDBID	$\langle C \rangle$	σ
1ANE	0,16	0,07	1AQ7	0,19	0,09
1ARB	0,22	0,10	1BEF	0,18	0,08
1BIO	0,24	0,11	1BRU	0,24	0,11
1CA8	0,23	0,10	1CGH	0,20	0,09
1DUA	0,19	0,08	1DY9	0,19	0,08
1EAX	0,24	0,11	1ETR	0,23	0,10
1G2L	0,24	0,11	1GVK	0,26	0,11
1GVZ	0,23	0,10	1H8I	0,22	0,10
1HPG	0,25	0,11	1HXE	0,23	0,10
1K2I	0,21	0,09	1KLI	0,12	0,05
1NN6	0,20	0,09	1NTP	0,18	0,08

Continua ...

Tabela 5.8 *– (Continuação)*

PDBID	$\langle C \rangle$	σ	PDBID	$\langle C \rangle$	σ
1OP0	0,24	0,11	1OP2	0,24	0,11
1OWE	0,20	0,09	1P3C	0,20	0,09
1PPZ	0,22	0,10	1PQ7	0,26	0,11
1QNJ	0,26	0,11	1QY6	0,22	0,10
1S0R	0,24	0,10	1S83	0,23	0,10
1SGP	0,23	0,10	1SGT	0,21	0,09
1SQT	0,18	0,08	1SSX	0,26	0,12
1TAW	0,20	0,09	1TON	0,18	0,08
1VR1	0,18	0,08	1VZQ	0,23	0,10
1WCZ	0,25	0,11	1Y8T	0,22	0,10
2SFA	0,21	0,09	2SGA	0,23	0,10
2TBS	0,23	0,10	5PTP	0,23	0,10

Média dos valores é 0,22 ± 0,06.

Os resultados obtidos para as serinoproteases leva às mesmas considerações e conclusões feitas para as globinas, quando comparados em relação aos demais índices de aglomeração apresentados na literatura [Newman (2003c)], para outras redes encontradas no mundo real.

Os índices de distância entre átomos, para as serinoproteases estudadas, são apresentados na figura 5.19. Seguindo a mesma convenção adotada neste trabalho, nas figuras cada uma das serinoproteases está representada por um grupo distinto de símbolos e cores. O valor que corresponde à cada um dos resíduos de aminoácido de cada proteína é marcado no eixo horizontal - i. De forma análoga, os valores referentes aos índices de distância média entre átomos estão marcados no eixo vertical - L_i. As linhas horizontais em vermelho, representam o valor médio dos índices para as proteínas analisadas.

De forma similar ao observado para os índices de aglomeração, percebe-se, nestes gráficos, a existência de duas regiões distintas para a distribuição dos valores dos índices de distância. Mais à esquerda do gráfico 5.19, entre as abscissas $1 \leq i \leq 262$ e coordenadas entre $1,5 \leq l_i \leq 30,0$ ($\langle L \rangle = 11,16$, $\sigma = 3,76$), estão representados os índices relativos aos resíduos que formam a estrutura primária das serinoproteases . Mais à direita, entre as abscissas $214 \leq i \leq 400$, estão representados os índices das moléculas que não fazem parte da estrutura primária das proteínas (água, grupos prostéticos, etc). A tabela 5.9 mostra coeficiente de aglomeração para as serinoproteases estudadas.

Tabela 5.9 *– Coeficiente médio de distância entre átomos para as serinoproteases em estudo*

PDBID	$\langle L \rangle$	σ	PDBID	$\langle L \rangle$	σ
1ANE	12,00	5,30	1AQ7	11,31	5,00

Continua ...

Tabela 5.9 *– (Continuação)*

PDBID	$\langle L \rangle$	σ	PDBID	$\langle L \rangle$	σ
1ARB	11,64	5,14	1BEF	11,62	5,13
1BIO	11,61	5,13	1BRU	15,13	6,69
1CA8	14,46	6,39	1CGH	10,90	4,82
1DUA	11,08	4,90	1DY9	16,26	7,18
1EAX	13,81	6,10	1ETR	15,07	6,66
1G2L	16,25	7,18	1GVK	15,47	6,83
1GVZ	12,84	5,67	1H8I	13,43	5,93
1HPG	14,05	6,21	1HXE	14,67	6,48
1K2I	7,52	3,32	1KLI	8,17	3,61
1NN6	9,04	3,99	1NTP	9,42	4,16
1OP0	14,02	6,19	1OP2	12,96	5,72
1OWE	11,90	5,25	1P3C	11,75	5,19
1PPZ	11,61	5,13	1PQ7	11,61	5,13
1QNJ	6,25	2,76	1QY6	6,63	2,93
1S0R	7,50	3,31	1S83	8,17	3,61
1SGP	9,52	4,21	1SGT	10,58	4,67
1SQT	10,87	4,80	1SSX	11,35	5,01
1TAW	11,89	5,25	1TON	12,98	5,73
1VR1	9,42	4,16	1VZQ	14,09	6,22
1WCZ	5,87	2,59	1Y8T	18,14	8,01
2SFA	6,40	2,83	2SGA	7,20	3,18
2TBS	10,97	4,85	5PTP	10,96	4,84

Média dos valores é 11,16 ± 3,76.

Os resultados apresentados na figura 5.19 mostram que, similar ao comportamento observado nas globinas, os índices de distância média para os resíduos de aminoácido das serinoproteases apresentam grande variabilidade. Esta variabilidade sugere que, também para este grupo de proteínas, existe assimetria na capacidade de percolação de sinais entre os resíduos. Isto sugere que algumas porções da proteína devem responder mais rapidamente às perturbações induzidas em uma região qualquer da estrutura, enquanto outras devem responder, ao mesmo sinal, com um retardo maior.

Contudo, observando os pontos compreendidos entre as abscissas $1 \leq i \leq 262$ (os quais são relativos aos resíduos que compõem a estrutura primária da proteína), vê-se que o perfil de distribuição sugere que as serinoproteases mostram-se mais homogêneas quanto as propriedades de percolação de sinais em suas estruturas. Tal evidência sugere que, diferente das globinas, as serinoproteases apresentam um tempo de resposta bem similar, quando tem suas estruturas perturbadas por algum sinal.

Ainda com relação às serinoproteases , vale relembrar que os índices relativos às distâncias

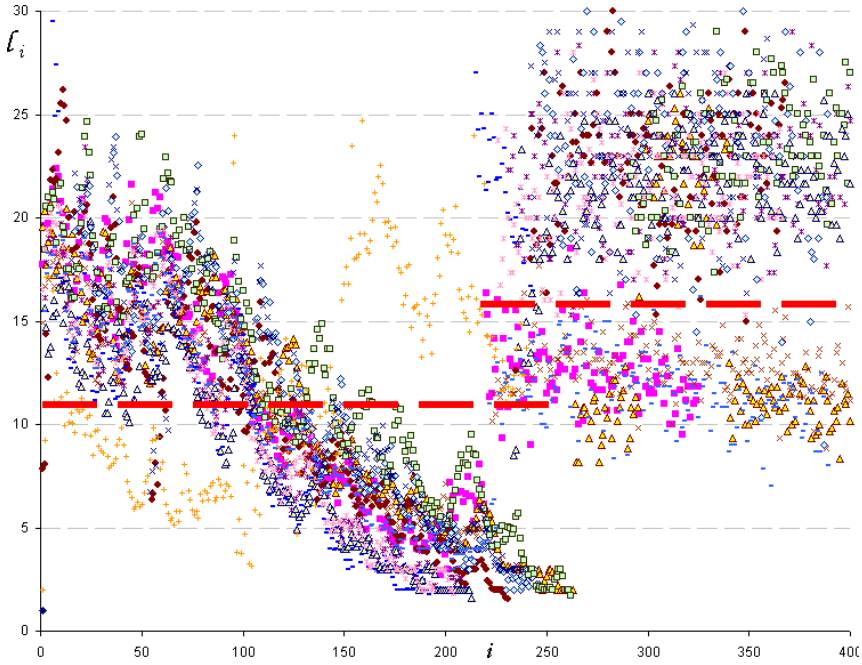


Figura 5.19 *– Distribuições, para as serinoproteases , dos índices de distância média (em passos) entre átomos, por resíduo na estrutura primária. O eixo das abscissas mostra os índices dos resíduos (abscissas $\leq$ 250), e dos demais grupos químicos (abscissas $>$ 250) das proteínas. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média está em vermelho. O aglomerado de pontos à esquerda mostra o comportamento dos resíduos das proteínas, o mais à direita mostra o comportamento das moléculas de água estrutural.*

entre nodos de uma rede exercem uma influência importante na emergência do efeito de "*small-world*" nas redes[4]. Este efeito tem implicações para a dinâmica dos processos que são influenciados pelos aspectos estruturais das rede. Para as proteínas, a propagação de estímulos localizados (como a acomodação de ligantes em sítios ativos), por toda a estrutura da proteína, pode implicar em alterações de toda a sua topologia e de seu comportamento. Ressalta-se que, quanto mais pronunciado for o efeito "*small-world*" na estrutura de uma proteína, mais rápida deve ser a propagação de perturbações por toda sua estrutura e mais rápida deverá ser sua resposta a estas perturbações.

Assim, outro ponto de interesse é avaliar em que grau as estruturas destas proteínas aderem aos modelos teóricos de redes complexas. Como apresentado na seção 3.3.1, uma rede apresenta propriedades que a definem como apresentando o fenômeno de "*small-world*", se para qualquer nodo desta rede, o número de nodos N, presentes dentro de um raio r crescer exponencialmente com r (ou seja $N \propto e^r$), e o valor da distância média entre nodos crescer logaritimicamente (ou a taxas menores) com N (ou seja $L \propto \log(N)$) [Newman (2003c)].

Desta forma, buscou-se avaliar para as serinoproteases , como o número de átomos presentes dentro de uma esfera com raio variável r com centro em um átomo qualquer, evolui à medida que este raio varia. Da mesma forma, avaliou-se como a distância média entre os

[4]Estes aspectos são abordados em maiores detalhes na seção 3.3.1

átomos varia em função deste raio. Os resultados obtidos estão apresentados nos gráficos das figuras 5.20 e 5.21.

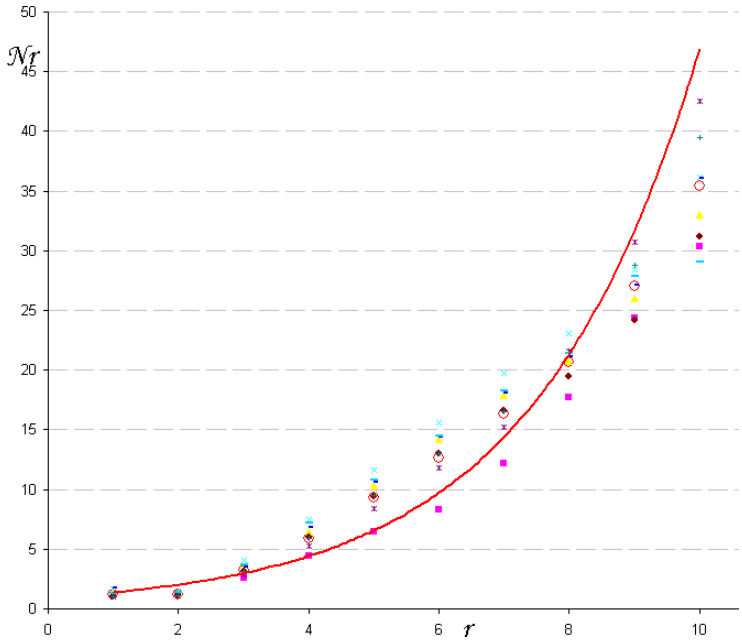


Figura 5.20 – *Distribuições, para as serinoproteases , da evolução do número de átomos $N(r)$, presentes dentro de um raio r em função do crescimento deste raio. Observa-se o crescimento exponencial de $N(r)$. Para as serinoproteases analisadas, $N(r)$ ajusta-se a uma curva exponencial $N(r) = 0,92e^{0,39r}$ com $R^2 = 0,95$.*

Os procedimentos de cálculo, que levaram aos resultados apresentados nestas figuras, são apresentados na seção 3.3.2. Os valores médios relativos ao número de átomos presentes em uma esfera de raio r $(N(r))$, para as serinoproteases estão apresentados na figura 5.20. A curva de tendência média apresentada no gráfico 5.20 revela que $N(r)$ varia com r segundo uma exponencial, de forma similar àquele apresentado pelas globinas. Este comportamento $(N(r) \propto \beta e^{\alpha r})$, ajusta-se ao comportamento previsto, para este parâmetro, do modelo de Watts [Watts e Strogatz (1998)], com $\beta = 0,92$ e $\alpha = 0,39$ e $R^2 = 0,95$.

Por outro lado, os valores médios de $L(r)$[5]para as serinoproteases estão apresentados na figura 5.21. Para este parâmetro, a curva de tendência média apresentada no gráfico 5.17 revela que $L(r)$ varia com r também segundo uma curva logarítmica ($i.e$ $N(r) \propto \beta \ln(r) + \gamma$). Também para as serinoproteases , percebe-se que o comportamento deste parâmetro ajusta-se ao previsto para o modelo de Watts [Watts e Strogatz (1998)], com $\beta = 1,69$ e $\gamma = 0,67$ e $R^2 = 0,98$.

A leitura destes resultados, permite afirmar que o comportamento dos índices de aglomeração e distância entre átomos, também para as serinoproteases , apresenta um comportamento típico das estruturas que apresentam o fenômeno de "*small-world*". Pode-se infe-

[5]Distância média (em passos) entre um átomo qualquer i a todos os $N(i,r)$ átomos compreendidos em uma esfera com centro em i e de raio r

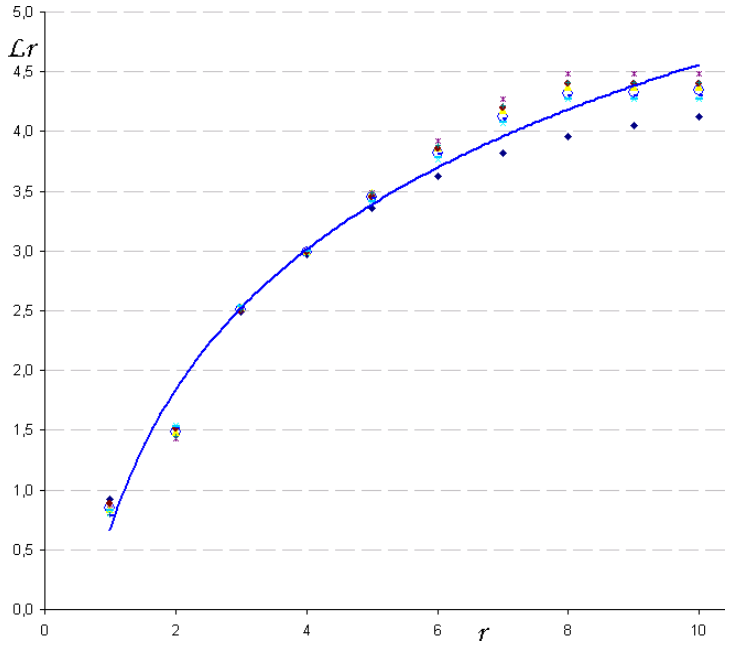


Figura 5.21 – *Distribuições normalizadas, para as serinoproteases , das evolução valor da distância média entre átomos $L(r)$, presentes dentro de um raio r em função do crescimento deste raio. Observa-se o crescimento logarítmico de $L(r)$. Para as serinoproteases analisadas, $L(r)$ ajusta-se a uma curva logarítmica $L(r) = 1,69\ln(r) + 0,67$ com $R^2 = 0,98$.*

rir, também neste caso, que a estrutura destas proteínas deve apresentar comportamentos análogos àqueles estimados para os modelos de redes complexas "*small-world*".

5.3 Análise Espectral das Estruturas das Proteínas

Tal como apresentado na seção 3.5, informações acerca dos atributos das redes como os relacionados aos processos de difusão e arranjo hierárquico dos vértices podem ser obtidos da aplicação da análise espectral de redes complexas. Nesta seção são apresentados os resultados obtidos da aplicação desta análise para o caso das globinas e serinoproteases.

Para proceder a análise espectral das redes de interações não-covalentes subjacentes às proteínas, os algoritmos apresentados na seção 3.5 foram implementados e avaliados quanto a qualidade dos resultados apresentados. Para a avaliação da qualidade das implementações foram analisados 20 grafos randômicos construídos segundo o modelo de Erdös e Rényi [Erdös e Rényi (1959)], 20 grafos construídos segundo o modelo "*small-world*" de Newman e Watts [Newman e Watts (1999)] e 20 grafos construídos segundo o modelo de Albert e Barabási [Albert et al. (1999), Albert e Barabasi (1999)]. Inicialmente estes grafos foram construídos fixando o número de vértices – $N_v = 2.900$[6], e com probabilidade – $p(k_i) = 50\%$. Para a construção dos grafos segundo o modelo de "*small-world*" de Newman e Watts [Newman e Watts (1999)], o número de contatos inicial de cada vértice foi fixado em – $k_0 = 2$. Tal conjunto de grafos permite que os padrões encontrados para as redes presentes nas proteínas possam ser comparadas com os padrões apresentados pelos modelos de referência estudados na literatura. Os espectros dos modelos teóricos construídos com base nestes parâmetros são apresentados na figura 5.22.

Da observação da figura 5.22 percebe-se que o algoritmo implementado replica corretamente os resultados apresentados por Farkas [Farkas et al. (2001)] (ver figura 3.15). O semi-círculo verde mostra a curva construída segundo o modelo proposto por Wigner [Wigner (1955)], que representa o caso de um sistema totalmente aleatório. O semi-círculo vermelho, mostra a curva de distribuição de densidades para o modelo de Erdös e Rényi, enquanto o semi-círculo azul mostra o espectro para o modelo de Watts e Strogatz – "*small-world*". Para estes dois últimos casos é possível perceber a existência do pico central, próximo ao eixo λ_0. Ao mesmo tempo a distribuição do espectro tende para a distribuição de Wigner, quando $N \rightarrow \infty$, o que está de acordo com os resultados apresentados por Farkas [Farkas et al. (2001)]. Por outro lado, o espectro relativo ao modelo de Barabási-Albert (curva laranja no formato triangular) apresenta as mesmas tendências relatadas por Farkas, apresentando um segmento que decai seguindo uma lei de potência – $\rho(\lambda) = \alpha\lambda^{-\gamma}$ onde $\alpha = 0,044$ e $\gamma = 3,41$ com $R^2 = 0,92$ (linha vermelha à direita).

Validada a implementação do algoritmo de cálculo de densidade espectral, o mesmo cálculo foi feito agora para os mesmos modelos de referência agora com os parâmetros estruturais apresentados tanto pelas globinas quanto pelas serinoproteases. No caso das globinas, tem-se o número médio de átomos por proteína – $\langle N_A \rangle = 2334$, com número médio de interações por átomo – $\langle k_i \rangle = 13,76$, implicando em uma probabilidade média – $\langle p_i \rangle = 0,006$. Os resultados desta análise são apresentados na figura 5.23. Para o caso das serinoproteases , tem-se o número médio de átomos por proteína – $\langle N_A \rangle = 3600$, com número médio de interações por

[6]Este valor foi adotado, considerando que este é o número médio de átomos das proteínas em estudo.

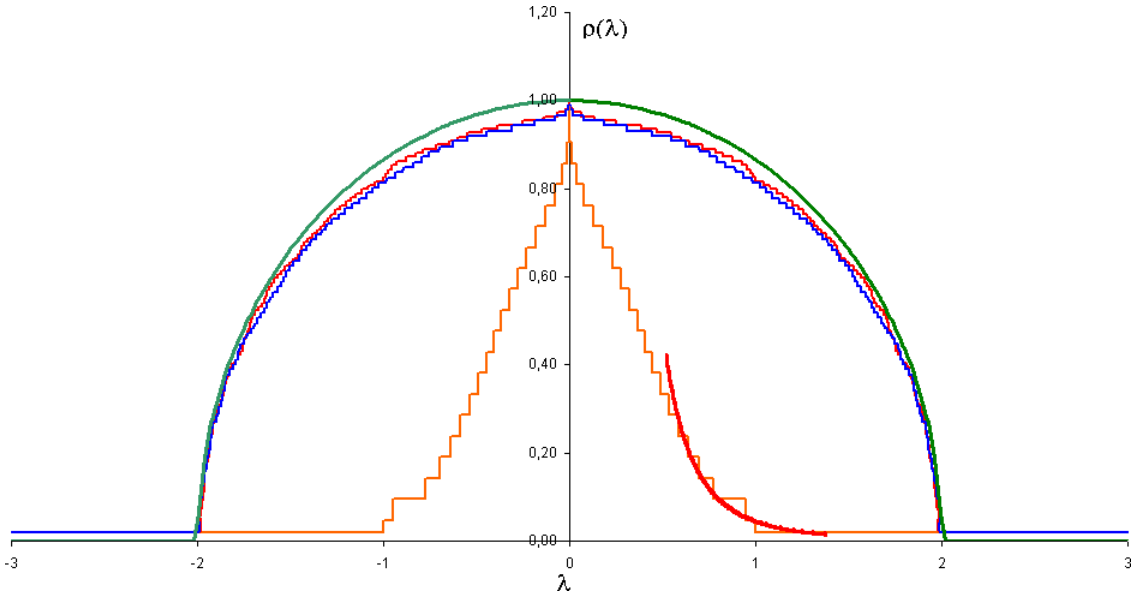


Figura 5.22 *– Distribuição de densidade – $\rho(\lambda)$, de três modelos: (1) Rede de Erdös e Rényi – semi-círculo vermelho; (2) Modelo de Newman e Watts ("*small-world*") com menor grau $k_0 = 8$ – semi-círculo azul; (3) Modelo de Barabási-Albert (livre de escala) – linha triangular laranja. A distribuição de densidade para o modelo teórico de Wigner, que representa um sistema de partículas aleatoriamente distribuídas em um dado volume, é representada pelo semi-círculo verde. Todos os modelos foram gerados considerando $p(k_i) = 50\%$. A linha vermelha à direita mostra o segmento do espectro do modelo Barabási-Albert que apresenta decaimento seguindo uma lei de potência.*

átomo – $\langle k_i \rangle = 11,56$, implicando em uma probabilidade média – $\langle p_i \rangle = 0,003$. Os resultados desta análise são apresentados na figura 5.24.

Os resultados apresentados nas figuras 5.23 e 5.24, quando comparados com os resultados da figura 5.22 permite perceber que os modelos teóricos, quando calculados com os parâmetros típicos tanto das globinas como das serinoproteases, apresentam um comportamento diferente daquele obtido anteriormente. Em todos os casos, o comportamento dos modelos afasta-se do comportamento apresentado por um sistema totalmente aleatório, que é representado pelo modelo de Wigner [Wigner (1955)]. De um ponto de vista geral, a própria conectividade química da estrutura primária das proteínas garante que a distribuição dos átomos neste sistema não pode ser aleatório.

Nestas mesmas condições, a curva de distribuição de densidades para o modelo de Erdös e Rényi (curva em vermelho), apesar de ainda apresentar uma grande dispersão no espectro, mostra uma rarefação na densidade dos laços internos, mostrando que a rede tende a ficar esparsa. Por sua vez, o espectro do modelo "*small-world*" (curva azul) degenera para uma distribuição similar àquela apresentada pelo modelo livre de escala (curva laranja). Nestas condições, percebe-se que no modelo "*small-world*" cessa a predominância topológica dos laços locais e emerge uma estrutura com características mais hierarquizadas. Ao mesmo tempo, observa-se que a topologia do modelo livre de escala não apresenta alterações topológicas apreciáveis, ressaltando contudo a tendência de decaimento em lei de potência –

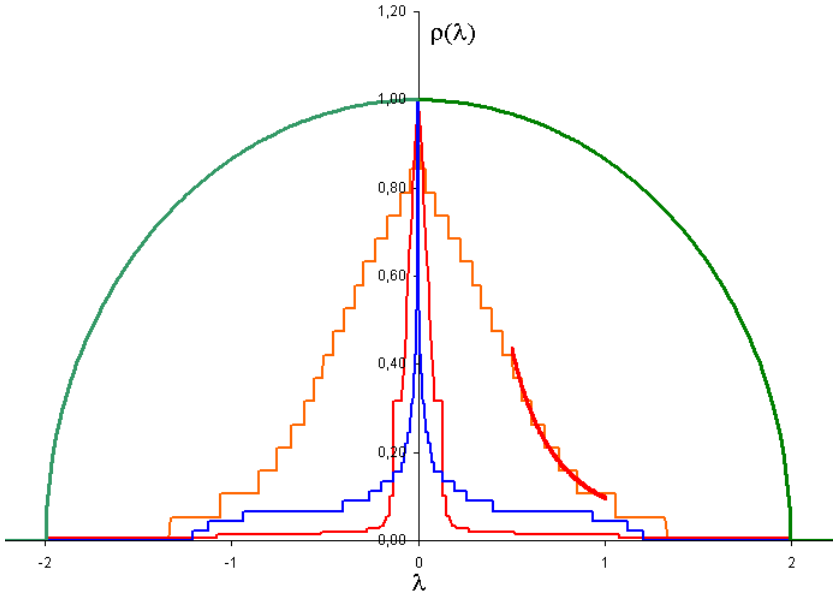


Figura 5.23 *– Distribuição de densidade – $\rho(\lambda)$, de três modelos calculados com os parâmetros médios típicos das globinas: (1) Rede de Erdös e Rényi – curva em vermelho; (2) Modelo de Newman e Watts ("small-world") com grau $k_0 = 7$ – curva azul; (3) Modelo de Barabási-Albert (livre de escala) – curva laranja. A densidade para o modelo de Wigner é representada pelo semi-círculo verde. Todos os modelos foram gerados considerando $p(k_i) = 0,06\%$. A linha vermelha à direita mostra o segmento do espectro do modelo Barabási-Albert que apresenta decaimento seguindo uma lei de potência.*

$\rho(\lambda) = \alpha\lambda^{-\gamma}$ do segmento onde $[0,5 \lesssim \lambda \lesssim 1,0]$, onde as globinas (figura 5.23) apresentam $\alpha = 0,116$ e $\gamma = 2,246$ com $R^2 = 0,96$ (linha vermelha à direita); e as serinoproteases (figura 5.24) apresentam $\alpha = 0,079$ e $\gamma = 2,697$ com $R^2 = 0,96$ (linha vermelha à direita). Percebe-se ainda a existência do pico central, próximo ao eixo λ_0. Segundo Farkas [Farkas et al. (2001)], tal pico indica a existência de um grande conjunto de vértices cujos respectivos vetores singulares compartilham o mesmo módulo, sendo assim indiferenciáveis do ponto de vista de importância topológica para a rede.

Após a obtenção dos modelos de referência, foram feitas as análises dos espectros tanto das globinas quanto das serinoproteases, conforme os métodos apresentados na seção 3.5. Entretanto, antes de apresentar e discutir os resultados relativos às proteínas, cabe ressaltar que os gráficos de distribuição espectral apresentados até aqui, mostram-se simétricos em relação ao eixo $\lambda = 0$, de forma a tornar fácil a comparação dos resultados apresentados com os padrões apresentados na literatura. Contudo, uma vez que os valores de λ derivados da decomposição dos valores singulares - SVD são sempre positivos ou zero (ver seção 3.5.1), e que a função $\delta(\lambda_i - \lambda_j) \geq 0$ (eq. 3.8), os gráficos seguintes serão apresentados somente para valores de $\lambda \geq 0$ e $\rho(\lambda) \geq 0$.

Com relação à análise das proteínas, inicialmente foram feitos os cálculos para a obtenção dos valores singulares relativos às globinas e serinoproteases estudadas. As distribuições das freqüências de ocorrência dos valores singulares das globinas são apresentadas na figura 5.25.

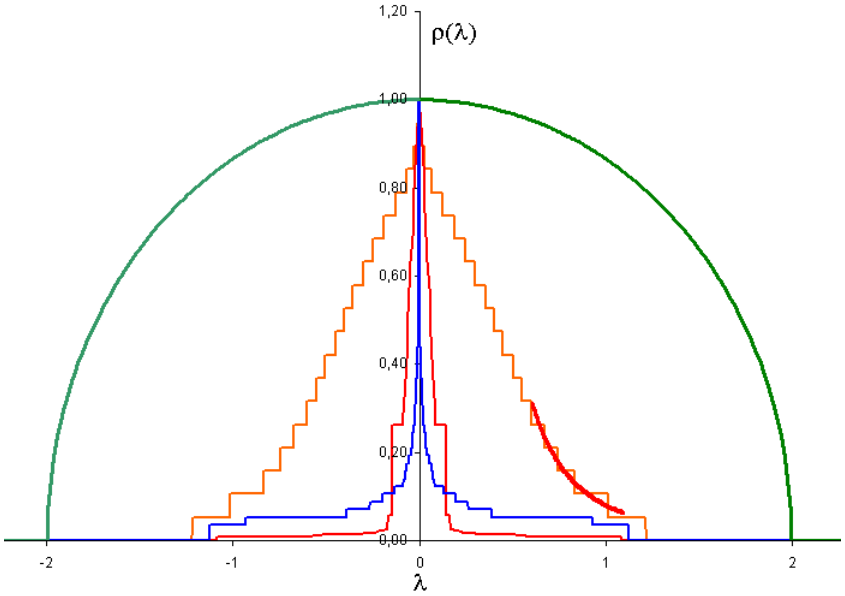


Figura 5.24 *– Distribuição de densidade – $\rho(\lambda)$, de três modelos calculados com os parâmetros médios típicos das serinoproteases: (1) Rede de Erdös e Rényi – curva em vermelho; (2) Modelo de Newman e Watts ("*small-world*") com grau $k_0 = 7$ – curva azul; (3) Modelo de Barabási-Albert (livre de escala) – curva laranja. A densidade para o modelo de Wigner é representada pelo semi-círculo verde. Todos os modelos foram gerados considerando $p(k_i) = 0,03\%$. A linha vermelha à direita mostra o segmento do espectro do modelo Barabási-Albert que apresenta decaimento seguindo uma lei de potência.*

O gráfico da figura 5.25 mostra toda a distribuição de freqüências médias, normalizadas, dos valores singulares $p(\lambda)$ para as globinas. Neste gráfico, os valores singulares foram calculados com base na matriz de adjacências não ponderadas representativas das globinas em estudo. Os valores apresentados foram normalizados em função do maior valor de freqüência. Nesta escala percebe-se que a freqüência de ocorrência dos valores singulares é degenerada, apresentando praticamente dois grupos distintos de valores singulares. O primeiro grande grupo é composto por todos os valores singulares próximos à origem do gráfico. A grande freqüência de valores nesta região mostra que a maioria dos átomos (e por conseqüência os respectivos resíduos), do ponto de vista dos atributos de percolação de informação na estrutura das globinas, não exercem uma função significativa para a proteína.

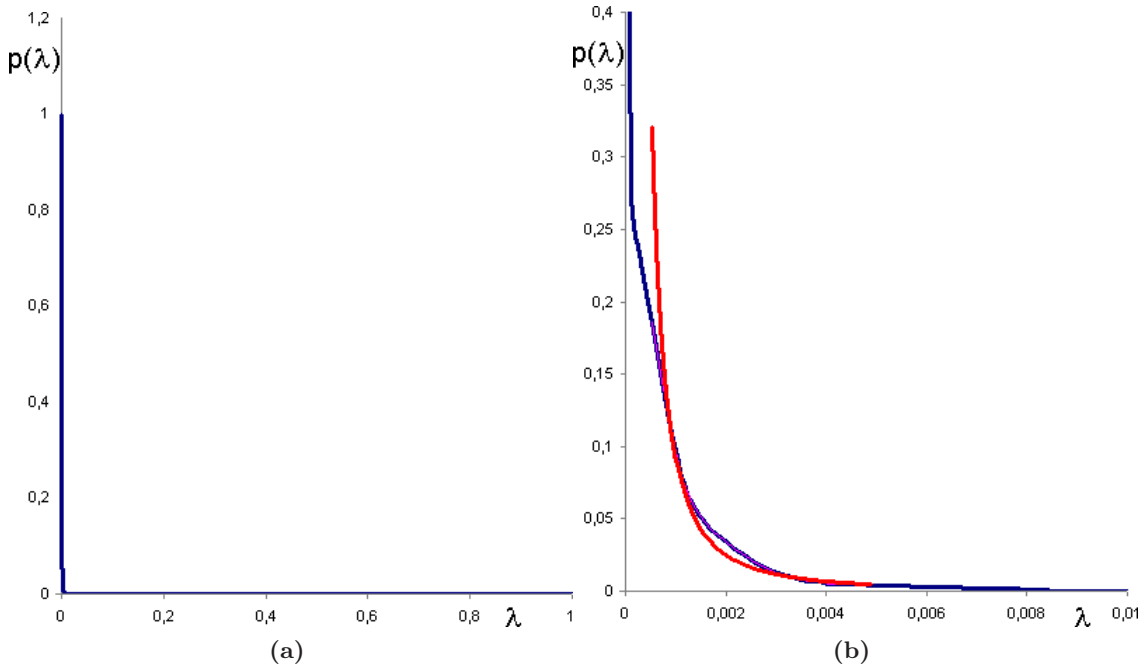


Figura 5.25 *– (a) – Distribuição de freqüências da média dos valores singulares – $p(\lambda)$, para as globinas calculados com base na matriz de adjacências. (b) – Detalhe da distribuição de freqüências – $p(\lambda)$, apresenta decaimento conforme uma lei de potência – $p(\lambda) = \alpha\lambda^{-\gamma}$, representada pela linha vermelha, com parâmetros $\alpha = 2,0 \times 10^{-7}$, $\gamma = 1,89$ com $R^2 = 0,97$.*

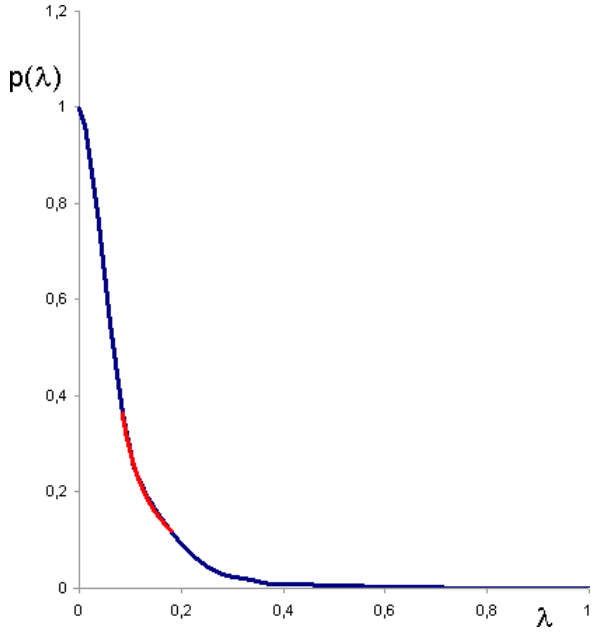


Figura 5.26 *– Detalhe da distribuição de freqüências da média dos valores singulares – $p(\lambda)$, para as serinoproteases calculados com base na matriz de adjacências. O decaimento dos valores singulares médios das proteínas segue uma lei de potência – $p(\lambda) = \alpha\lambda^{-\gamma}$, representada pela linha vermelha, com parâmetros $\alpha = 0,01$, $\gamma = -1,46$ com $R^2 = 0,99$.*

Observa-se ainda na figura 5.25(b) que um grupo pequeno de átomos/resíduos possui valores singulares com grande módulo. O módulo destes valores indica que tais átomos/resíduos exercem uma função significativa nestas estruturas. Ao mesmo tempo, observa-se um reduzido grupo de valores singulares cujo decaimento segue uma lei de potência, o qual é apresentado em detalhes na figura 5.25(b). Tais padrões estão de acordo com os apresentados em Newman [Newman (2003a)] e Brede [Brede e Sinha (2005)]. Um comportamento similar é apresentado pelas serinoproteases , tal como atesta a figura 5.26.

Ressalta-se que os gráficos apresentados da figura 5.25 à 5.26, foram obtidas a partir das matrizes de adjacências não ponderadas representativas das proteínas em estudo. Com base nos padrões apresentados por estes gráficos, constata-se que estes resultados demonstram que a geometria da rede, formada pelas interações não-covalentes entre os átomos destas proteínas, é compatível com os conceitos apresentados na literatura [Newman (2003a), Brede e Sinha (2005)] para as redes com grande discrepância entre os valores de λ_1, λ_2 e λ_N.

Como apresentado na seção 5.2.1, cada uma das interações não covalente existentes entre os átomos das proteínas estudadas, possui um valor associado de energia potencial, medida em $[kcal/mol]$. Este valor será usado como peso das arestas na composição das matrizes de adjacências ponderadas, representativas das proteínas em estudo. Sabe-se que redes com comportamento similar ao das redes livre de escala, apresentam uma distribuição dos valores singulares com formato triangular [Newman (2003a) ,Brede e Sinha (2005)]. Em Brede [Brede e Sinha (2005)], observa-se que quando o cálculo dos valores singulares é feito com base em matrizes ponderadas, o formato triangular da distribuição destes valores torna-se ainda mais aguçada. Tal padrão é também observado para as proteínas estudadas, como mostram os gráficos da figura 5.27.

A figura 5.27 mostra que o perfil das curvas de distribuição dos valores singulares, relativos às matrizes de adjacências ponderadas, tanto para globinas quanto para serinoproteases são bem diferentes das curvas apresentadas nas figuras 5.25 a 5.26. O padrão apresentado, pelos valores singulares derivados das matrizes ponderadas, mostra que a relevância dos átomos/resíduos das proteínas para os processos de percolação de informação na estrutura das proteínas sofre significativa influência das propriedades das interações não-covalentes. Ainda assim, é possível perceber na figura 5.27, que o decaimento das curvas continua seguindo um padrão de lei de potência, estando de acordo com as previsões teóricas apresentadas na seção 3.5.4.

A análise espectral dos valores singulares das matrizes ponderadas pode ainda revelar outros padrões bastante informativos sobre os aspectos dinâmicos dependentes da estrutura tridimensional das proteínas em análise. Com base nos métodos apresentados na seção 3.5.4, esta análise foi feita e os resultados são apresentados na figura 5.28.

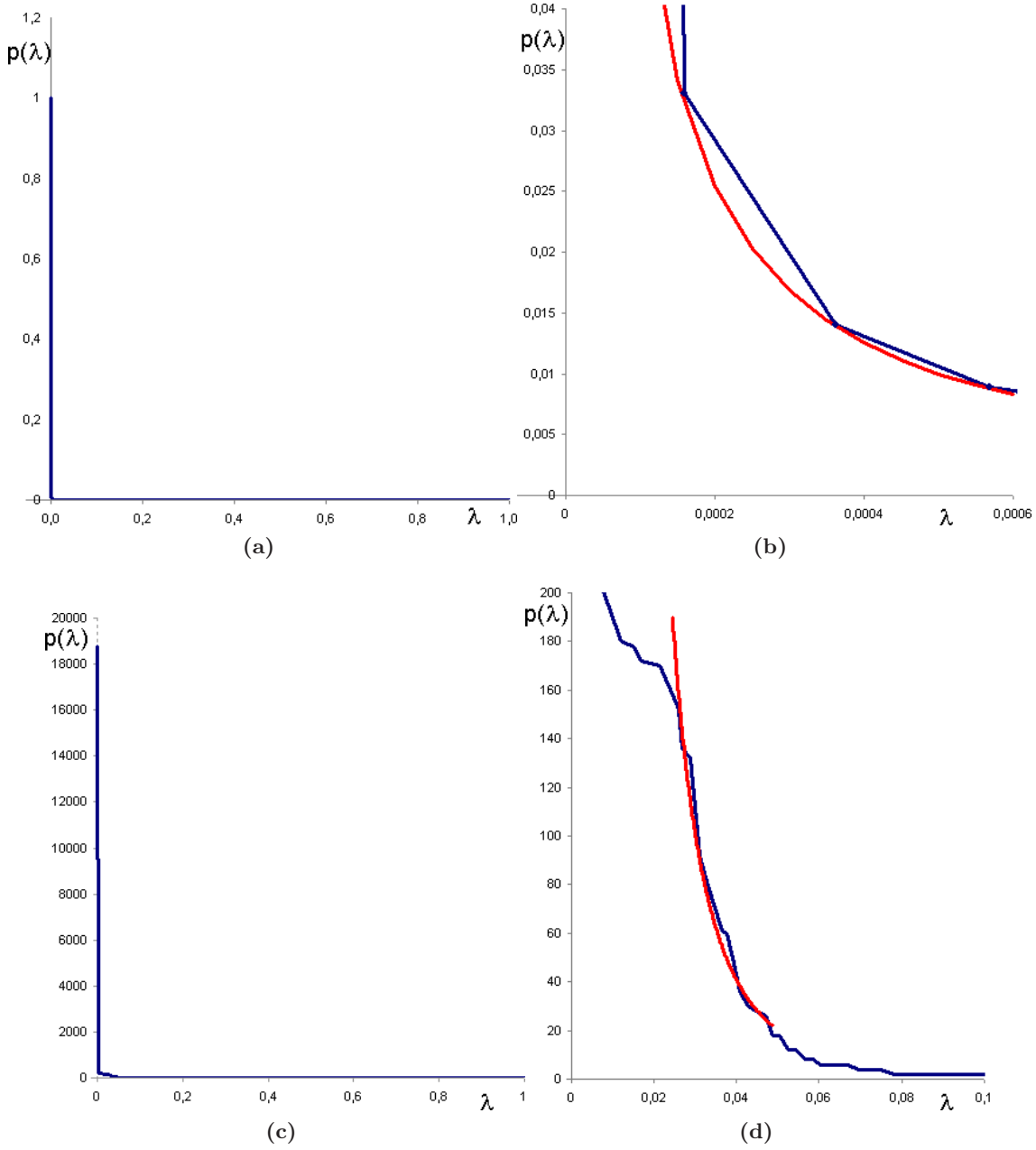


Figura 5.27 *– Distribuição de freqüências da média dos valores singulares – $p(\lambda)$, para as globinas (a) (b) e serinoproteases (c) (d), calculados com base na matriz de interações não-covalentes entre átomos, ponderadas pelas respectivas energias potenciais. Em (b) o decaimento da distribuição dos valores singulares – λ, para as globinas segue uma lei de potência – $p(\lambda) = \alpha\lambda^{-\gamma}$, representado pela linha vermelha, com parâmetros $\alpha = 4,24 \times 10^{-6}$, $\gamma = 1,02$ com $R^2 = 0,96$. Em (d) o decaimento da distribuição de λ, para as serinoproteases , representado pela linha vermelha, apresenta parâmetros $\alpha = 1,7 \times 10^{-3}$, $\gamma = 3,14$ com $R^2 = 0,97$.*

A leitura dos gráficos da figura 5.28 mostra que, para todas as proteínas analisadas, a quase totalidade dos átomos destas não apresenta diferença entre os respectivos valores singulares, enquanto um pequeno número deles apresenta valores singulares bem distintos dos demais. Tal distribuição permite inferir que existe um pequeno número de átomos/resíduos nestas proteínas, que determinam os processos de percolação (ou difusão) de impulsos através da estrutura das mesmas.

A existência de um pico central nestas distribuições indica, ao mesmo tempo, que a percolação de impulsos pela estrutura das proteínas deve ocorrer seguindo uma estrutura organizada de forma hierárquica, seguindo um percurso similar a uma "árvore". Assim os átomos/resíduos associados aos maiores valores singulares devem atuar como elementos constituintes da raiz. Diante disto pode-se inferir que o número de passos dos átomos constituintes desta "raiz" até os demais átomos seria o mínimo, número este representado pelo diâmetro médio em passos – $\langle L \rangle$, apresentado por cada uma das famílias topológicas estudadas: globinas – 9,16 ± 4,05 (tabela 5.5); serinoproteases – 11,16 ± 3,76 (tabela 5.9).

Os resultados derivados da análise espectral, dos valores singulares relativos às matrizes de adjacências ponderadas, tanto das globinas quanto das serinoproteases, podem ser comparados com os conceitos relativos à análise espectral de redes complexas, apresentados nas páginas 52 a 58.

Os perfis de distribuição dos espectros obtidos neste trabalho, indicam que as proteínas apresentam uma rede subjacente com forte caráter associativo. O pronunciado perfil dos espectros indica que a densidade de interações não covalente entre átomos, necessária para a emergência do grupo-núcleo ("*core group*") das proteínas é proporcionalmente muito baixa [Newman (2002), Newman (2003b), Newman (2003a), Brede e Sinha (2005)]. Uma vez que as proteínas apresentam características fortemente associativas, é de se esperar que os átomos/resíduos mais energeticamente conectados estejam preferencialmente ligados a outros átomos/resíduos similares.

Para avaliar esta hipótese, foram observados os maiores componentes do primeiro vetor singular das globinas analisadas. Esta análise permite identificar quais os átomos/resíduos mais relevantes para a composição do primeiro vetor singular, o qual deve mostrar uma correlação positiva com o caráter associativo apresentado pelas redes [Newman (2002), Newman (2003b), Newman (2003a), Brede e Sinha (2005)] e, por conseqüência, pelas proteínas. A análise dos dados revela que, para o caso geral das globinas, os resíduos identificados que mais concentram átomos relevantes na análise espectral são: o grupo **HEME**; a histidina proximal; e a fenilalanina conservada ligada ao grupo Heme. Para o caso das serinoproteases estudadas, os resíduos mais relevantes identificados, tomando como referência a proteína PDBID 1ANE, foram: histidina 57; asparagina 102; serina 195; aspártico 143; cisteina 191; serina 214; cisteina 220 . Estes resultados são apresentados nas figuras 5.29 e 5.30.

Além dos resíduos da tríade catalítica (Ser 195, His 57, e Asp 102), os demais resíduos identificados, são citados na literatura [Nienaber et al. (2000), Krem e Cera (2001), Krem et al. (2002), Milgotina et al. (2003), Wangikar et al. (2003), Bobofchak et al. (2005)], pela relevância destes para a função das proteínas desta família.

Conforme apresentado nas figuras 5.29 e 5.30, foi possível identificar os resíduos que formam o grupo-núcleo tanto para a família das globinas, quanto para as serinoproteases. A constatação da existência de grupos núcleo na estrutura das proteínas tem implicações na interpretação dos processos dinâmicos. Com base nos conceitos apresentados em [Newman (2003a), Brede e Sinha (2005)], pode-se conjecturar que, nas proteínas, a existência destes grupos-núcleo faculta a rápida percolação de informação para toda a estrutura das proteínas. Constata-se que estes grupos são restritos e que a densidade de interações é grande. Ao mesmo tempo, os grupos núcleo não se estendem pela estrutura das proteínas, estando confinado ao núcleo da rede.

Como estes padrões são típicos de uma rede com elevada associabilidade, é de se esperar que também para estas proteínas, estes núcleos devam prover robustez às estruturas das proteínas, ao concentrar todos os resíduos estruturalmente vitais, em uma região restrita destas proteínas. Assim, é possível que este arranjo estrutural seja aquele que foi evolutivamente selecionado, capaz de compatibilizar a flexibilidade necessária para permitir a evolução destes grupos-núcleo, com a necessidade da manutenção da estabilidade estrutural das proteínas. Isto porque uma mutação destes resíduos não deve ser suficiente para comprometer a conectividade estrutural das proteínas. Pode-se conjecturar que estas mutações não devam comprometer significativamente a capacidade percolação de informação pela estrutura das proteínas.

Os valores singulares λ_1 e λ_2, mostram-se bem maiores que os demais valores de λ, presentes nos espectros das proteínas estudadas. Este comportamento dos valores singulares sugere que as proteínas devem apresentar forte tendência à instabilidade. Contudo, considerando que os modelos teóricos [Newman (2003a),Brede e Sinha (2005)] sejam plenamente aplicáveis ao caso das proteínas, estima-se que as perturbações ambientais devam surtir pouco efeito na percolação de informação através da estrutura das proteínas. Ao mesmo tempo, as proteínas devem ser capazes de dar respostas rápidas aos impulsos percebidos pelos grupos-núcleo, saindo com muita facilidade, de uma conformação para outra mais adequada ao novo contexto a que a proteína está exposta.

Estes achados mostram-se relevantes quando os problemas relacionados aos fenômenos alostéricos[7] apresentados pelas proteínas. Os fenômenos alostéricos exercem um papel fundamental em diferentes processos de sinalização celular. As perturbações causadas por fatores como a entrada de um ligante em um sítio funcional afetam outros sítios distantes, e desta forma regulam a afinidade e a atividade desta proteína. Apesar de métodos experimentais já terem mostrado alguns padrões associados à comunicação alostérica, o entendimento dos princípios gerais de transmissão de informação entre sítios funcionais distantes permanece como um desafio. Os indícios aqui identificados sugerem a existência de resíduos chaves, nas proteínas, responsáveis pela geração e transmissão de tais sinais, ao mesmo tempo em que a topologia destas redes a transmissão destes sinais deve ocorrer seguindo certos atalhos entre as diferentes regiões das estruturas.

[7] Alosteria é o fenômeno relacionado à mudança na forma e na atividade de uma proteína (especialmente nas enzimas) que decorre da combinação de outra molécula em um ponto quimicamente ativo de sua estrutura.

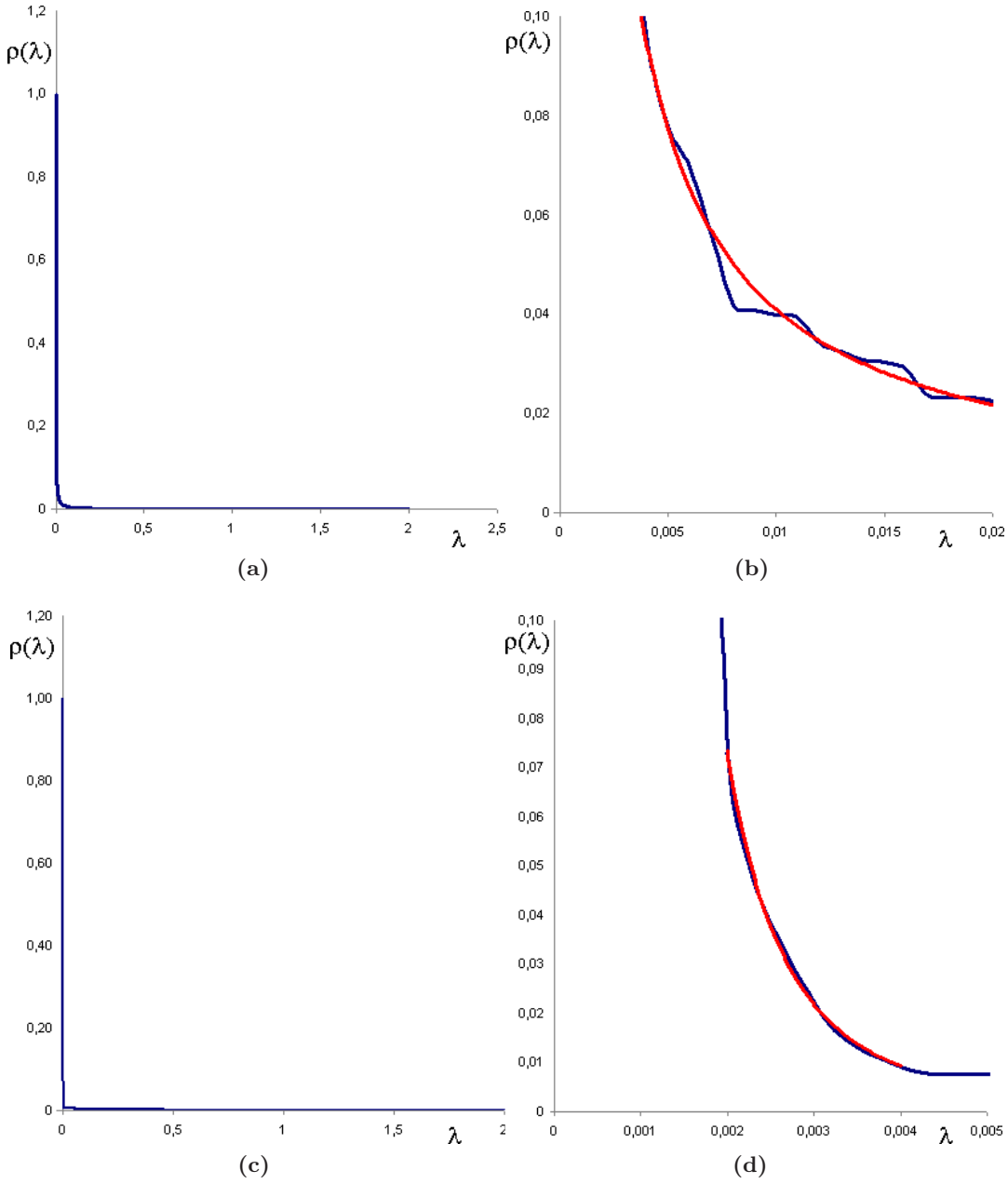


Figura 5.28 *– Distribuição média do espectro dos valores singulares – $\rho(\lambda)$, para as globinas (a) (b) e serinoproteases (c) (d), calculados com base na matriz de interações não-covalentes entre átomos, ponderadas pelas respectivas energias potenciais. Em (b) o decaimento da distribuição dos valores singulares – λ, para as globinas segue uma lei de potência – $p(\lambda) = \alpha\lambda^{-\gamma}$, representado pela linha vermelha, com parâmetros $\alpha = 6,0 \times 10^{-4}$, $\gamma = 1,01$ com $R^2 = 0,98$. Em (d) o decaimento da distribuição de λ, para as serinoproteases , representado pela linha vermelha, apresenta parâmetros $\alpha = 7,0 \times 10^{-10}$, $\gamma = 2,98$ com $R^2 = 0,99$.*

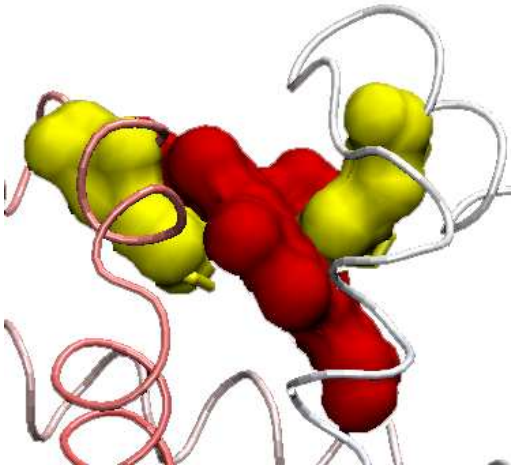
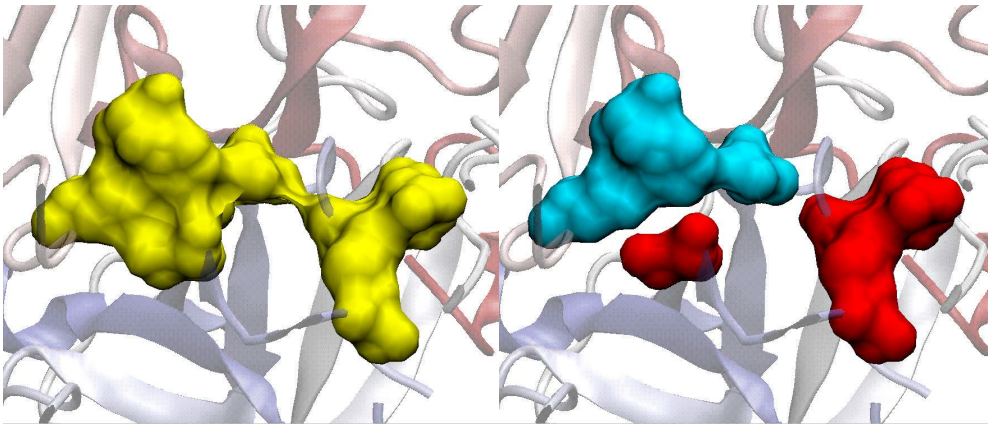


Figura 5.29 *– Resíduos pertencentes ao grupo-núcleo das globinas estudadas. O grupo* ***HEME*** *aparece em vermelho. A histidina proximal, está em amarelo à esquerda enquanto a fenilalanina está à direita.*



(a) **(b)**

Figura 5.30 *– Resíduos pertencentes ao grupo-núcleo das serinoproteases estudadas. (a) – Todos os resíduos do grupo-núcleo estão em amarelo. Dentre estes encontram-se os resíduos que formam a tríade catalítica das serinoproteases estudadas. É possível observar o encadeamento existente entre eles. (b) – Os resíduos da tríade catalítica são apresentados em azul. Os resíduos de cisteína e aspártico são apresentados em vermelho.*

5.4 Identificação dos Nodos Concentradores

Os parâmetros estruturais importantes para a caracterização das proteínas, vão além dos parâmetros já analisados. Assim sendo, faz-se necessária a tipificação de outros fatores relevantes da rede, e dentre estes o conjunto de resíduos que exercem o papel de vértices concentradores da estrutura das proteínas.

Avaliar a existência destes elementos críticos de uma rede, permite identificar quais são os vértices que apresentam maior relevância para a estabilidade estrutural e para os processos dinâmicos que aí ocorrem. Neste particular, um dos processos de maior interesse para o entendimento das proteínas é o da propagação de impulsos[8] através da rede formada pelo conjunto de interações não-covalentes subjacente á estrutura tridimensional destas proteínas.

Como apresentado na seção , os vértices críticos da rede são usualmente aqueles vértices que apresentam maior grau de conectividade ("*hubs*"), quando comparados aos demais vértices da rede. Tais vértices, são cruciais para a estabilidade da rede, apresentando maior relevância dentro do concerto sistêmico.

Visando a identificação destes átomos, outras análises foram feitas no conjunto de globinas estudadas, e os resultados obtidos são apresentados nas subseções seguintes. Para tanto, foi feita a identificação das interações não-covalentes (átomo a átomo) que são estabelecidas na estrutura tridimensional destas proteínas. Posteriormente, com base na identificação destas interações, foram também identificados aqueles resíduos com maior conectividade. Por seu turno, as interações não-covalentes foram caracterizadas com base na distância euclidiana entre os átomos e na energia potencial inerente às mesmas. Os conceitos e métodos adotados para a realização destas análises foram previamente apresentados e discutidos no capítulo 4.

5.4.1 Identificação dos Resíduos mais conectados nas globinas

A partir da identificação das ligações não-covalentes presentes na estrutura tridimensional das globinas em análise, foi feita a contagem do número de interações vinculadas à cada um dos átomos (e por conseguinte à cada um dos resíduos) presentes em cada uma das globinas estudadas. Os resultados da análise das interações, ponderadas pelos respectivos valores de energia, são apresentados no gráfico da figura 5.31.

Na figura 5.31, o decaimento da freqüência média dos valores energia $f(E_A)$ dos átomos apresenta um decaimento em lei de potência, tal como apresentado na seção 5.2.1. A observação da cauda do gráfico mostra a existência, na estrutura das proteínas, de um pequeno número de átomos fortemente conectados.

Os vértices de um grafo que apresentam alta conectividade (ou *strength*[9]) e baixa freqüência, apresentam a propriedade de permitir que aglomerados de vértices, topologicamente distantes, possam se comunicar. Em outras palavras, sem o auxílio das interações estabelecidas por estes vértices aglomerados topologicamente distantes estariam se comunicando com muita

[8] O termo *impulsos* refere-se genericamente às perturbações induzidas em uma proteína por outros agentes presentes no seu ambiente, tal como a interação de um ligante em um sítio ativo.

[9] A apresentação do parâmetro *strength* é feita na seção 3.4.2.

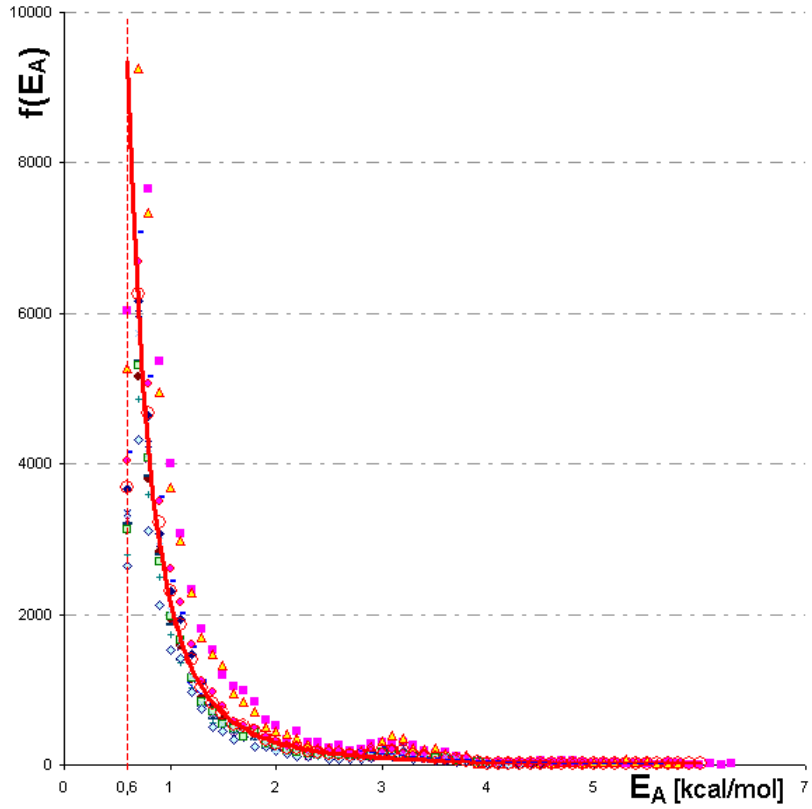


Figura 5.31 – *Distribuições, para as globinas, da freqüência de níveis de energia por átomo $f(E_A)$, e níveis de energia por átomo E_A (em kcal/mol) para as proteínas solvatadas com oclusão. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média (em vermelho) segue uma curva de potência $f(E_A) = \alpha(E_A)^{\gamma}$, com parâmetros $\alpha = 2178$ e $\gamma = -2,851$ com $R^2 = 0,96$.*

dificuldade ou estariam privados de se comunicar [Albert e Barabasi (1999), Barabasi et al. (1999), Newman (2003c), Gol'dshtein et al. (2004)].

No caso das proteínas, tais vértices estariam facultando a ligação entre resíduos que não compartilham a mesma vizinhança dentro da estrutura primária. Nas globinas, resíduos que compartilham a mesma vizinhança imediata só estabelecem ligações não-covalentes energeticamente relevantes ($E > 0,6\ kcal/mol$) quando formam hélices. Topologicamente, as hélices apresentam um passo com $\approx$ 4,0 resíduos. Desta forma, foi feita a identificação das interações cujos resíduos estejam a mais de 4 posições de distância. A rede formada por estas interações foi analisada e os resultados apresentados na figura 5.32.

A partir da análise dos dados da figura 5.32, observa-se que o comportamento de $f(N_{CA})$ apresenta uma tendência média de decaimento seguindo uma lei de potência na forma $f(N_{CA}) = \alpha(N_{CA})^{\gamma}$, com parâmetros $\alpha = 142$ e $\gamma = -1,57$ com $R^2 = 0,88$. Contudo, é necessário ir além da quantificação das interações e avaliar as energias associadas à estas interações de longo alcance. O cálculo das energias associadas às interações do conjunto analisado na figura 5.32, foi feito e está apresentado na figura 5.33.

Os dados relativos ao gráfico da figura 5.33, mostram que a freqüência média das interações

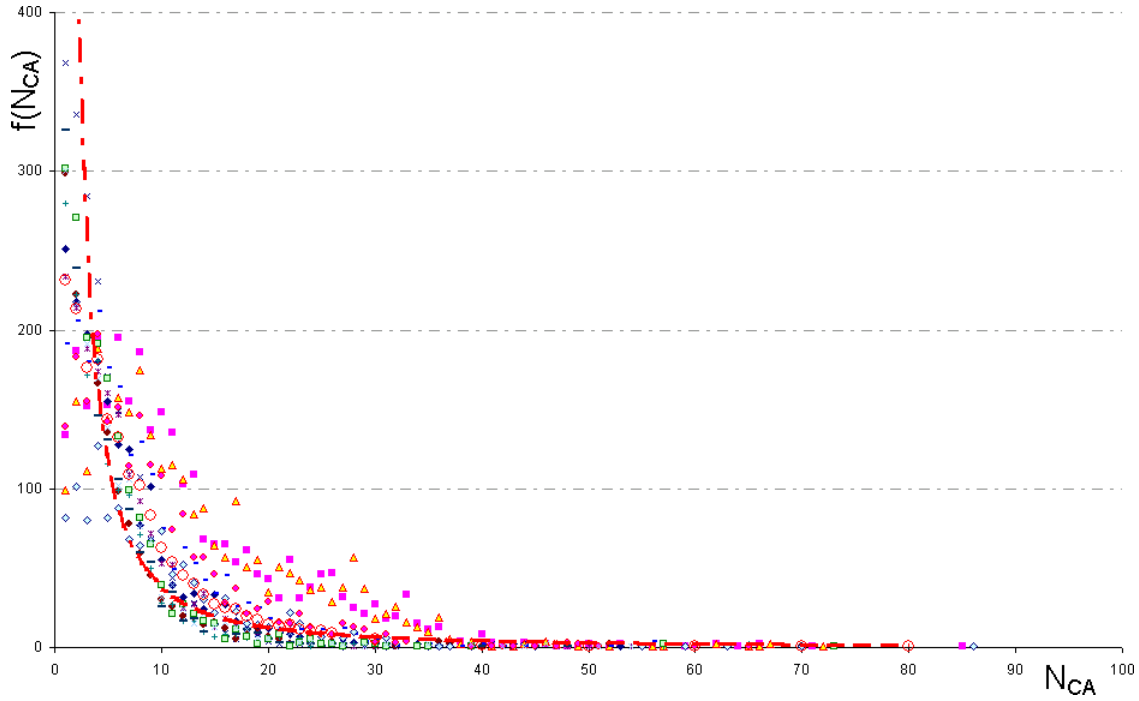


Figura 5.32 – *Distribuições, para as globinas, de $f(N_{CA})$ e N_{CA}, considerando somente as interações de resíduos distantes entre si mais de 4 posições, com as proteínas solvatadas e adotando os critérios de oclusão. Cada uma das proteínas estudadas é representada por um conjunto de pontos com símbolos e cores diferentes. O decaimento médio das curvas ajusta-se a uma distribuição de lei de potência (curva em vermelho) na forma $f(N_{CA}) = \alpha(N_{CA})^{\gamma}$, com parâmetros $\alpha = 142$ e $\gamma = -1,57$ com $R^2 = 0,88$.*

de longa distância, ponderadas pelas respectivas energias potenciais - $f(E_A)$, apresenta uma tendência ainda mais acentuada decai seguindo uma lei de potência ($f(E_A) = \alpha(E_A)^{\gamma}$, apresentando parâmetros $\alpha = 961,84$, $\gamma = -3,03$ com $R^2 = 0,98$.

É interessante observar que o comportamento médio das redes formadas pelas interações não-covalentes de longa distância, subjacentes às globinas estudadas, mostra-se similar ao comportamento do modelo de Barabasi [Albert e Barabasi (1999), Barabasi et al. (1999)]. Enquanto o expoente da função de decaimento, da relação $f(N_{CA}) \times N_{CA}$, para as globinas é $\gamma = -1,57$, o expoente do decaimento de $f(E_A) \times E_A$ é $\gamma = -3,03$. Notavelmente, este último expoente é similar àquele identificado, por diferentes autores, para as estruturas de outros sistemas encontrados no mundo real.

A análise desses dois gráficos mostra que também as estruturas tridimensionais das proteínas, apresentam um conjunto restrito de elementos (neste caso os átomos), que responde por um grande número de interações não locais no seio das proteínas. Quando se leva em conta as energias potenciais intrínsecas a estas interações, observa-se que relação $f(E_A) \times E_A$ mostra um decaimento em lei de potência muito mais acentuado que aquele apresentado pela relação $f(N_{CA}) \times N_{CA}$. Esta diferença de comportamentos mostra que a análise baseada nas energias potenciais das interações não-covalentes, informa mais sobre a estrutura das globinas, que a análise baseada somente no comprimento dessas interações.

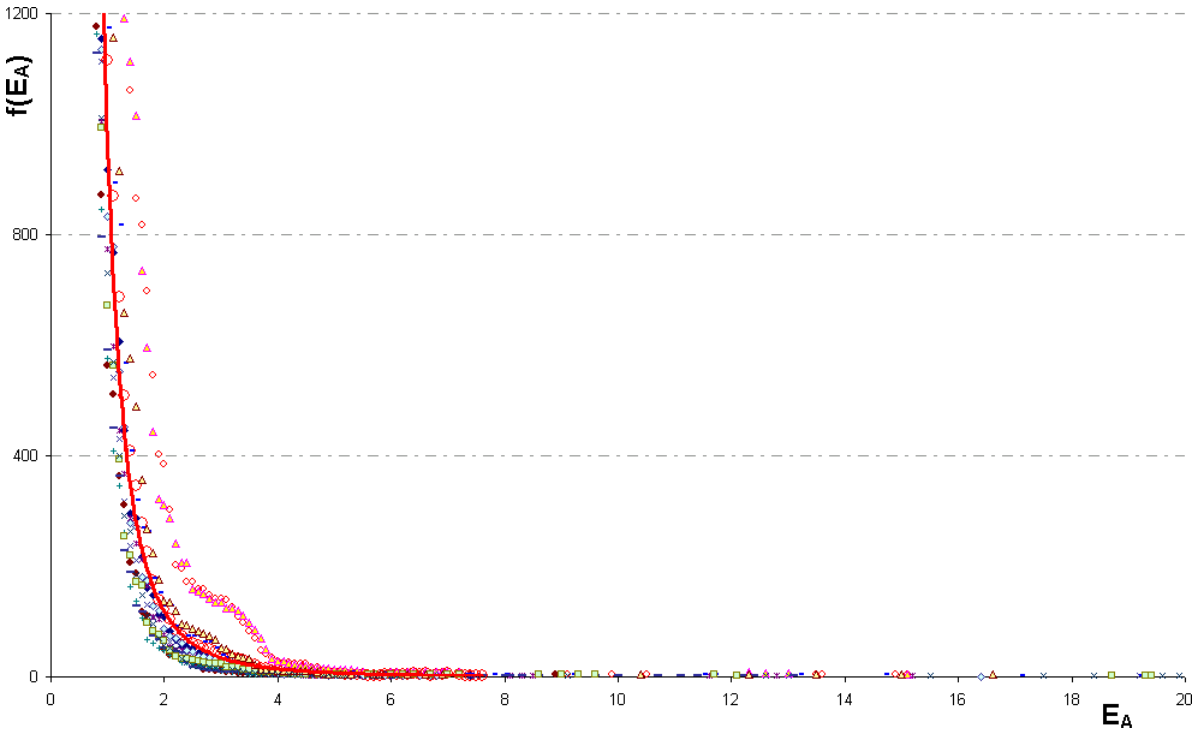


Figura 5.33 – *Distribuições, para as globinas, da freqüência de níveis de energia por átomo $f(E_A)$, e níveis de energia por átomo E_A (em kcal/mol), considerando somente as interações de resíduos distantes entre si mais de 4 posições, para as proteínas solvatadas com oclusão. Cada uma das proteínas estudadas é representada por um conjunto de pontos com cores diferentes. A curva de tendência média (em vermelho) segue uma curva de potência $f(E_A) = \alpha(E_A)^{\gamma}$, tendo $\alpha = 961,84$, $\gamma = -3,025$ com $R^2 = 0,98$.*

Voltando a atenção para os átomos relacionados aos valores de energia potencial das interações - E_A, presentes na cauda da curva mostrada na figura 5.33, lembramos que estes são os átomos que exercem um papel crítico para a estabilidade e a percolação dentro das globinas. Em face disto, torna-se necessário determinar o limiar mínimo para E_A a partir do qual os átomos podem formalmente ser considerados como átomos críticos (ou "*hubs*") dentro do conjunto de globinas estudado.

Para tanto, pode-se recorrer à análise da derivada segunda da curva de decaimento $f(E_A)$, buscando um valor de E_A tal que

$$\frac{\partial^2 f(E_A)}{\partial E_A} = 0.$$

O estudo da derivada segunda justifica-se, pois este permite identificar o valor de E_A a partir do qual o decaimento de $f(E_A)$ "desacelera", tornando-se praticamente invariável ou seja, $f(E_A)$ torna-se praticamente constante. A análise da derivada segunda de $f(E_A)$ mostrou que o decaimento dessa função assume valores tais que para $E_A \geq 4\ kcal/mol$, $f(E_A) < 10^{-2}$, tornando-se praticamente imutável. Desta forma, os átomos que apresentam valores de $E_A \geq 4\ kcal/mol$ serão selecionados como átomos/resíduos "*hubs*".

Uma vez identificados quais são estes átomos, para cada uma das globinas estudadas,

foi feita a identificação dos resíduos nos quais estes átomos estão presentes. Com base no alinhamento estrutural das proteínas estudadas, tal como mostrado na figura 5.34, obteve-se um alinhamento de resíduos e sobre este alinhamento foram identificados os resíduos nos quais foram encontrados os átomos "*hub*". A tabela 5.10 apresenta uma amostra deste conjunto de resíduos "*hub*" identificados em algumas das globinas estudadas.

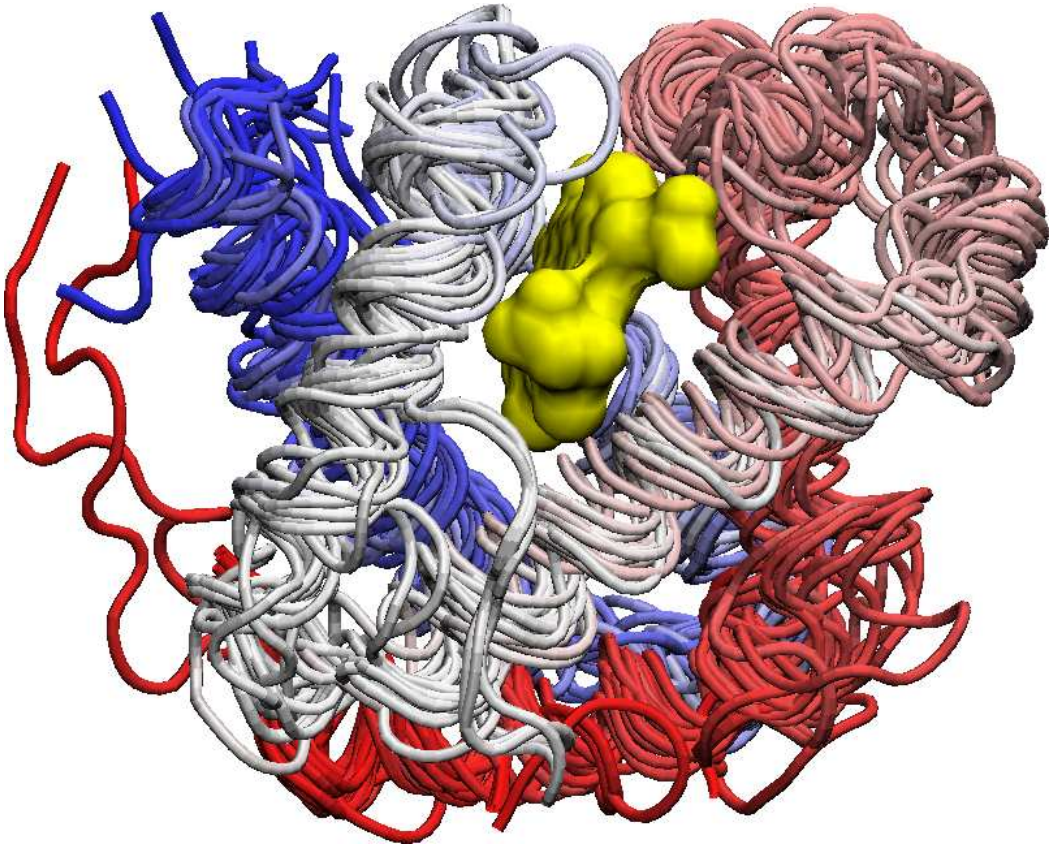


Figura 5.34 – *Alinhamento estrutural das globinas estudadas. O alinhamento dos grupos* ***HEME*** *está destacado em laranja.*

Tabela 5.10 – *Seqüências resultantes do alinhamento estrutural das globinas, ressaltando (em amarelo) os resíduos mais energeticamente ligados.*

	-133	-132	-131	-130	-129	-128	-127	-126	-125	-124	-123	-122	-121	-120	-119	-118	-117	-116	-115	-114	-113	-112	-111	-110	-109	-108	-107	-106	-105	-104
1A6G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	V	-	L	S	E	G	E	W	Q	L	V	L	-	H	V
1ASH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	N	K	T	R	E	L	C	M	-	K	S
1B0B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	L	S	A	A	Q	K	D	N	V	K	-	S	S
1BZP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	V	-	L	S	E	G	E	W	Q	L	V	L	-	H	V
1DLW	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	L	-	F	-	E	Q
1DLY	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	L	-	F	-	A	K
1ECD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	L	S	A	D	Q	I	S	T	V	Q	-	A	S
1GDJ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	A	-	L	T	E	S	Q	A	A	L	V	K	-	S	S
1HLM	-	G	-	A	T	Q	S	F	-	Q	S	V	-	-	-	G	D	L	T	P	A	E	K	D	L	I	R	-	S	T
1JF3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	A	A	Q	R	Q	V	V	A	-	S	T
1JF4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	A	A	Q	R	Q	V	V	A	-	S	T
1KR7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1LHS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	D	D	E	W	N	H	V	L	-	G	I
1MBS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	D	G	E	W	H	L	V	L	-	N	V
1MYT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	D	F	D	A	V	L	-	K	C
1Q1F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	R	P	E	S	E	L	I	R	-	Q	S
1RTX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	T	-	L	Y	E	K
1UVX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	L	-	F	-	A	K
1V5H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	E	-	A	E	R	K	A	V	Q	-	A	M
2FAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	-	L	S	A	A	E	A	D	L	A	G	-	K	S
2MM1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	D	G	E	W	Q	L	V	L	-	N	V

(a)

Continua . . .

Tabela 5.10 – *(continuação)*

	-103	-102	-101	-100	-99	-98	-97	-96	-95	-94	-93	-92	-91	-90	-89	-88	-87	-86	-85	-84	-83	-82	-81	-80	-79	-78	-77	-76	-75	-74	-73	-72	-71	-70	-69	-68	-67
1A6G	W	A	-	K	V	-	E	A	-	-	-	-	D	-	V	A	G	-	H	G	Q	D	I	L	I	R	L	F	K	S	H	P	E	T	L	E	K
1ASH	L	E	-	H	A	K	V	D	-	T	S	N	E	-	A	R	Q	-	D	G	I	D	L	Y	K	H	M	F	E	N	Y	P	P	L	R	K	Y
1B0B	W	A	-	K	A	-	S	A	-	-	-	-	A	-	W	G	T	-	A	G	P	E	F	F	M	A	L	F	D	A	H	D	D	V	F	A	K
1BZP	W	A	-	K	V	-	E	A	-	-	-	-	D	-	V	A	G	-	H	G	Q	D	I	L	I	R	L	F	K	S	H	P	E	T	L	E	K
1DLW	L	-	-	G	G	-	Q	A	-	-	-	-	A	-	V	Q	A	-	V	T	A	Q	F	Y	A	N	I	Q	A	D	-	A	T	V	A	T	F
1DLY	L	-	-	G	G	-	R	E	-	-	-	-	A	-	V	E	A	-	A	V	D	K	F	Y	N	K	I	V	A	D	-	P	T	V	S	T	Y
1ECD	F	D	-	K	V	-	K	G	-	-	-	-	D	-	-	-	-	-	-	P	V	G	I	L	Y	A	V	F	K	A	D	P	S	I	M	A	K
1GDJ	W	E	-	E	F	-	N	A	-	-	-	-	N	-	I	P	K	-	H	T	H	R	F	F	I	L	V	L	E	I	A	P	A	A	K	D	L
1HLM	W	D	-	Q	L	-	M	T	-	-	-	-	H	-	R	T	G	-	F	V	A	D	V	F	I	R	I	F	H	N	D	P	T	A	Q	R	K
1JF3	W	K	-	D	I	-	A	G	A	-	-	-	D	N	G	A	G	-	V	G	K	E	C	L	S	K	F	I	S	A	H	P	E	M	A	A	V
1JF4	W	K	-	D	I	-	A	G	S	-	-	-	D	N	G	A	G	-	V	G	K	E	C	F	T	K	F	L	S	A	H	H	D	M	A	A	V
1KR7	-	M	-	V	-	-	-	-	-	-	-	-	N	-	W	A	A	-	V	V	D	D	F	Y	Q	E	L	F	K	A	H	P	E	Y	Q	N	K
1LHS	W	A	-	K	V	-	E	P	-	-	-	-	D	-	L	S	A	-	H	G	Q	E	V	I	I	R	L	F	Q	L	H	P	E	T	Q	E	R
1MBS	W	G	K	V	E	-	-	T	-	-	-	-	D	-	L	A	G	-	H	G	Q	E	V	L	I	R	L	F	K	S	H	P	E	T	L	E	K
1MYT	W	G	-	P	V	-	E	A	-	-	-	-	D	-	Y	T	T	-	M	G	G	L	V	L	T	R	L	F	K	E	H	P	E	T	Q	K	L
1Q1F	W	R	-	V	V	-	S	R	-	-	-	-	S	-	P	L	E	-	H	G	T	V	L	F	A	R	L	F	A	L	E	P	S	L	L	P	L
1RTX	L	-	-	G	G	-	T	T	-	-	-	-	A	-	V	D	L	-	A	V	D	K	F	Y	E	R	V	L	Q	D	-	D	R	I	K	H	F
1UVX	L	-	-	G	G	-	R	E	-	-	-	-	A	-	V	E	A	-	A	V	D	K	F	Y	N	K	I	V	A	D	-	P	T	V	S	T	Y
1Y5H	W	A	-	R	L	-	Y	A	-	-	-	-	N	-	-	C	E	D	V	G	V	A	I	L	V	R	F	F	V	N	F	P	S	A	K	Q	Y
2FAL	W	A	-	P	V	-	F	A	-	-	-	-	N	-	K	N	A	-	N	G	L	D	F	L	V	A	L	F	E	K	F	P	D	S	A	N	F
2MM1	W	G	-	K	V	-	E	A	-	-	-	-	D	-	I	P	G	-	H	G	Q	E	V	L	I	R	L	F	K	G	H	P	E	T	L	E	K

(b)

	-66	-65	-64	-63	-62	-61	-60	-59	-58	-57	-56	-55	-54	-53	-52	-51	-50	-49	-48	-47	-46	-45	-44	-43	-42	-41	-40	-39	-38	-37	-36	-35	-34	-33	-32	-31	-30	-29
1A6G	F	-	D	R	-	-	F	-	K	H	L	K	T	E	A	E	M	K	A	S	E	D	L	K	K	H	G	V	T	V	L	T	A	L	G	A	I	L
1ASH	F	-	K	S	-	-	R	-	E	E	Y	T	-	A	E	D	V	Q	N	D	P	F	F	A	K	Q	G	Q	K	I	L	L	A	C	H	V	L	C
1B0B	F	S	G	L	-	-	F	-	S	G	A	A	-	K	G	T	V	K	N	T	P	E	M	A	A	Q	A	Q	S	F	K	G	L	V	S	N	W	V
1BZP	F	-	D	R	-	-	F	-	K	H	L	K	T	E	A	E	M	K	A	S	E	D	L	K	K	H	G	V	T	V	L	T	A	L	G	A	I	L
1DLW	F	N	G	I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D	M	P	N	Q	T	N	K	T	A	A	F	L	C	A	A
1DLY	F	S	N	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D	M	K	V	Q	R	S	K	Q	F	A	F	L	A	Y	A	L
1ECD	F	-	T	Q	-	-	F	-	A	G	K	-	D	L	E	S	I	K	G	T	A	P	F	E	T	H	A	N	R	I	V	G	F	K	S	K	I	I
1GDJ	F	-	S	F	-	-	L	-	K	G	T	S	E	V	-	-	P	Q	N	N	P	E	L	Q	A	H	A	G	K	V	F	K	L	V	Y	E	A	A
1HLM	F	-	P	Q	-	-	M	A	G	L	S	P	-	A	E	-	L	R	T	S	R	Q	M	H	A	H	A	I	R	V	S	A	L	M	T	T	Y	I
1JF3	F	-	G	-	-	-	F	-	-	S	-	G	-	A	S	-	-	-	-	-	D	P	G	V	A	E	L	G	A	K	V	L	A	Q	I	G	V	A	V
1JF4	F	-	G	-	-	-	F	-	-	S	-	G	-	A	S	-	-	-	-	-	D	P	G	V	A	D	L	G	A	K	V	L	A	Q	I	G	V	A	V
1KR7	F	-	-	G	-	-	F	-	K	G	V	-	A	L	G	S	L	K	G	N	A	A	Y	K	T	Q	A	G	K	T	V	D	Y	I	N	A	A	I
1LHS	F	-	A	K	-	-	F	-	K	N	L	T	T	I	D	A	L	K	S	S	E	E	V	K	K	H	G	T	T	V	L	T	A	L	G	R	I	L
1MBS	F	-	D	K	-	-	F	-	K	H	L	K	S	E	D	D	M	R	R	S	E	D	L	R	K	H	G	N	T	V	L	T	A	L	G	G	I	L
1MYT	F	-	P	K	-	-	F	-	A	G	I	A	-	Q	A	D	I	A	G	N	A	A	I	S	A	H	G	A	T	V	L	K	K	L	G	E	L	L
1Q1F	F	-	Q	Y	N	G	R	-	-	Q	F	S	S	P	E	D	S	L	S	S	P	E	F	L	D	H	I	R	K	V	M	L	V	I	D	A	A	V
1RTX	F	A	D	V	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D	M	A	K	Q	R	A	H	Q	K	A	F	L	T	Y	A	F
1UVX	F	S	N	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D	M	K	V	Q	R	S	K	Q	F	A	F	L	A	Y	A	L
1Y5H	F	-	S	Q	-	-	F	-	K	H	M	E	D	P	L	E	M	E	R	S	P	Q	L	R	K	H	A	C	R	V	M	G	A	L	N	T	V	V
2FAL	F	-	A	D	-	-	F	-	K	G	K	S	-	V	A	D	I	K	A	S	P	K	L	R	D	V	S	S	R	I	F	T	R	L	N	E	F	V
2MM1	F	-	D	R	-	-	F	-	K	H	L	K	S	E	D	E	M	K	A	S	E	D	L	K	K	H	G	A	T	V	L	T	A	L	G	G	I	L

(c)

	-28	-27	-26	-25	-24	-23	-22	-21	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1A6G	-	-	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-	E	L	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-	-	-	K	I	P	I	K
1ASH	-	-	-	A	T	Y	-	D	D	R	E	T	F	-	-	-	-	N	A	-	Y	T	R	E	L	L	D	R	H	A	-	R	D	H	V	-	-	-	-	-	H	M	P	P	E
1B0B	-	-	-	D	N	L	-	D	N	A	G	A	L	-	-	-	-	E	G	-	Q	C	K	T	F	A	A	N	H	K	-	A	-	-	R	-	-	-	-	-	G	I	S	A	G
1BZP	-	-	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-	E	L	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-	-	-	K	I	P	I	K
1DLW	G	G	P	N	-	-	-	-	-	-	-	A	W	T	G	R	-	-	-	-	N	-	-	-	L	K	E	V	H	A	-	N	-	-	M	-	-	-	-	-	G	V	S	N	A
1DLY	G	G	-	A	-	-	-	-	-	S	E	W	K	G	K	-	-	-	-	-	D	-	-	-	M	R	T	A	H	K	-	D	-	-	L	-	-	V	P	-	H	L	S	D	V
1ECD	-	-	-	G	E	L	-	-	-	P	N	-	I	-	-	-	-	E	A	-	D	V	N	T	F	V	A	S	H	K	-	P	-	-	R	-	-	-	-	-	G	V	T	H	D
1GDJ	-	-	-	I	Q	L	E	V	T	G	V	-	V	-	-	-	V	T	D	A	T	L	K	N	L	G	S	V	H	V	-	S	-	-	K	-	-	-	-	-	G	V	A	D	A
1HLM	-	-	-	D	E	M	-	D	T	-	E	V	L	-	-	-	-	P	E	-	L	L	A	T	L	T	R	T	H	D	-	K	-	-	N	-	-	-	-	-	H	V	G	K	K
1JF3	-	-	-	S	H	L	-	G	D	E	G	K	M	-	-	-	-	V	A	-	E	M	K	A	V	G	V	R	H	K	-	G	Y	-	G	-	-	-	N	K	H	I	K	A	E
1JF4	-	-	-	S	H	L	-	G	D	E	G	K	M	-	-	-	-	V	A	-	E	M	K	A	V	G	V	R	H	K	-	G	Y	-	G	-	-	-	N	K	H	I	K	A	E
1KR7	-	-	-	G	G	-	-	-	-	-	-	S	A	-	-	-	-	-	-	-	D	A	A	G	L	A	S	R	H	K	-	G	-	-	R	-	-	-	-	-	N	V	G	S	A
1LHS	-	-	-	K	Q	K	-	-	-	N	N	-	H	-	-	-	-	E	Q	-	E	L	K	P	L	A	E	S	H	A	-	T	K	-	H	-	-	-	-	-	K	I	P	V	K
1MBS	-	-	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-	E	L	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-	-	-	K	I	P	I	K
1MYT	-	-	-	K	A	K	-	-	-	G	S	-	H	-	-	-	-	A	A	-	I	L	K	P	L	A	N	S	H	A	-	T	K	-	H	-	-	-	-	-	K	I	P	I	N
1Q1F	-	-	-	T	N	V	-	E	D	L	S	S	L	-	-	-	-	E	E	-	Y	L	T	S	L	G	R	K	H	R	-	A	-	-	V	-	-	-	-	-	G	V	R	L	S
1RTX	G	-	-	G	-	-	-	-	-	T	D	K	Y	-	-	-	-	-	-	-	D	G	R	Y	M	R	E	A	H	K	-	E	-	-	L	V	E	N	H	-	G	L	N	G	E
1UVX	G	G	-	A	-	-	-	-	-	-	S	E	W	K	G	K	-	-	-	-	D	-	-	-	M	R	T	A	H	K	-	D	-	-	L	-	-	V	P	-	H	L	S	D	V
1Y5H	-	-	-	E	N	L	-	H	D	P	D	K	V	-	-	-	-	S	S	-	V	L	A	L	V	G	K	A	H	A	-	L	K	-	H	-	-	-	-	-	K	V	E	P	V
2FAL	-	-	-	N	N	A	-	A	N	A	G	K	M	-	-	-	-	S	A	-	M	L	S	Q	F	A	K	E	H	V	-	G	-	-	F	-	-	-	-	-	G	V	G	S	A
2MM1	-	-	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-	E	I	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-	-	-	K	I	P	V	K

(d)

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61
1A6G	Y	L	E	F	I	S	E	A	I	I	H	V	L	H	S	R	H	P	-	G	N	F	G	A	D	A	Q	G	A	M	N	K	A	L	E	L	-	-	F	R	K	D	I	A	A
1ASH	V	W	T	D	F	W	K	L	F	E	E	Y	L	G	K	K	T	-	-	T	-	L	D	E	P	T	K	Q	A	W	H	E	I	G	R	E	-	-	F	A	K	E	I	N	K
1B0B	Q	L	E	A	A	F	K	V	L	A	G	F	M	K	S	Y	-	-	-	G	-	G	-	D	-	-	E	G	A	W	T	A	V	A	G	A	-	-	L	M	G	M	I	R	P
1BZP	Y	L	E	F	I	S	E	A	I	I	H	V	L	H	S	R	H	P	-	G	D	F	G	A	D	A	Q	G	A	M	N	K	A	L	E	L	-	-	F	R	K	D	I	A	A
1DLW	Q	F	T	T	V	I	G	H	L	R	S	A	L	T	G	A	-	-	-	G	-	V	A	A	A	L	V	E	Q	T	V	A	V	A	E	T	-	-	V	R	G	D	V	V	-
1DLY	H	F	Q	A	V	A	R	H	L	S	D	T	L	T	E	L	-	-	-	G	-	V	P	P	E	D	I	T	D	A	M	A	Y	V	A	S	-	-	T	R	T	E	V	L	-
1ECD	Q	L	N	N	F	R	A	G	F	V	S	Y	M	K	A	H	T	-	-	D	-	F	-	A	G	A	E	A	A	W	G	A	T	L	D	T	-	-	F	F	G	M	I	F	S
1GDJ	H	F	P	V	V	K	E	A	I	L	K	T	I	K	E	V	V	G	-	A	K	W	S	E	E	L	N	S	A	W	T	I	A	Y	D	E	-	-	L	A	I	V	I	K	K
1HLM	N	Y	D	L	F	G	K	V	L	M	E	A	I	K	A	-	E	L	G	V	G	F	T	K	Q	V	H	D	A	V	A	K	T	F	A	I	-	-	V	Q	G	V	-	L	I
1JF3	Y	F	E	P	L	G	A	S	L	L	S	A	M	E	H	R	I	G	-	G	K	M	N	A	A	A	K	D	A	W	A	A	A	Y	G	D	-	-	I	S	G	A	L	I	S
1JF4	Y	F	E	P	L	G	A	S	L	L	S	A	M	E	H	R	I	G	-	G	K	M	N	A	A	A	K	D	A	W	A	A	A	Y	G	D	-	-	I	S	G	A	L	I	S
1KR7	E	F	H	N	A	K	A	C	L	A	K	A	C	S	A	H	-	-	-	G	-	A	-	-	-	P	D	-	-	L	G	-	-	-	-	H	-	-	A	I	D	D	I	L	S
1LHS	Y	L	E	F	I	C	E	I	I	V	K	V	I	A	E	K	H	P	-	S	D	F	G	A	D	S	Q	A	A	M	K	K	A	L	E	L	-	-	F	R	N	D	M	A	S
1MBS	Y	L	E	F	I	S	E	A	I	I	H	V	L	H	S	K	H	P	-	A	E	F	G	A	D	A	Q	A	A	M	K	K	A	L	E	L	-	-	F	R	N	D	I	A	A
1MYT	N	F	K	L	I	S	E	V	L	V	K	V	M	H	E	K	A	-	-	G	-	L	D	A	G	G	Q	T	A	L	R	N	V	M	G	I	-	-	I	I	A	D	L	E	A
1Q1F	S	F	S	T	V	G	E	S	L	L	Y	M	L	E	K	S	L	G	-	P	D	F	T	P	A	T	R	T	A	W	S	R	L	Y	G	A	-	-	V	V	Q	A	M	S	R
1RTX	H	F	D	A	V	A	E	D	L	L	A	T	L	K	E	M	-	-	-	G	-	V	P	E	D	L	I	A	E	V	A	A	V	A	G	A	P	A	H	K	R	D	V	L	-
1UVX	H	F	Q	A	V	A	R	H	L	S	D	T	L	T	E	L	-	-	-	G	-	V	P	P	E	D	I	T	D	A	M	A	V	V	A	S	-	-	T	R	T	E	V	L	-
1Y5H	Y	F	K	I	L	S	G	V	I	L	E	V	V	A	E	E	F	A	-	S	D	F	P	P	E	T	Q	R	A	W	A	K	L	R	G	L	-	-	I	Y	S	H	V	T	A
2FAL	Q	F	E	N	V	R	S	M	F	P	G	F	V	A	S	V	A	-	-	A	-	P	P	A	G	A	D	A	A	W	T	K	L	F	G	L	-	-	I	I	D	A	L	K	A
2MM1	Y	L	E	F	I	S	E	A	I	I	Q	V	L	Q	S	K	H	P	-	G	D	F	G	A	D	A	Q	G	A	M	N	K	A	L	E	L	-	-	F	R	K	D	M	A	S

(e)

Continua . . .

Tabela 5.10 – *(continuação)*

	62	63	64	65	66	67	68	69	70	71	72	73	74	75
1A6G	K	Y	-	K	-	E	L	G	Y	-	-	-	-	-
1ASH	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1B0B	D	-	-	-	-	-	-	-	M	-	-	-	-	-
1BZP	K	Y	-	K	-	E	L	G	Y	Q	G	-	-	-
1DLW	-	T	-	V	-	-	-	-	-	-	-	-	-	-
1DLY	-	N	-	M	-	P	-	-	-	-	-	-	Q	Q
1ECD	K	M	-	-	-	-	-	-	-	-	-	-	-	-
1GDJ	E	M	-	D	-	D	A	A	-	-	-	-	-	-
1HLM	T	K	-	H	-	A	-	S	-	-	-	-	-	-
1JF3	G	L	-	Q	-	S	-	-	-	-	-	-	-	-
1JF4	G	L	-	Q	-	S	-	-	-	-	-	-	-	-
1KR7	H	L	-	-	-	-	-	-	-	-	-	-	-	-
1LHS	K	Y	-	K	-	E	F	G	F	-	Q	G	-	-
1MBS	K	Y	-	K	E	L	G	F	H	-	G	-	-	-
1MYT	N	Y	-	K	-	E	L	G	F	-	S	G	-	-
1Q1F	G	W	D	G	-	-	-	-	-	-	-	-	-	-
1RTX	-	N	-	Q	-	-	-	-	-	-	-	-	-	-
1UVX	-	N	-	M	-	P	-	-	-	-	-	-	Q	Q
1V5H	A	Y	-	K	-	E	V	G	W	-	-	-	-	-
2FAL	A	-	-	-	-	-	-	G	A	-	-	-	-	-
2MM1	N	Y	-	K	-	E	L	G	F	-	Q	G	-	-

(f)

Para identificar padrões recorrentes em toda as globinas, foram selecionados os resíduos "*hubs*" existentes em regiões espacialmente alinhadas, onde não existe a ocorrência de "*gaps*". O resultado desta seleção é mostrado na figura 5.35.

Para unificar a forma de referenciar os resíduos, neste alinhamento estrutural, ao resíduo de histidina conservado, comum à todas as globinas, foi atribuído o índice 0 (zero) fazendo deste o centro do sistema de referência para este estudo. É interessante observar que tanto os resíduos de histidina quanto de fenilalanina conservados aparecem também como "*hubs*" em todas as globinas estudadas. Tendo esta histidina como origem do sistema de referência, observa-se que em todas as globinas estudadas esta fenilalanina encontra-se sempre a 66 posições a montante[10] da **HIS0**.

Na tabela apresentada na figura 5.35, a segunda coluna mostra a posição de cada resíduo relativo a este resíduo de HIS conservado. Nesta mesma tabela, a última coluna indica qual o elemento estrutural onde se localizam os resíduos listados em cada linha. Já a penúltima coluna indica qual a probabilidade de ocorrência de um "*hub*" naquela posição estrutural.

Para facilitar a identificação dos resíduos, elegendo o resíduo conservado de HIS como origem de referência, adotada-se neste trabalho a seguinte forma de indicação:

- O resíduo de HIS, tomado como origem, será indicado como **HIS0**;
- Caso um resíduo mantenha-se conservado em uma posição, ele será explicitamente indicado (ex. **PHE66-**). Senão, será utilizada a indicação genérica **POS**;
- Um resíduo qualquer localizado n posições distante da **HIS0** em direção à porção N terminal da globina será indicado como $[\langle res \rangle \mid \mathbf{POS}]\boldsymbol{n}$-;
- Um resíduo qualquer localizado n posições distante da **HIS0** em direção à porção C terminal da globina será indicado como $[\langle res \rangle \mid \mathbf{POS}]\boldsymbol{n}+$.

Seguindo essa convenção, esta fenilalanina será referenciada como **PHE66-**[11]

[10] Em direção à posição N terminal da proteína

[11] Esta notação faz indica que o resíduo de PHE está distante 66 posições, na direção N terminal, da **HIS0** de referência.

Seq	Seq	1A6G	1ASH	1BOB	1BZP	1DLW	1DLY	1ECD	1GDJ	1HLM	1JF3	1JF4	1KR7	1LHS	1MBS	1MYT	1Q1F	1RTX	1UVX	1V5H	2FAL	2MM1	%	Elem Estrutural
68	-66	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	100%	CD7
134	0	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	100%	F8
126	-8	E	Y	Q	E	N	D	D	T	L	E	E	D	E	E	I	Y	D	D	V	M	E	52%	EF7
89	-45	D	F	E	D	D	D	P	E	Q	G	G	A	E	D	A	E	D	D	Q	K	D	48%	D10
52	-82	D	D	E	D	Q	K	G	R	D	E	E	D	E	E	L	V	K	K	A	D	E	38%	B7
93	-41	H	Q	Q	H	Q	Q	H	H	H	L	L	Q	H	H	H	H	Q	Q	H	V	H	38%	E3
151	17	Y	V	Q	Y	Q	H	Q	H	N	Y	Y	E	Y	Y	N	S	H	H	Y	Q	Y	38%	G8
51	-83	Q	I	P	Q	A	D	V	H	A	K	K	D	Q	Q	G	T	D	D	V	L	Q	33%	B6
186	52	L	E	A	L	T	S	T	E	I	D	D	H	L	L	I	A	A	S	L	L	L	33%	H14
192	58	D	E	M	D	D	E	M	V	V	A	A	D	D	D	D	A	D	E	H	A	D	33%	H18
103	-31	A	V	N	A	A	Y	K	E	T	V	V	A	R	G	E	A	Y	Y	T	E	G	29%	E13
118	-16	H	F	L	H	W	W	I	V	L	M	M	A	H	H	H	L	Y	W	V	M	H	29%	EF4
149	15	I	P	A	I	N	D	H	D	K	A	A	S	V	I	I	L	G	D	P	S	V	29%	G6
153	19	E	T	E	E	T	Q	N	P	D	E	E	H	E	E	K	S	D	Q	K	E	E	29%	G10
165	31	S	K	S	S	G	E	A	E	A	H	H	A	E	S	E	K	E	E	E	S	S	29%	G22
177	43	Q	K	E	Q	V	I	E	N	H	K	K	D	Q	Q	Q	R	I	I	Q	D	Q	29%	H5
181	47	N	H	T	N	V	M	G	T	A	A	A	G	K	K	R	S	A	M	A	T	N	29%	H9
43	-91	D	E	A	D	A	A	D	N	H	D	D	N	D	D	D	S	A	A	N	N	D	24%	AB3
56	-78	R	H	A	R	N	K	A	L	R	K	K	E	R	R	R	R	R	K	R	A	R	24%	B11
63	-71	E	P	D	E	T	T	S	A	T	E	D	E	E	E	E	S	R	T	S	D	E	24%	C2
92	-42	K	K	A	K	N	V	T	A	A	E	D	T	K	K	A	D	K	V	K	D	K	24%	E2
102	-32	G	H	S	G	C	A	S	Y	T	G	G	N	G	G	G	D	T	A	N	N	G	24%	E12
132	-2	Q	D	A	Q	E	T	A	S	R	V	V	S	E	Q	N	R	E	T	K	K	Q	24%	F6
150	16	K	E	G	K	A	V	D	A	K	E	E	A	K	K	N	S	E	V	V	A	K	24%	G7
190	56	R	A	M	R	R	R	F	A	Q	S	S	I	R	R	I	V	K	R	Y	I	R	24%	H16
66	-68	E	K	A	E	T	T	A	D	R	A	A	N	E	E	K	P	H	T	Q	N	E	19%	C5
131	-3	A	L	A	A	K	R	V	G	T	G	G	A	A	A	A	G	R	R	G	A	A	19%	F5
154	20	F	D	A	F	T	A	N	V	L	P	P	N	F	F	L	T	A	A	I	N	F	19%	G11
194	60	A	N	R	A	V	L	F	K	L	I	I	L	A	A	E	S	L	L	T	K	A	19%	H20
34	-100	K	H	K	K	G	G	K	E	Q	D	D	V	K	V	P	V	G	G	R	P	K	14%	A14
54	-80	L	Y	F	L	Y	Y	L	F	F	L	F	Y	I	L	L	F	Y	Y	L	L	L	14%	B9

Figura 5.35 – *Resíduos* "hubs" *(em amarelo) identificados em regiões espacialmente alinhadas nas globinas estudadas, onde não há ocorrência de* gaps. *A penúltima coluna indica qual a probabilidade da referida posição ser* "hub"*, no conjunto das globinas estudadas. A primeira coluna indica a posição dos resíduos no conjunto do alinhamento global das globinas. A segunda coluna indica a posição dos resíduos em relação à HIS conservada. A última coluna indica a posição, na estrutura secundária (tomando a globina 1A6G como referência) onde os* "hubs" *ocorrem. A histidina conservada é tomada como referência para numeração das posições.*

É possível então mapear os resíduos identificados na figura 5.35 sobre o conjunto de globinas estruturalmente alinhadas, como mostra a figura 5.36. Nesta figura os resíduos "*hub*" estão destacados em amarelo. Ao mesmo tempo, o grupo **HEME** encontra-se destacado em laranja. Desta figura é possível inferir o grau de conservação espacial destes resíduos "*hub*" e do grupo **HEME**. É notável o fato de que mesmo havendo o alinhamento estrutural de globinas de diferentes tamanhos, existe uma conservação acentuada não só da função de "*hub*" de cada posição estrutural, como também da interação par-a-par entre estes resíduos topologicamente conservados. Com o intuito de facilitar a observação deste fato, a figura 5.37 mostra o mesmo conjunto de resíduos mapeados somente sobre a globina de cachalote (PDBID 1A6G).

O mapeamento destes "*hubs*", apresentados na figura 5.35, sobre o conjunto de globinas espacialmente alinhadas, permite definir regiões do espaço topologicamente conservadas e importantes para todas as globinas estudadas. Neste trabalho, o termo *locus*[12] foi adotado para fazer referência a estas regiões. Cabe ressaltar que, neste trabalho, o conceito de *locus* faz referência à uma vizinhança no entorno de um "*hub*", e não a um ponto determinado. Visto que mesmo as proteínas de uma mesma família topológica geralmente apresentam variações no número de resíduos, a localização espacial de uma posição "*hub*" conservada,

[12] O termo latino *locus* (pl. *loci*), é semanticamente equivalente a "lugar onde algo está situado ou acontece" [Valle (2004)].

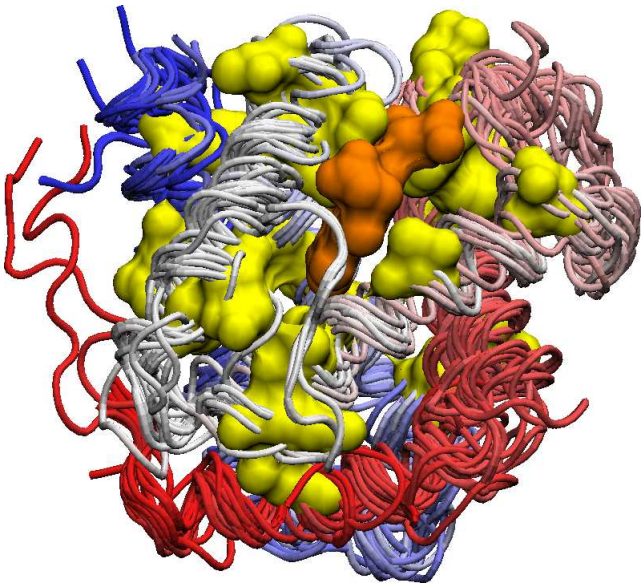


Figura 5.36 – *Mapeamento dos resíduos* "hubs" *(em amarelo) sobre o conjunto de globinas espacialmente alinhadas. Em laranja encontra-se destacado o grupo* ***HEME****.*

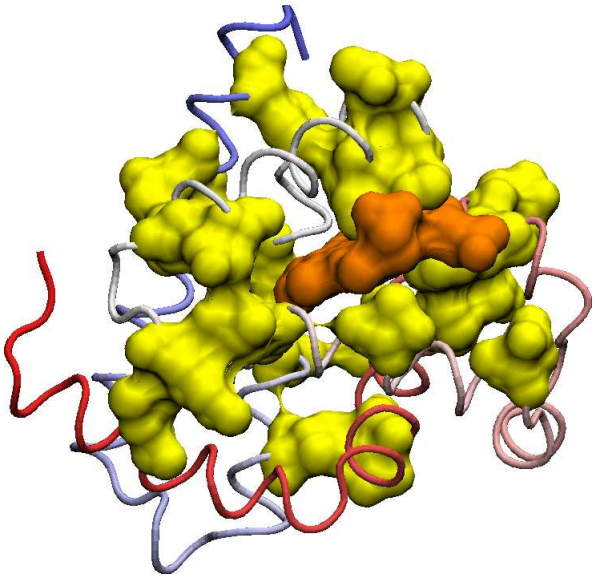


Figura 5.37 – *Mapeamento dos resíduos* "hubs" *(em amarelo) sobre a globina 1A6G.*

acaba variando dentro de uma vizinhança estreita. Este conceito de vizinhança recebe, neste trabalho, o nome de *locus*.

Em todas as globinas estudadas, foram identificados os seguintes *loci* (figura 5.38):

1. **locus 1** – entorno do grupo **HEME**;
2. **locus 2** – ligação entre as hélices **B** e **E**;
3. **locus 3** – ligação entre as hélices **A** e **H**;
4. **locus 4** – ligação entre as hélices **F** e **H**.

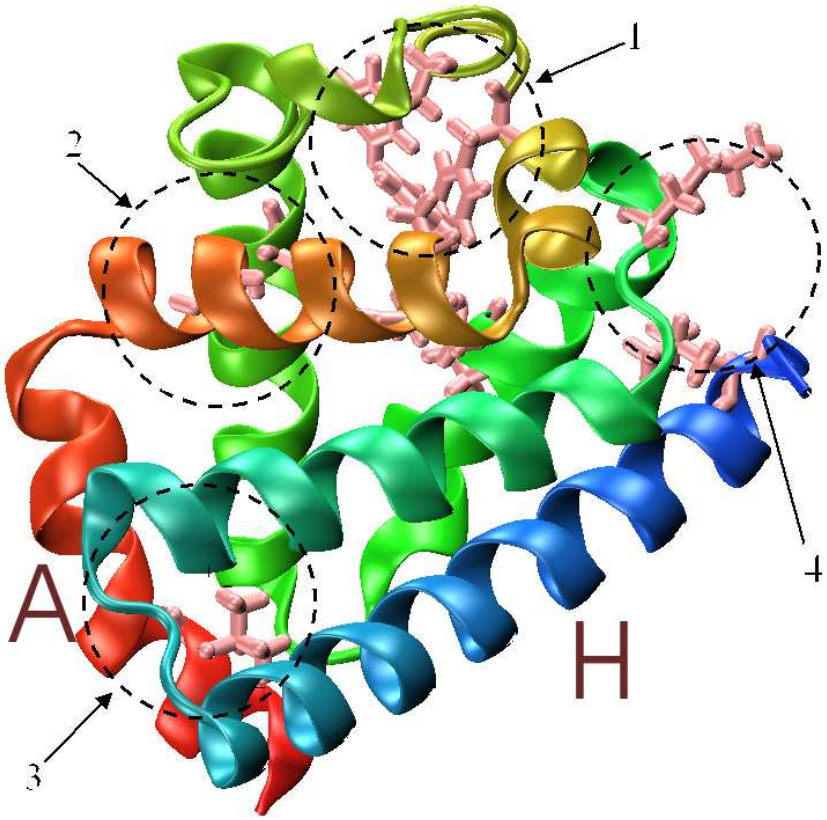


Figura 5.38 *– Resíduos "hub" na estrutura 3D da mioglobina 1A6G. As regiões estruturais marcadas como* ***loci*** *de 1 a 4, indicam onde os dez resíduos conectados com maior energia, foram encontrados. Figura gerada com o uso do VMD.*

Analisando o **locus 1**, 7 elementos encontram-se conservados em 100% das globinas analisadas:

- Grupo **HEME**;
- Resíduo **PHE66-** (res 43 PHE pos CD1 em 1A6G) é "*hub*" em 100% dos casos;
- Resíduo **XXX38-** (res 67 LEU pos E06 em 1A6G) é "*hub*" em 62% dos casos;
- Resíduo **XXX33-** (res 72 LEU pos E11 em 1A6G) é "*hub*" em 62% dos casos;
- Resíduo **HIS0** (res 93 HIS pos F08 em 1A6G) é "*hub*" em 100% dos casos;

- Resíduo **XXX06+** (res 97 HIS pos FG02 em 1A6G) é "*hub*" em 71% dos casos;
- Resíduo **XXX13+** (res 99 ILE pos FG04 em 1A6G) é "*hub*" em 62% dos casos;

Na figura 5.39 observa-se que a presença do grupo **HEME** nas globinas, une firmemente as hélices **C**, **E** e **F**. É possível inferir que a ausência do grupo **HEME** nesta posição tornaria impraticável a aproximação e a estabilização destas porções da proteína. De fato, o grupo **HEME** aparece nos cálculos das energias de interações não-covalentes, como um "*hub*" de alta energia.

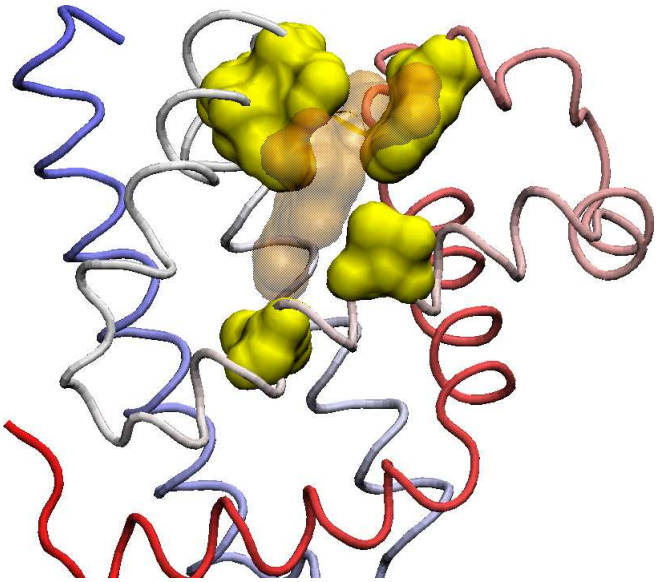
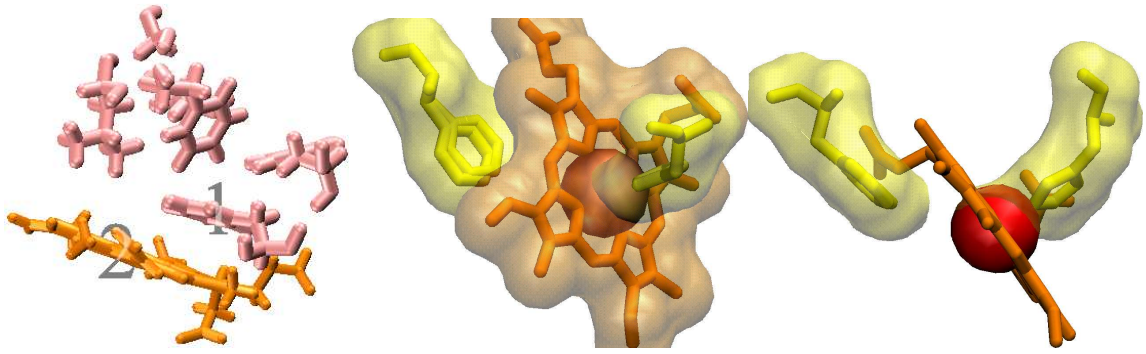


Figura 5.39 *– Resíduos "*hub*" presentes no* ***locus 1*** *da mioglobina 1A6G. Os resíduos são apresentados em amarelo. O grupo* ***HEME*** *aparece em laranja. Figura gerada com o uso do VMD.*



(a) **(b)** **(c)**

Figura 5.40 *– Representação do sítio de ligação de uma globina. Em (a) e (b) o anel aromático da PHE66- encontra-se sempre posicionado paralelo e apresenta forte interação com o grupo* ***HEME****. Em (b) e (c), a HIS proximal mostra forte interação com o átomo de ferro do grupo* ***HEME****. Figura gerada com o uso do VMD.*

Na figura 5.40, são mostradas diferentes vistas da interação do grupo **HEME** com a histidina e a fenilalanina conservadas em todas as globinas. A conservação do resíduo de PHE na posição CD1 é um fato bem descrito na literatura, havendo uma forte relação entre a ancoragem e ligação do grupo **HEME** [Dickerson e Geis (1983), Hargrove et al. (1994)]. Este resíduo compartilha com a HIS (F8) proximal uma conservação absoluta em todas as globinas conhecidas até o momento [Kapp et al. (1995)].

A análise do **locus 2**, devido às flutuações estruturais observadas nesta região da estrutura terciária da proteína, deve ser feita detidamente. A figura 5.41 apresenta duas vistas (frente e topo) do **locus 2**.

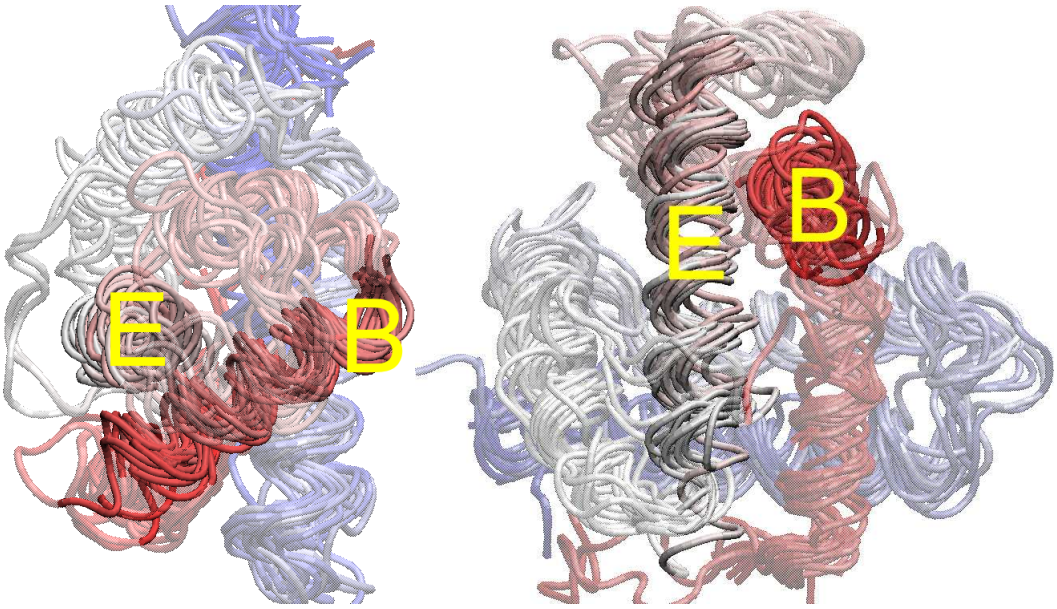


Figura 5.41 *– Vistas de topo (a) e frente (b) do* ***locus 2*** *visto no alinhamento estrutural conjunto das globinas estudadas. As hélices E e B encontram-se indicadas. Figura gerada com o uso do VMD.*

Na região onde a hélice **B** mais aproxima-se da hélice **E**, os "*hubs*" presentes neste *locus*, aparecem como "*hubs*" no conjunto de globinas estudadas com os índices apresentados na tabela 5.11.

Na figura 5.42, todas as posições identificadas para o **locus 2** aparecem mapeadas sobre o alinhamento estrutural conjunto das globinas estudadas.

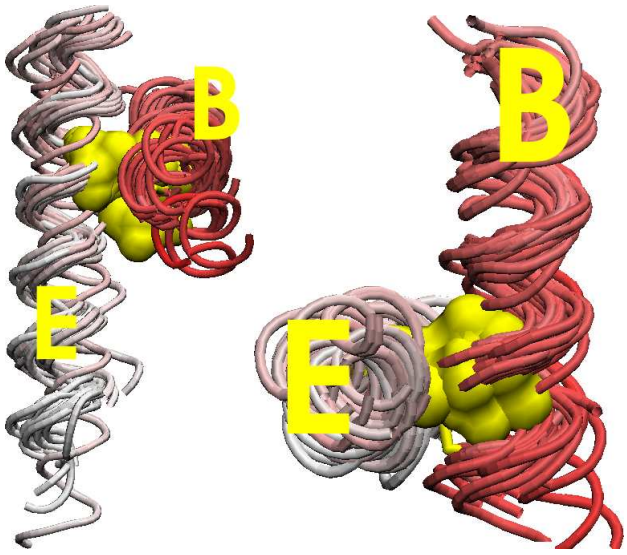
No **locus 3** existe a conservação da conexão das hélices A e H. A figura 5.43 apresenta uma visão desta região.

No **locus 4** existe a conservação da conexão das hélices F e H. A figura 5.44 apresenta uma visão desta região.

É interessante observar que, alguns dos resíduos identificados como "*hubs*", já haviam sido identificados como sendo os mais conservados em estudos de alinhamentos estruturais realizados anteriormente em globinas [Lesk e Chothia (1980), Dickerson e Geis (1983), Bash-

Resíduo 1	Ocorrência como ‘‘*hub*’’	Ocorrência nas proteínas estudadas	Resíduo 2	Ocorrência como ‘‘*hub*’’	Ocorrência nas proteínas estudadas
POS88-	60%	100%	POS43-	40%	100%
POS88-	60%	100%	POS39-	60%	100%
POS85-	50%	100%	POS36-	50%	100%
POS84-	30%	100%	POS44-	30%	100%
POS84-	30%	100%	POS40-	30%	100%
POS84-	30%	100%	POS36-	50%	100%
POS83-	30%	100%	POS44-	30%	100%
POS83-	30%	100%	POS43-	40%	100%
POS81-	50%	100%	POS36-	50%	100%
POS80-	30%	100%	POS44-	30%	100%
POS80-	30%	100%	POS40-	30%	100%
POS76-	60%	100%	POS44-	30%	100%

Tabela 5.11 *– Posições mais freqüentes responsáveis pela estabilidade do* ***locus 2****.*



(a) **(b)**

Figura 5.42 *– Vistas de topo (b) e frente (a) do* ***locus 2****. Observa-se a sobreposição dos resíduos nas posições “*hub*” para as globinas estudadas. Figura gerada com o uso do VMD.*

ford et al. (1987), Kapp et al. (1995), Ptitsyn e Ting (1999), Süel et al. (2002)]. Observa-se ainda a importância do grupo **HEME** para estrutura das globinas, já que ela também aparece como um “*hub*” em todas as proteínas estudadas. O alinhamento estrutural das globinas mostra a conservação da posição espacial deste grupo.

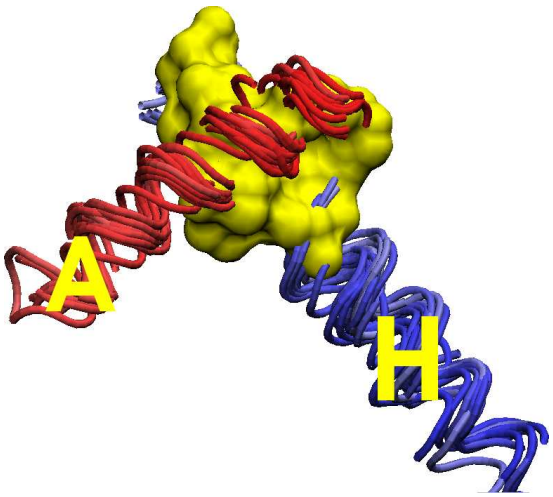


Figura 5.43 – *Vista do* ***locus 3****. As hélices A e H encontram-se indicadas. Figura gerada com o uso do VMD.*

Resíduo 1	Ocorrência como ‘‘*hub*’’	Ocorrência nas proteínas estudadas	Resíduo 2	Ocorrência como ‘‘*hub*’’	Ocorrência nas proteínas estudadas
POS112-	76,19%	50%	POS45+	95,24%	45,00%
POS112-	76,19%	50%	POS48+	95,24%	20,00%
POS109-	95,24%	40%	POS41+	90,48%	47,37%
POS109-	95,24%	40%	POS42+	95,24%	60,00%
POS108-	76,19%	25%	POS45+	95,24%	45,00%
POS108-	76,19%	25%	POS46+	100,00%	47,62%
POS108-	76,19%	25%	POS49+	95,24%	40,00%

Tabela 5.12 – *Posições mais freqüentes responsáveis pela estabilidade do* ***locus 3****.*

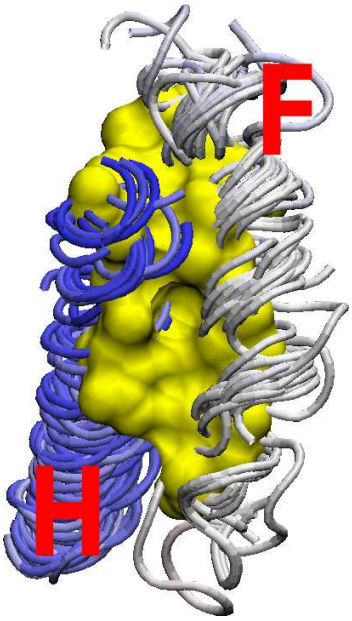


Figura 5.44 – *Vista do* ***locus 4****. As hélices F e H encontram-se indicadas. Figura gerada com o uso do VMD.*

Resíduo 1	Ocorrência como ‘‘*hub*’’	Ocorrência nas proteínas estudadas	Resíduo 2	Ocorrência como ‘‘*hub*’’	Ocorrência nas proteínas estudadas
POS58+	38,10%	100,00%	POS019-	62,50%	76,19%
POS58+	38,10%	100,00%	POS016-	100,00%	66,67%
POS58+	38,10%	100,00%	POS007-	44,44%	85,71%
POS59+	65,00%	95,24%	POS007-	44,44%	85,71%
POS59+	65,00%	95,24%	POS003-	57,14%	100,00%
POS63+	55,56%	85,71%	HIS0	100,00%	100,00%
POS63+	55,56%	85,71%	POS013+	57,14%	100,00%

Tabela 5.13 – *Posições mais freqüentes responsáveis pela estabilidade do* ***locus 4****.*

Da mesma forma, algumas moléculas de água (dados não mostrados) forma identificadas como sendo "*hub*", apresentando altos níveis de energia potencial. As figuras 5.45 e 5.46 mostram a superposição das moléculas de água da primeira camada de solvatação, para todas as globinas estudadas. Nas figuras 5.47 e 5.48 são mostradas a superposição das moléculas de água da primeira camada de solvatação para a globina PDBID 1A6G. A dificuldade em se mapear se estas águas mais energeticamente ligadas está na necessidade de proceder um mapeamento das coordenadas espaciais destas moléculas e avaliar a conservação espacial destas moléculas para todas as globinas. Tal procedimento exige um tratamento dos dados disponíveis, que vai além do escopo deste trabalho. Contudo, trabalhos já existentes [Levy e Onuchic (2004), Papoian et al. (2004)] mostram não só a conservação de moléculas de água na estrutura das proteínas, bem como a importância destas para a dinâmica das proteínas.

A título de exemplo, a figura 5.49 mostra, para a mioglobina PDBID 1A6G, uma das moléculas de água identificadas e a sua localização entre dois resíduos "*hub*" já identificados. Como apresentado por Papoian [Papoian et al. (2004)], as moléculas de água podem atuar estabelecendo ligações de "longa distância" entre resíduos hidrofílicos, o que pode representar um papel importante tanto para a estabilidade estrutural dos proteínas, como para os aspectos dinâmicos das mesmas.

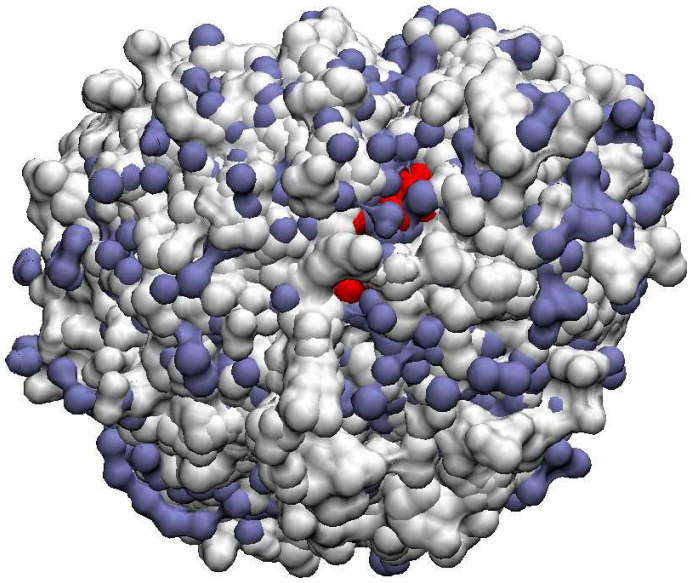


Figura 5.45 *– Vista da sobreposição de todas as águas de solvatação em conjunto com a sobreposição de todas as globinas estudadas. Os grupos* ***HEME*** *aparecem sobrepostos e em vermelho. Figura gerada com o uso do VMD.*

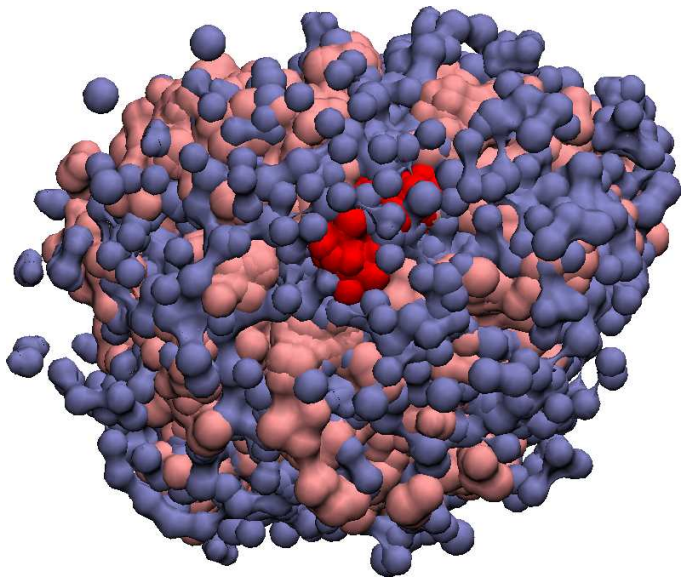


Figura 5.46 *– Vista da sobreposição de todas as águas de solvatação em conjunto com a sobreposição do núcleo hidrofóbico de as globinas estudadas. Percebe-se a formação de arranjos de água estruturada em torno do núcleo hidrofóbico. Os grupos* ***HEME*** *aparecem sobrepostos e em vermelho. Figura gerada com o uso do VMD.*

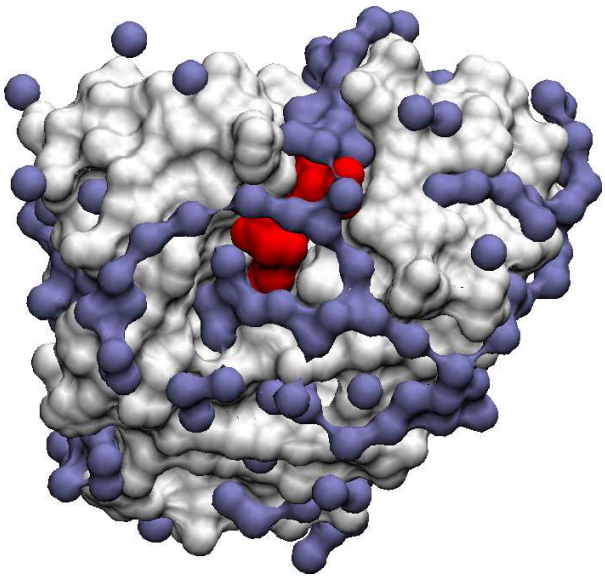


Figura 5.47 *– Vista da proteína PDBID 1A6G mostrando a distribuição das águas de solvatação sobre esta proteína. Percebe-se a formação de arranjos de água estruturada. O grupo* ***HEME*** *aparece em vermelho. Figura gerada com o uso do VMD.*

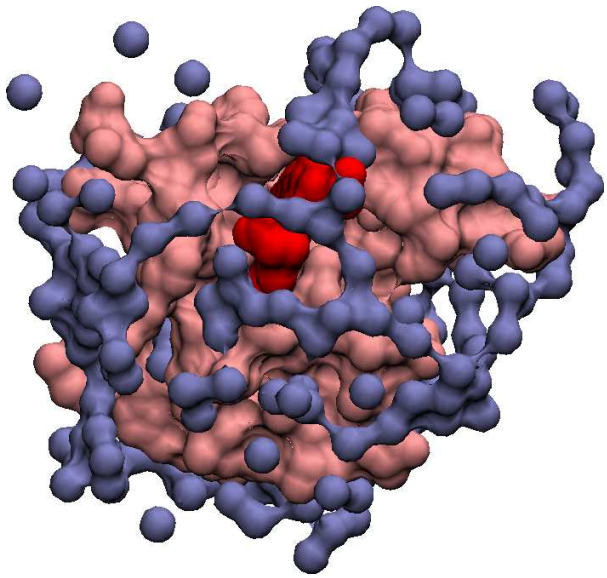


Figura 5.48 *– Vista da proteína PDBID 1A6G mostrando a distribuição das águas de solvatação sobre o núcleo hidrofóbico desta proteína. A presença dos arranjos de água estruturada mostra-se mais nítido. O grupo* ***HEME*** *aparece em vermelho. Figura gerada com o uso do VMD.*

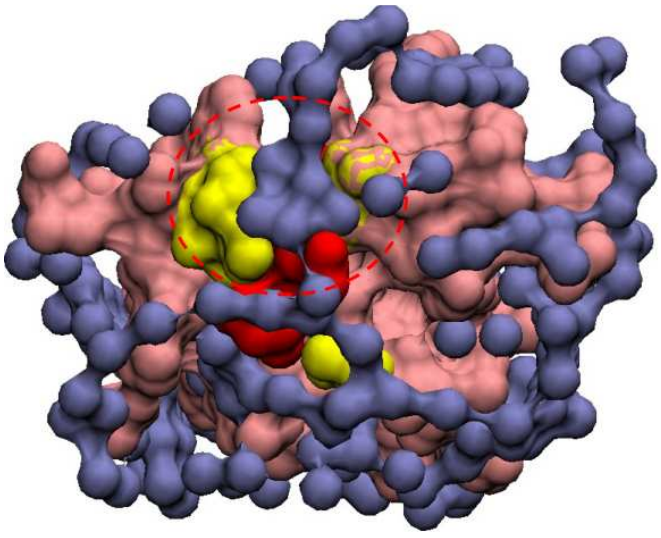


Figura 5.49 *– Na globina PDBID 1A6G, alguns resíduos "*hub*" do* ***locus 1*** *(em amarelo) têm contribuição de águas de solvatação que os mantém próximos. Figura gerada com o uso do VMD.*

5.4.1.1 Identificação das Propriedades Conservadas para as Posições mais Conectadas nas globinas

Para proceder esta análise, duas medidas de interesse estatístico (confiança e suporte) foram feitas para cada uma das posições existentes no alinhamento estrutural das globinas. A medida "suporte" é definida, neste caso, como o número de vezes, no alinhamento global, que uma posição do alinhamento apresenta um resíduo de aminoácido (ou não apresenta "gaps"). Já a medida "confiança" informa quantas vezes esta posição apareceu como "*hub*" em uma proteína, dado que ela não está vazia.

Para esta análise, foram selecionadas as posições "*hub*", com índice de suporte de 100% e confiança mínima de 50%. Para avaliar as propriedades, o método proposto por Lockless e Ranganathan [Lockless e Ranganathan (1999)], descrito na seção 4.5, foi adotado. Neste estudo, foi feita uma extensão da proposta original apresentada em [Lockless e Ranganathan (1999)], e detalhada na seção 4.5, onde não só a conservação dos resíduos é avaliada, mas também a conservação de outras propriedades a eles relacionadas como: o grau de hidrofobicidade, a polaridade do resíduo, o peso molecular, a natureza da cadeia lateral (se alifática, sulfúrica, não aromática com radical hidroxil, aromática, ácida ou básica).

Tabela 5.14 – *A análise das posições "*hub*", identificadas no alinhamento estrutural das globinas, com base no critério de ganho de informação revela quais as propriedades mais conservadas em cada uma destas posições. As diferentes cores ressaltam as propriedades mais significativas para cada posição: Em amarelo o nome do resíduo, em azul o grau de hidrofobicidade, em lilás a polaridade, em vermelho o peso molecular, em laranja a natureza da cadeia lateral. O volume molecular também foi um dos atributos avaliados, mas não mostrou relevância para qualquer posição das seqüências analisadas.*

	-132	-131	-130	-129	-128	-127	-126	-125	-124	-123	-122	-121	-120	-119	-118	-117	-116	-115	-114	-113	-112	-111	-110	-109	-108	-107	-106	-105	-104	-103	-102	-101	-100	-99	-98	-97
1A6G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	V	-	L	S	E	G	E	W	Q	L	V	L	-	H	V	W	A	-	K	V	-	E
1ASH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	N	K	T	R	E	L	C	M	-	K	S	L	E	-	H	A	K	V
1B0B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	L	S	A	A	Q	K	D	N	V	K	-	S	S	W	A	-	K	A	-	S
1BZP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	V	-	L	S	E	G	E	W	Q	L	V	L	-	H	V	W	A	-	K	V	-	E
1DLW	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	L	-	F	-	E	Q	L	-	-	G	G	-	Q
1DLY	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	L	-	F	-	A	K	L	-	-	G	G	-	R
1ECD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	L	S	A	D	Q	I	S	T	V	Q	-	A	S	F	D	-	K	V	-	K
1GDJ	-	-	-	-	-	-	-	-	-	-	-	-	-	G	A	-	L	T	E	S	Q	A	A	L	V	K	-	S	S	W	E	-	E	F	-	N
1HLM	G	-	A	T	Q	S	F	-	Q	S	V	-	-	-	G	D	L	T	P	A	E	K	D	L	I	R	-	S	T	W	D	-	Q	L	-	M
1JF3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	A	A	Q	R	Q	V	V	A	-	S	T	W	K	-	D	I	-	A
1JF4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	A	A	Q	R	Q	V	V	A	-	S	T	W	K	-	D	I	-	A
1KR7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M	-	V	-	-	-
1LHS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	D	D	E	W	N	H	V	L	-	G	I	W	A	-	K	V	-	E
1MBS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	D	G	E	W	H	L	V	L	-	N	V	W	G	K	V	E	-	-
1MYT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	D	F	D	A	V	L	-	K	C	W	G	-	P	V	-	E
1Q1F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	R	P	E	S	E	L	I	R	-	Q	S	W	R	-	V	V	-	S
1RTX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	T	-	L	Y	E	K	L	-	-	G	G	-	T
1UYX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	L	-	F	-	A	K	L	-	-	G	G	-	R
1V5H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	E	-	A	E	R	K	A	V	Q	-	A	M	W	A	-	R	L	-	Y
2FAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	-	L	S	A	A	E	A	D	L	A	G	-	K	S	W	A	-	P	V	-	F
2MM1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	L	S	D	G	E	W	Q	L	V	L	-	N	V	W	G	-	K	V	-	E
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
NR	1,02	0	1,63	0,98	1,07	1,04	0,98	0	1,07	1,04	1,16	0	0	1,02	6,41	1	15,8	15	3,54	5,57	15,3	11,8	8,61	15,6	18,1	5,71	1,15	4,47	6,44	48,5	3,87	1,12	6,36	8,86	1,12	6,34
HDR	1,02	0	1,63	1,66	1,01	1,66	0,98	0	1,01	1,66	1,63	0	0	1,02	6,32	1	15,8	13,8	3,57	5,41	14,8	7,84	6,96	15,7	18,3	5,84	1,15	2,59	3,54	35,5	3,96	1,12	6,4	9,3	1,12	6,05
PM	1,02	0	1,63	0,98	2,65	1,04	0,98	0	2,65	1,04	1,81	0	0	1,02	6,41	3,63	7,43	14,3	2,52	5,09	13,7	4,88	5,23	8,94	12,8	2,7	1,15	4,14	5,48	17,4	3,39	2,65	4,75	5,73	2,65	4,36
POL	4,25	0	10,6	4,25	4,25	4,25	10,6	0	4,25	4,25	10,6	0	0	4,25	5,7	1,52	2,13	10,2	4,32	1,79	10	4,34	3,34	5,04	4,31	2,19	10,6	4,54	5,01	3,31	1,9	2,2	4,76	4,11	2,2	2,68
CLT	8,17	0	8,17	1,7	2,48	1,14	1,7	0	2,48	1,14	8,17	0	0	8,17	2,04	2,48	2,96	14,3	3,62	3,2	16,6	5,93	11,4	7,11	6,01	2,41	1,7	3,2	4,67	24,1	2,11	2,2	4	10,7	2,2	4,34
VM	1,02	0	2,75	2,54	1,51	2,75	1,12	0	1,51	2,75	1,51	0	0	1,02	6,26	2,54	6,34	7,74	2,26	4,04	14,2	11,6	2,13	8,1	14,5	4,41	1,12	4,99	2,37	48,5	2,39	3,31	3,63	6,87	3,31	4,63

(a)

Continua . . .

Tabela 5.14 – *(continuação)*

	-98	-97	-96	-95	-94	-93	-92	-91	-90	-89	-88	-87	-86	-85	-84	-83	-82
1A6G	-	E	A	-	-	-	-	D	-	V	A	G	-	H	G	Q	D
1ASH	K	V	D	-	T	S	N	E	-	A	R	Q	-	D	G	I	D
1B0B	-	S	A	-	-	-	-	A	-	W	G	T	-	A	G	P	E
1BZP	-	E	A	-	-	-	-	D	-	V	A	G	-	H	G	Q	D
1DLW	-	Q	A	-	-	-	-	A	-	V	Q	A	-	V	T	A	Q
1DLY	-	R	E	-	-	-	-	A	-	V	E	A	-	A	V	D	K
1ECD	-	K	G	-	-	-	-	D	-	-	-	-	-	-	P	V	G
1GDJ	-	N	A	-	-	-	-	N	-	I	P	K	-	H	T	H	R
1HLM	-	M	T	-	-	-	-	H	-	R	T	G	-	F	V	A	D
1JF3	-	A	G	A	-	-	-	D	N	G	A	G	-	V	G	K	E
1JF4	-	A	G	S	-	-	-	D	N	G	A	G	-	V	G	K	E
1KR7	-	-	-	-	-	-	-	N	-	W	A	A	-	V	V	D	D
1LHS	-	E	P	-	-	-	-	D	-	L	S	A	-	H	G	Q	E
1MBS	-	-	T	-	-	-	-	D	-	L	A	G	-	H	G	Q	E
1MYT	-	E	A	-	-	-	-	D	-	Y	T	T	-	M	G	G	L
1Q1F	-	S	R	-	-	-	-	S	-	P	L	E	-	H	G	T	V
1RTX	-	T	T	-	-	-	-	A	-	V	D	L	-	A	V	D	K
1UVX	-	R	E	-	-	-	-	A	-	V	E	A	-	A	V	D	K
1V5H	-	Y	A	-	-	-	-	N	-	-	C	E	D	V	G	V	A
2FAL	-	F	A	-	-	-	-	N	-	K	N	A	-	N	G	L	D
2MM1	-	E	A	-	-	-	-	D	-	I	P	G	-	H	G	Q	E
	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NR	1,12	6,94	8,11	1	0,98	1,04	1,07	13,9	2,25	6,12	5,85	8,71	1	11,1	24,2	8,67	9,67
HDR	1,12	6,05	8,07	1,34	1,66	1,66	1,01	13,3	1,46	6,95	4,73	8,76	1	10,9	24,1	6,1	9,6
PM	2,65	4,96	7,7	1	0,98	1,04	3,63	9,11	2,54	7,27	4,54	8,12	3,63	10,3	24,1	4,29	6,84
POL	2,2	2,68	1,47	10,5	4,25	4,25	4,25	10,1	3,01	2,33	1,67	3,94	1,52	3,54	11,1	1,86	15
CLT	2,2	4,34	3,31	7,49	1,7	1,14	2,48	14,2	1,75	4,22	3,12	4,51	2,48	3,71	10,8	5,37	11,9
VM	3,91	4,63	4,77	1,92	2,54	2,75	2,54	11,8	1,79	4,25	4,22	7,94	2,54	3,08	24	5,29	7,18

	-81	-80	-79	-78	-77	-76	-75	-74	-73	-72	-71	-70	-69	-68	-67	-66	-65	-64	-63
1A6G	I	L	I	R	L	F	K	S	H	P	E	T	L	E	K	F	-	D	R
1ASH	L	Y	K	H	M	F	E	N	Y	P	P	L	R	K	Y	F	-	K	S
1B0B	F	F	M	A	L	F	D	A	H	D	D	V	F	A	K	F	S	G	L
1BZP	I	L	I	R	L	F	K	S	H	P	E	T	L	E	K	F	-	D	R
1DLW	F	Y	A	N	I	Q	A	D	-	A	T	V	A	T	F	F	N	G	I
1DLY	F	Y	N	K	I	V	A	D	-	P	T	V	S	T	Y	F	S	N	T
1ECD	I	L	Y	A	V	F	K	A	D	P	S	I	M	A	K	F	-	T	Q
1GDJ	F	F	I	L	V	L	E	I	A	P	A	A	K	D	L	F	-	S	F
1HLM	V	F	I	R	I	F	H	N	D	P	T	A	Q	R	K	F	-	P	Q
1JF3	C	L	S	K	F	I	S	A	H	P	E	M	A	A	V	F	-	G	-
1JF4	C	F	T	K	F	L	S	A	H	H	D	M	A	A	V	F	-	G	-
1KR7	F	Y	Q	E	L	F	K	A	H	P	E	Y	Q	N	K	F	-	-	G
1LHS	V	I	I	R	L	F	Q	L	H	P	E	T	Q	E	R	F	-	A	K
1MBS	V	L	I	R	L	F	K	S	H	P	E	T	L	E	K	F	-	D	K
1MYT	V	L	T	R	L	F	K	E	H	P	E	T	Q	K	L	F	-	P	K
1Q1F	L	F	A	R	L	F	A	L	E	P	S	L	L	P	L	F	-	Q	Y
1RTX	F	Y	E	R	V	L	Q	D	-	D	R	I	K	H	F	F	A	D	V
1UVX	F	Y	N	K	I	V	A	D	-	P	T	V	S	T	Y	F	S	N	T
1V5H	I	L	V	R	F	F	V	N	F	P	S	A	K	Q	Y	F	-	S	Q
2FAL	F	L	V	A	L	F	E	K	F	P	D	S	A	N	F	F	-	A	D
2MM1	V	L	I	R	L	F	K	G	H	P	E	T	L	E	K	F	-	D	R
	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
NR	12,2	14,2	11,8	20,5	11,2	30	7,28	5,18	19,1	47,2	11,9	7,54	5,78	6,39	9,94	62,3	1,77	5,58	5,88
HDR	13,4	14,1	8,31	20,5	11,4	29,9	7,24	4,61	19,1	47,2	12,2	5,07	4,54	6,05	9,93	62,3	1,13	4,79	4,56
PM	11,5	11,9	4,33	20,2	10,4	29,8	9,57	4,77	19	26,2	4,92	7,37	4,36	4,51	6,19	62,3	2,63	3,48	4,47
POL	9,98	15,6	2,78	18,1	15,6	12,5	4,76	2,26	8,11	8,07	13,4	2,99	1,29	3,89	5,56	15,6	9,14	5,27	3,68
CLT	5,48	10,9	3,38	17,8	8,85	18	4,46	3,59	8,94	47,4	9,64	4,49	2,76	5,53	6,64	43,7	6,77	4,03	3,27
VM	9,42	16	11	20,2	6,01	24,5	2,26	5,09	4,37	28,1	11,8	2,35	4,02	5,41	8,37	53,9	1,6	6,62	3,37

(b)

	-62	-61	-60	-59	-58	-57	-56	-55	-54	-53	-52	-51	-50	-49	-48	-47	-46	-45
1A6G	-	-	F	-	K	H	L	K	T	E	A	E	M	K	A	S	E	D
1ASH	-	-	R	-	E	E	Y	T	-	A	E	D	V	Q	N	D	P	F
1B0B	-	-	F	-	S	G	A	A	-	K	G	T	V	K	N	T	P	E
1BZP	-	-	F	-	K	H	L	K	T	E	A	E	M	K	A	S	E	D
1DLW	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D
1DLY	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D
1ECD	-	-	F	-	A	G	K	-	D	L	E	S	I	K	G	T	A	P
1GDJ	-	-	L	-	K	G	T	S	E	V	-	-	P	Q	N	N	P	E
1HLM	-	-	M	A	G	L	S	P	-	A	E	-	L	R	T	S	R	Q
1JF3	-	-	F	-	-	S	-	G	-	A	S	-	-	-	-	D	P	G
1JF4	-	-	F	-	-	S	-	G	-	A	S	-	-	-	-	D	P	G
1KR7	-	-	F	-	K	G	V	-	A	L	G	S	L	K	G	N	A	A
1LHS	-	-	F	-	K	N	L	T	T	I	D	A	L	K	S	S	E	E
1MBS	-	-	F	-	K	H	L	K	S	E	D	D	M	R	R	S	E	D
1MYT	-	-	F	-	A	G	I	A	-	Q	A	D	I	A	G	N	A	A
1Q1F	N	G	R	-	-	Q	F	S	S	P	E	D	S	L	S	S	P	E
1RTX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D
1UVX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D
1V5H	-	-	F	-	K	H	M	E	D	P	L	E	M	E	R	S	P	Q
2FAL	-	-	F	-	K	G	K	S	-	V	A	D	I	K	A	S	P	K
2MM1	-	-	F	-	K	H	L	K	S	E	D	E	M	K	A	S	E	D
	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89
NR	1,07	1,02	26,6	1,63	11,4	8,6	4,99	3,12	2,4	4,23	3,26	5,43	9,15	9,02	3,08	12,3	14,7	11,1
HDR	1,01	1,02	26,6	1,63	11,3	8,27	3,93	3,43	3,5	3,75	3,27	5,17	6,89	9,03	1,22	11,3	14,7	11,1
PM	3,63	1,02	26,5	1,63	6,64	7,9	3,75	1,62	2,23	1,47	1,14	2,05	2,05	7,87	2,09	12,1	5,42	3,96
POL	4,25	4,25	4,03	10,6	8,78	8,7	1,89	5,35	8,41	2,2	5,41	11,5	3,54	9,82	4,79	9,25	2,31	14,7
CLT	2,48	8,17	16,6	8,17	6,78	1,61	2,56	3,04	5,92	1,49	2,83	7,7	7,62	8,44	1,75	10,7	14,2	14,3
VM	2,54	1,02	21,5	2,75	3,85	6,2	3,05	1,99	1,88	3,21	3,23	3,31	2,8	3,54	1,55	5,6	5,96	5,75

	-44	-43	-42	-41	-40	-39	-38	-37	-36	-35	-34	-33	-32	-31	-30	-29	-28	-27
1A6G	L	K	K	H	G	V	T	V	L	T	A	L	G	A	I	L	-	-
1ASH	F	A	K	Q	G	Q	K	I	L	L	A	C	H	V	L	C	-	-
1B0B	M	A	A	Q	A	Q	S	F	K	G	L	V	S	N	W	V	-	-
1BZP	L	K	K	H	G	V	T	V	L	T	A	L	G	A	I	L	-	-
1DLW	M	P	N	Q	T	N	K	T	A	A	F	L	C	A	A	L	G	G
1DLY	M	K	V	Q	R	S	K	Q	F	A	F	L	A	Y	A	L	G	G
1ECD	F	E	T	H	A	N	R	I	V	G	F	F	S	K	I	I	-	-
1GDJ	L	Q	A	H	A	G	K	V	F	K	L	V	Y	E	A	A	-	-
1HLM	M	H	A	H	A	I	R	V	S	A	L	M	T	T	Y	I	-	-
1JF3	V	A	E	L	G	A	K	V	L	A	Q	I	G	V	A	V	-	-
1JF4	V	A	D	L	G	A	K	V	L	A	Q	I	G	V	A	V	-	-
1KR7	Y	K	T	Q	A	G	K	T	V	D	Y	I	N	A	A	I	-	-
1LHS	V	K	K	H	G	T	T	V	L	T	A	L	G	R	I	L	-	-
1MBS	L	R	K	H	G	N	T	V	L	T	A	L	G	G	I	L	-	-
1MYT	I	S	A	H	G	A	T	V	L	K	K	L	G	E	L	L	-	-
1Q1F	F	L	D	H	I	R	K	V	M	L	V	I	D	A	A	V	-	-
1RTX	M	A	K	Q	R	A	H	Q	K	A	F	L	T	Y	A	F	G	-
1UVX	M	K	V	Q	R	S	K	Q	F	A	F	L	A	Y	A	L	G	G
1V5H	L	R	K	H	A	C	R	V	M	G	A	L	N	T	V	V	-	-
2FAL	L	R	D	V	S	S	R	I	F	T	R	L	N	E	F	V	-	-
2MM1	L	K	K	H	G	A	T	V	L	T	A	L	G	G	I	L	-	-
	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107
NR	12,6	8,17	8,93	25,1	13,2	4,62	13,4	18	9,24	8,24	6,98	15,9	10,8	4,56	11,4	9,89	3,52	2,3
HDR	10	8,06	8,97	23,3	12,8	4,71	12	21,5	9,11	5,25	6,93	15,8	10,6	4,68	8,94	11,5	3,52	2,3
PM	3,08	5,75	4,53	22	13	3,65	9,27	12,6	4,63	7,85	6,56	14,9	10,5	3,66	8,03	7,69	3,52	2,3
POL	15,6	9,02	4,4	9,97	4,76	6	15,2	4,92	8,09	3,08	6,38	12,5	14	1,51	12,5	12,5	1,79	2,23
CLT	9,72	9,25	4,18	9,46	7,33	2,85	15,5	5,54	3,17	3,48	3,07	10,7	2,74	1,89	10,8	13,1	3,7	4,81
VM	8,33	4,14	3,16	9,92	12,8	4,39	5,28	22,5	8,15	2,88	5,27	8,99	10,3	2,59	9,36	4,65	3,52	2,3

(c)

	-26	-25	-24	-23	-22	-21	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9
1A6G	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-
1ASH	-	A	T	Y	-	D	D	R	E	T	F	-	-	-	-	N	A	-
1B0B	-	D	N	L	-	D	N	A	G	A	L	-	-	-	-	E	G	-
1BZP	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-
1DLW	P	N	-	-	-	-	-	-	-	A	W	T	G	R	-	-	-	-
1DLY	-	A	-	-	-	-	-	-	S	E	W	K	G	K	-	-	-	-
1ECD	-	G	E	L	-	-	-	P	N	-	I	-	-	-	-	E	A	-
1GDJ	-	I	Q	L	E	V	T	G	V	-	V	-	-	-	V	T	D	A
1HLM	-	D	E	M	-	D	T	-	E	V	L	-	-	-	-	P	E	-
1JF3	-	S	H	L	-	G	D	E	G	K	M	-	-	-	-	V	A	-
1JF4	-	S	H	L	-	G	D	E	G	K	M	-	-	-	-	V	A	-
1KR7	-	G	G	-	-	-	-	-	-	S	A	-	-	-	-	-	-	-
1LHS	-	K	Q	K	-	-	-	N	N	-	H	-	-	-	-	E	Q	-
1MBS	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-
1MYT	-	K	A	K	-	-	-	G	S	-	H	-	-	-	-	A	A	-
1Q1F	-	T	N	V	-	E	D	L	S	S	L	-	-	-	-	E	E	-
1RTX	-	G	-	-	-	-	-	T	D	K	Y	-	-	-	-	-	-	-
1UVX	-	A	-	-	-	-	-	-	S	E	W	K	G	K	-	-	-	-
1V5H	-	E	N	L	-	H	D	P	D	K	V	-	-	-	-	S	S	-
2FAL	-	N	N	A	-	A	N	A	G	K	M	-	-	-	-	S	A	-
2MM1	-	K	K	K	-	-	-	G	H	-	H	-	-	-	-	E	A	-
	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125
NR	1	5,86	5,51	6,54	0,98	2,45	4,64	6,73	4,61	3,93	9,8	0,9	2,3	0,9	1,16	11,9	10	1,63
HDR	1	5,29	6,58	6,49	0,98	2,29	4,58	6,6	4,17	3,41	11,8	1,06	2,3	0,9	1,63	11,8	9,94	1,63
PM	1,81	3,1	4,39	3,29	2,65	1,68	2,05	6,4	4,29	3,14	8,21	1,46	2,3	1,46	1,81	4,14	9,74	1,63
POL	10,6	4,23	8,22	2,23	1,52	7,18	2,54	3,61	8,43	5,4	2,22	3,37	2,23	1,43	10,6	7,29	1,41	10,6
CLT	1	2,49	5,84	2,9	2,48	2,99	3,06	1,73	2,42	3,68	2,86	1,33	4,81	1,43	8,17	6,27	2,59	8,17
VM	2,54	3,03	2,3	8,5	0,98	2,31	5,42	6,17	2,52	1,32	7,97	2,75	2,3	2,35	1,51	11,4	7,76	2,75

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
1A6G	E	L	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-
1ASH	Y	T	R	E	L	L	D	R	H	A	-	R	D	H	V	-	-	-
1B0B	Q	C	K	T	F	A	A	N	H	K	-	A	-	-	R	-	-	-
1BZP	E	L	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-
1DLW	N	-	-	-	L	K	E	V	H	A	-	N	-	-	M	-	-	-
1DLY	D	-	-	-	M	R	T	A	H	K	-	D	-	-	L	-	-	V
1ECD	D	V	N	T	F	V	A	S	H	K	-	P	-	-	R	-	-	-
1GDJ	T	L	K	N	L	G	S	V	H	V	-	S	-	-	K	-	-	-
1HLM	L	L	A	T	L	T	R	T	H	D	-	K	-	-	N	-	-	-
1JF3	E	M	K	A	V	G	V	R	H	K	-	G	Y	-	G	-	-	-
1JF4	E	M	K	A	V	G	V	R	H	K	-	G	Y	-	G	-	-	-
1KR7	D	A	A	G	L	A	S	R	H	K	-	G	-	-	R	-	-	-
1LHS	E	L	K	P	L	A	E	S	H	A	-	T	K	-	H	-	-	-
1MBS	E	L	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-
1MYT	I	L	K	P	L	A	N	S	H	A	-	T	K	-	H	-	-	-
1Q1F	Y	L	T	S	L	G	R	K	H	R	-	A	-	-	V	-	-	-
1RTX	D	G	R	Y	M	R	E	A	H	K	-	E	-	-	L	V	E	N
1UVX	D	-	-	-	M	R	T	A	H	K	-	D	-	-	L	-	-	V
1V5H	V	L	A	L	V	G	K	A	H	A	-	L	K	-	H	-	-	-
2FAL	M	L	S	Q	F	A	K	E	H	V	-	G	-	-	F	-	-	-
2MM1	E	I	K	P	L	A	Q	S	H	A	-	T	K	-	H	-	-	-
	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
NR	10,7	11,4	13,5	9,65	15,5	9	5,55	9,07	66,6	11,4	0	8,86	7,21	0,98	10,9	1,16	0,98	0,88
HDR	9,98	11	13,3	9,1	15,4	8,78	4,67	7,37	66,6	11,4	0	6,55	7,21	0,98	11	1,63	0,98	0,92
PM	5,36	6,45	6,57	4,14	6,94	8,54	4,75	8,41	66,6	8,17	0	7,79	3,19	0,98	10,4	1,81	2,65	2,72
POL	14,4	3,96	11,9	2,09	15,6	1,29	4,32	3,5	39,1	5,68	0	6,91	7,63	2,2	8,86	10,6	1,52	8,6
CLT	14,4	5,07	11,5	8,78	5,88	7,06	5,32	6,43	39,1	5,64	0	3,81	3,39	2,2	9,11	8,17	2,48	6,02
VM	9,29	7,26	3,91	7,19	11,7	5,94	2,92	8,38	29,8	5,11	0	6,98	1,83	3,91	6,98	1,51	0,98	1,69

(d)

Continua ...

Tabela 5.14 – *(continuação)*

	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1A6G	-	-	K	I	P	I	K	Y	L	E	F	I	S	E	A	I	I	H	V	L	H	S	R	H	P	-	G	N	F	G	A	D	A	Q	G	A
1ASH	-	-	H	M	P	P	E	V	W	T	D	F	W	K	L	F	E	E	Y	L	G	K	K	T	-	-	T	-	L	D	E	P	T	K	Q	A
1B0B	-	-	G	I	S	A	G	Q	L	E	A	A	F	K	V	L	A	G	F	M	K	S	Y	-	-	-	G	-	G	-	D	-	-	E	G	A
1BZP	-	-	K	I	P	I	K	Y	L	E	F	I	S	E	A	I	I	H	V	L	H	S	R	H	P	-	G	D	F	G	A	D	A	Q	G	A
1DLW	-	-	G	V	S	N	A	Q	F	T	T	V	I	G	H	L	R	S	A	L	T	G	A	-	-	-	G	-	V	A	A	A	L	V	E	Q
1DLY	P	-	H	L	S	D	V	H	F	G	A	V	A	R	H	L	S	D	T	L	T	E	L	-	-	-	G	-	V	P	P	E	D	I	T	D
1ECD	-	-	G	V	T	H	D	Q	L	N	N	F	R	A	G	F	V	S	Y	M	K	A	H	T	-	-	D	-	F	-	A	G	A	E	A	A
1GDJ	-	-	G	V	A	D	A	H	F	P	V	V	K	E	A	I	L	K	T	I	K	E	V	V	G	-	A	K	W	S	E	E	L	N	S	A
1HLM	-	-	H	V	G	K	K	N	Y	D	L	F	G	K	V	L	M	E	A	I	K	A	-	E	L	G	V	G	F	T	K	Q	V	H	D	A
1JF3	N	K	H	I	K	A	E	Y	F	E	P	L	G	A	S	L	L	S	A	M	E	H	R	I	G	-	G	K	M	N	A	A	A	K	D	A
1JF4	N	K	H	I	K	A	E	Y	F	E	P	L	G	A	S	L	L	S	A	M	E	H	R	I	G	-	G	K	M	N	A	A	A	K	D	A
1KR7	-	-	N	V	G	S	A	E	F	H	N	A	K	A	C	L	A	K	A	C	S	A	H	-	-	-	G	-	A	-	-	-	P	D	-	-
1LHS	-	-	K	I	P	V	K	Y	L	E	F	I	C	E	I	I	V	K	V	I	A	E	K	H	P	-	S	D	F	G	A	D	S	Q	A	A
1MBS	-	-	K	I	P	I	K	Y	L	E	F	I	S	E	A	I	I	H	V	L	H	S	K	H	P	-	A	E	F	G	A	D	A	Q	A	A
1MYT	-	-	K	I	P	I	N	N	F	K	L	I	S	E	V	L	V	K	V	M	H	E	K	A	-	-	G	-	L	D	A	G	G	Q	T	A
1Q1F	-	-	G	V	R	L	S	S	F	S	T	V	G	E	S	L	L	Y	M	L	E	K	S	L	G	-	P	D	F	T	P	A	T	R	T	A
1RTX	H	-	G	L	N	G	E	H	F	D	A	V	A	E	D	L	L	A	T	L	K	E	M	-	-	-	G	-	V	P	E	D	L	I	A	E
1UVX	P	-	H	L	S	D	V	H	F	Q	A	V	A	R	H	L	S	D	T	L	T	E	L	-	-	-	G	-	V	P	P	E	D	I	T	D
1Y5H	-	-	K	V	E	P	V	Y	F	K	I	L	S	G	V	I	L	E	V	V	A	E	E	F	A	-	S	D	F	P	P	E	T	Q	R	A
2FAL	-	-	G	V	G	S	A	Q	F	E	N	V	R	S	M	F	P	G	F	V	A	S	V	A	-	-	A	-	P	P	A	G	A	D	A	A
2MM1	-	-	K	I	P	V	K	Y	L	E	F	I	S	E	A	I	I	Q	V	L	Q	S	K	H	P	-	G	D	F	G	A	D	A	Q	G	A
	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179
NR	1,8	1,3	12,6	18,3	11,8	5,71	6,44	17,7	23,5	14,5	6,02	10,1	6,62	14	4,89	16,7	6,26	4,16	6,35	13,6	5,5	10,4	6,26	6,41	6,34	1,02	21,2	5,03	14,4	7,35	12,1	6,88	6,85	13	5,33	24,6
HDR	1,33	1,3	12,6	27,8	11,6	5,18	6,31	18,1	23,6	14,2	5,15	16	5,45	13,9	4,46	13,4	6,63	3,85	4,64	12,3	5,08	9,31	6,15	6,25	6,34	1,02	21,1	5,04	14,4	7,14	12,1	6,9	6,39	10,6	4,4	24,6
PM	1,81	1,86	10,3	8,42	5,04	4,29	6,49	17,1	23	11,6	4,57	4,66	6,19	3,46	2,32	20	4,8	4,5	4,57	8,67	6,12	7,11	5,23	5,31	3,35	1,02	21	1,93	14,3	4,55	11,3	1,54	6,3	3,73	4,54	24,5
POL	7,8	1,59	12,3	15,6	3,08	1,81	2,63	2,04	12,5	14,1	4,08	15,6	6,42	8,54	2,26	15,6	6,57	6,83	6,3	12,5	7,15	6,41	3,13	2,59	2,26	4,25	11,4	3,67	3,38	4,2	3,34	10,2	2,41	3,39	4,46	5,14
CLT	1,46	1,59	12,3	15,7	11,2	2,21	2,86	7,86	15,6	14,3	3,78	10,6	6,63	5,2	2,52	10,6	6,35	4,46	6,05	11	6,04	5,25	3,4	1,71	6,56	8,17	7,45	6,45	7,48	6,28	4,01	8,24	4,33	3,93	2,32	6,38
VM	2,65	2,75	11	17,4	5,45	4,26	3,26	8,8	21,4	13,9	5	9,03	5,71	13,8	4,32	11,7	4,61	2,39	4,54	11,9	3,77	3,83	5,39	2,78	3,14	1,02	20,9	2,81	11	8,69	7,32	3,66	5,06	6,45	4,45	18

(e)

	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
1A6G	M	N	K	A	L	E	L	-	-	F	R	K	D	I	A	A	K	Y	-	K	-	E	L	G	Y	-	-	-	-
1ASH	W	H	E	I	G	R	E	-	-	F	A	K	E	I	N	K	-	-	-	-	-	-	-	-	-	-	-	-	-
1B0B	W	T	A	V	A	G	A	-	-	L	M	G	M	I	R	P	D	-	-	-	-	-	-	-	M	-	-	-	-
1BZP	M	N	K	A	L	E	L	-	-	F	R	K	D	I	A	A	K	Y	-	K	-	E	L	G	Y	Q	G	-	-
1DLW	T	V	A	V	A	E	T	-	-	V	R	G	D	V	V	-	-	T	-	V	-	-	-	-	-	-	-	-	-
1DLY	A	M	A	V	V	A	S	-	-	T	R	T	E	V	L	-	-	N	-	M	-	P	-	-	-	-	-	-	Q
1ECD	W	G	A	T	L	D	T	-	-	F	F	G	M	I	F	S	K	M	-	-	-	-	-	-	-	-	-	-	-
1GDJ	W	T	I	A	Y	D	E	-	-	L	A	I	V	I	K	K	E	M	-	D	-	D	A	A	-	-	-	-	-
1HLM	W	A	K	T	F	A	I	-	-	V	Q	G	V	-	L	I	T	K	-	H	-	A	-	S	-	-	-	-	-
1JF3	W	A	A	A	Y	G	D	-	-	I	S	G	A	L	I	S	G	L	-	Q	-	S	-	-	-	-	-	-	-
1JF4	W	A	A	A	Y	A	D	-	-	I	S	G	A	L	I	S	G	L	-	Q	-	S	-	-	-	-	-	-	-
1KR7	L	G	-	-	-	-	H	-	-	A	I	D	D	I	L	S	H	L	-	-	-	-	-	-	-	-	-	-	-
1LHS	M	K	K	A	L	E	L	-	-	F	R	N	D	M	A	S	K	Y	-	K	-	E	F	G	F	-	Q	G	-
1MBS	M	K	K	A	L	E	L	-	-	F	R	N	D	I	A	A	K	Y	-	K	E	L	G	F	H	-	G	-	-
1MYT	L	R	N	V	M	G	I	-	-	I	I	A	D	L	E	A	N	Y	-	K	-	E	L	G	F	-	S	G	-
1Q1F	W	S	R	L	Y	G	A	-	-	V	V	Q	A	M	S	R	G	W	D	G	-	-	-	-	-	-	-	-	-
1RTX	V	A	A	V	A	G	A	P	A	H	K	R	D	V	L	-	-	N	-	Q	-	-	-	-	-	-	-	-	-
1UVX	A	M	A	V	V	A	S	-	-	T	R	T	E	V	L	-	-	N	-	M	-	P	-	-	-	-	-	-	Q
1Y5H	W	A	K	L	R	G	L	-	-	I	Y	S	H	V	T	A	A	Y	-	K	-	E	V	G	W	-	-	-	-
2FAL	W	T	K	L	F	G	L	-	-	I	I	D	A	L	K	A	A	-	-	-	-	-	-	G	A	-	-	-	-
2MM1	M	N	K	A	L	E	L	-	-	F	R	K	D	M	A	S	N	Y	-	K	-	E	L	G	F	-	Q	G	-
	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208
NR	31,1	4,33	10,8	7,25	6,84	10,2	5,48	1	1,63	10,7	14,7	7,17	12,3	14	5,88	6,87	4,78	14	1	7,77	0,98	7,2	2,43	8,58	2,56	1,07	1,53	2,3	2,24
HDR	33,6	3,73	10,7	6,84	6,89	10,2	5,43	1	1,63	10,7	14,6	6,42	12,3	16,1	4,31	4,81	4,6	13,8	1	7,37	0,98	7,21	2,29	8,51	2,35	1,01	0,75	2,3	1,46
PM	10,5	4,15	7,32	5,34	5,8	8,01	5,33	1,81	1,63	9,2	14,2	6,41	3,75	7,02	4,3	6,45	3,06	13,7	3,63	9,87	2,65	2,07	1,86	8,58	2,36	2,65	1,22	2,3	1,86
POL	4,61	2,62	5,82	8,04	8,14	7,08	1,84	10,6	10,6	8,03	5,43	8,5	14,3	12,5	3,29	1,55	6,67	1,89	1,52	6,63	1,52	6,34	3,27	7,18	3,1	4,25	1,63	2,23	3,01
CLT	14,6	2,17	5,5	10,7	4	3,33	2,94	1	8,17	7,16	5,98	2,43	3,73	8,91	3,97	4,98	2,72	6,44	2,48	5,91	2,48	5,19	1,82	2,01	6,5	2,48	5,43	4,81	1,75
VM	23,8	3,14	4,35	4,62	5,52	10,2	2,01	2,54	2,75	8,48	14	6,52	4,74	13,6	4,14	9,5	2,07	6,66	2,54	4,43	0,98	6,38	1,35	8,44	3,91	1,51	1,3	2,3	1,23

(f)

Os resultados da aplicação deste método de análise de ganho de informação, são mostrados na tabela 5.14. Nesta tabela cada uma das propriedades estudadas é representada por uma cor distinta: em amarelo estão ressaltadas as posições onde um resíduo específico é significativamente conservado; em azul aparecem as posições onde o grau de hidrofobicidade é a propriedade que acarreta maior ganho de informação; em lilás está indicado onde a polaridade do resíduo é mais relevante; em vermelho as posições onde o peso molecular é mais significativo e; em laranja a natureza da cadeia lateral. Cada uma das posições ressaltadas na tabela 5.14 mostra, para cada posição "*hub*" com maior significância estatística no conjunto do alinhamento estrutural das globinas, as propriedades conservadas que apresentam o maior ganho de informação[13] em cada posição.

Das globinas estudadas, os resíduos **HIS0** e **PHE66-** são os únicos conservados em todas elas, o que é muito bem descrito na literatura. Por outro lado, em outras posições os atributos que parecem ser conservados, de uma forma geral, nas globinas não estão restritos à conservação dos mesmos resíduos em cada posição. Na tabela 5.14(a), a posição **RES104-**, que participa do **locus 2**, parece conservar preferencialmente o valor de hidrofobicidade. Ao mesmo tempo, a posição **RES83-** equilibra entre a tendência de conservar o resíduo de glu-

[13] O termo "ganho de informação", no contexto deste trabalho é apresentado na seção 4.5.

tamina - Q, ou manter o grau de hidrofobicidade local. Já na posição **RES82-**, prevalece a conservação do grau de polaridade.

Mas se tais atributos mostram-se conservados para as globinas analisadas, eles seriam conservados também para as globinas mutantes que se mantiveram estáveis o suficiente para serem cristalografadas? Para isto, um conjunto de 159 mutantes de mioglobina de cachalote – PDBID 1A6G, identificados no PDB [Berman et al. (2000)], foram estudados seguindo o mesmo algoritmo de análise de ganho de informação. O resultado do alinhamento estrutural deste conjunto é mostrado na figura 5.50.

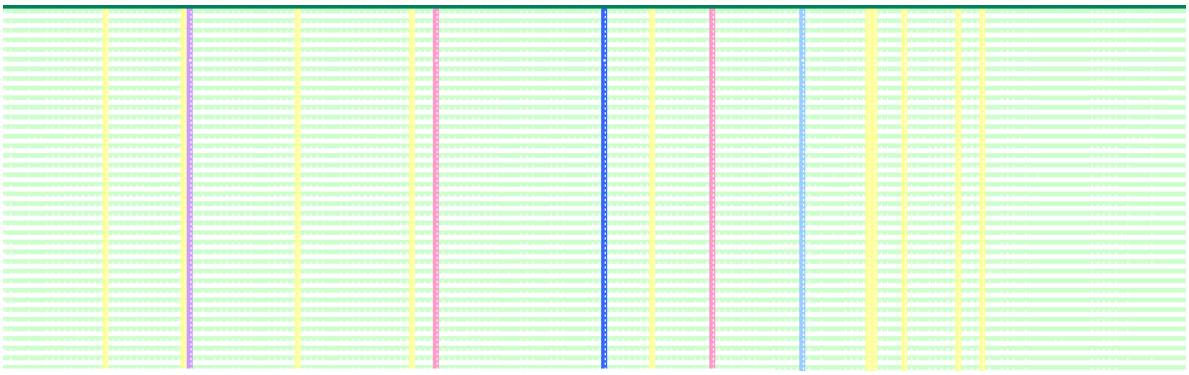


Figura 5.50 *– Visão de todas as posições de parte dos 159 mutantes da mioglobina PDBID 1A6G. Mesmo que as seqüências não sejam distinguíveis, a figura mostra que a análise do ganho de informação feito sobre o alinhamento estrutural do conjunto de mutantes em conjunto com a mioglobina nativa, revela que as propriedades das posições "*hub*", são conservadas em todos os mutantes dos quais foi possível obter algum cristal viável. As cores indicam as propriedades conservadas em cada posição: em amarelo o nome do resíduo, em azul o grau de hidrofobicidade, em lilás a polaridade, em vermelho o peso molecular, em laranja a natureza da cadeia lateral.*

Obviamente o resultado do alinhamento estrutural de 159 globinas é impossível de ser visto com alto grau de acuidade. Contudo, a figura 5.50 mostra que, mesmo para esta gama diversificada de mutantes da mesma mioglobina, a formação de estruturas estáveis demandou a conservação de certos atributos. Identificou-se que o resíduo **HIS0** é conservado em 153 mutantes, enquanto o resíduo **PHE66-** é conservado em 158 mutantes. Ao mesmo tempo, posição **RES82-** mantém-se ambíguo entre a tendência de conservar o resíduo de glutamina - Q, ou manter o grau de hidrofobicidade local. As demais posições analisadas, mesmo apresentando baixa variação de especificidade dos resíduos, conservam significativamente as tendências apresentadas na análise conduzida com as representantes de toda a família.

É interessante observar que, para a quase totalidade dos mutantes de mioglobina de cachalote estudados, foi necessária a conservação dos resíduos determinantes do sítio ativo. As demais propriedades, necessárias à formação da rede de interações não covalente mínima característica das globinas, parecem estar conservadas a despeito das mutações. Tudo leva a crer que a resilência estrutural das proteínas, no que concerne às mutações dos resíduos, se deve à manutenção de propriedades mais gerais compartilhadas por um grupo maior de aminoácidos, que podem ser permutados.

A existência de propriedades pontuais (resíduos "*hub*" e propriedades físico-químicas dos

aminoácidos) e topológicas (existência de "*locus*" de forte interação intra-cadeia) significativamente conservadas nas globinas estudadas, deixa acreditar que o processo de evolução das globinas levou à seleção de padrões que permitem coadunar a necessidade de manter a estabilidade e os atributos dinâmicos destas proteínas, com a liberdade necessária para promover mutações que permitam a evolução das mesmas proteínas.

Capítulo 6

Conclusões e Perspectivas

Nos sistemas vivos, as proteínas são as operárias por excelência, exercendo funções específicas em todos os processos biológicos. A ampla gama de funções exercidas por elas responde pela importância que as mesmas têm para a sustentação da vida. Dado a esta relevada importância, o domínio dos conhecimentos que facultem a regulação de sua expressão e atividade, abre possibilidades de que as mesmas sejam objeto de novas drogas, ou agentes reguladores capazes de induzir novas propriedades em plantas ou animais.

Ao fundamentar a visão das proteínas na perspectiva sistêmica, este trabalho chama a atenção para a propriedade desta outra visão das proteínas – proteínas como sistemas. Acreditamos que esta visão representa um novo paradigma capaz de por em evidência aspectos relativos às proteínas que não são claramente percebidos. Esperamos que esta abordagem amplie a forma de observar e estudar as proteínas.

Ressalta-se desde já que a versão atual da abordagem sistêmica das proteínas aqui apresentada, é incipiente e certamente não resolve, ainda, muitos dos problemas relacionados às mesmas. Cônscios disto, não há neste momento a pretensão de apresentar soluções perfeitas aos problemas aqui estudados. Ao contrário, muito ainda está por ser feito, tanto do ponto de vista teórico, quanto da extensão das análises aqui conduzidas, para outras famílias topológicas com o intuito de avaliar a generalidade das observações aqui relatadas.

Contudo, a visão das proteínas como sistema abre perspectivas de análise das proteínas, em quatro diferentes níveis, relacionados às propriedades fundamentais dos sistemas:

Estrutura – que no caso específico das proteínas, inclui os átomos que constituem os resíduos de aminoácidos e das moléculas de água, e a rede de interações covalentes e não covalentes que une estes átomos;

Dinâmica – que deriva dos mecanismos pelos quais as interações atômicas modulam as propriedades físicas da estrutura protéica, e levam à emergência do comportamento sistêmico da proteína ao longo do tempo, o qual é contingenciado tanto pelos diferentes fatores relacionados ao ambiente onde a proteína encontra-se, quanto pelas características estruturais desta proteína;

Controle – que é feito por diferentes "mecanismos" que a todo momento regulam os estados

da proteína, modulando seu comportamento.

Neste trabalho, a atenção concentrou-se nos estudos propedêuticos relacionados à análise estatística e espectral da estrutura de uma amostra representativa das proteínas de duas diferentes famílias topológicas – globinas e serinoproteases. Focar o estudo na identificação de caracterização da estrutura das proteínas, emerge naturalmente como o primeiro desafio no estudo metódico das proteínas quando vistas como sistemas complexos. Diante disto, a adoção dos modelos de redes complexas, para estudo da estrutura das proteínas, aparece como a alternativa apropriada já que estes modelos vêm contribuindo para o entendimento de diferentes fenômenos sistêmicos observáveis no mundo real. Em outras áreas do conhecimento, a aplicação destes modelos tem sido objeto de atenção, na medida em que tal aplicação vem auxiliando no entendimento dos sistemas reais.

Apesar dos avanços observados nas técnicas de análise protéica, muito do comportamento das proteínas ainda não pode ser diretamente observado. Tais dificuldades têm levado ao uso de modelos computacionais das proteínas como alternativa para o entendimento dos fenômenos de elevada complexidade. A despeito da sofisticação e da capacidade apresentada pelos métodos correntes de simulação computacional da dinâmica molecular, tais métodos demandam muito esforço computacional. Face a presente tecnologia computacional, estes métodos não se mostram apropriados para mimetizar os fenômenos de "longo"período[1] típicos das proteínas, e nem para prever eventuais comportamentos ainda não observados para as mesmas.

Neste trabalho, buscamos apresentar contribuições no sentido do aprimoramento de modelos de proteínas que sejam computacionalmente mais eficientes, sem perda da qualidade dos resultados. Aqui, o estudo da estrutura das proteínas, com o apoio dos modelos de redes complexas, demandou ainda uma abordagem multi e transdisciplinar, onde conhecimentos oriundos da física, da matemática, da biologia molecular e da ciência da computação, foram aplicados na análise e caracterização das redes de interações não covalentes que estabilizam a estrutura terciária das proteínas estudadas.

Estimar os atributos estruturais de uma proteína que respondam pela manutenção da estabilidade, e do seu comportamento alostérico, exigiu a determinação correta das relações que os constituintes das proteínas estabelecem entre si. Assim, este estudo centrou na identificação e caracterização da rede formada pelas interações não-covalentes existentes entre os átomos/resíduos que constituem as proteínas. Estas interações de natureza não covalente, são relevantes pois influenciam na estabilidade tridimensional da proteína após esta ter se enovelado. Neste trabalho, as proteínas foram inicialmente vistas como sendo uma malha de interações entre átomos, deixando a abordagem no nível dos resíduos de aminoácidos, para um momento posterior.

A determinação das interações não-covalentes entre os átomos, nas proteínas, levou em consideração não só a distância euclidiana entre os pares de átomos, mas também a existência de oclusão entre os átomos e a necessária "solvatação" das proteínas, o que evitou o trata-

[1] Neste contexto, fenômenos observáveis em períodos iguais ou maiores que 1 nanosegundo são considerados longos.

mento de interações entre os átomos das cadeias laterais, dos resíduos expostos na superfície da proteína, que não ocorrem na realidade. Outro aspecto da análise das interações não-covalentes entre os átomos, foi a determinação da energia potencial associada a cada uma delas, considerando os potenciais de Coulomb e de Lennard-Jones. Isto permitiu associar um atributo físico significativo a cada uma destas interações. Esta abordagem emprestou aos modelos uma semântica física mais realista, diferindo conceitualmente das abordagens adotadas nos trabalhos por nós identificados até agosto de 2007.

Enquanto os trabalhos correntes vêm adotando diferentes limiares arbitrários, para as distâncias entre os átomos, sem considerar as eventuais interferências estéricas entre estes, a abordagem aqui apresentada inova ao não estipular nenhum limiar para estas distâncias. Como dito anteriormente, as interações entre átomos, aqui identificadas, emergem naturalmente da metódica observação de critérios físicos fundamentados. Contudo, estamos cientes das limitações do método de análise de oclusão, apresentado neste trabalho. Tais limitações devem ser objeto de futuros trabalhos, onde este método deve ser aprimorado.

A análise das distribuições de freqüências do número de contatos por átomo, tanto para as globinas quanto para as serinoproteases, mostrou que estas distribuições apresentam dois regimes de comportamento bem distintos. Observamos que, para ambas famílias, existe um limiar de distância entre átomos dentro do qual o fenômeno mais relevante é o do empacotamento dos átomos ("*atomic packing*"), o qual se mostra mais acentuado na região do núcleo hidrofóbico das proteínas, o que ainda não havia sido relatado. Tais resultados permitem acreditar que o núcleo hidrofóbico das proteínas apresenta um grau de compactação equivalente ao máximo alcançável por qualquer sistema de partículas nas mesmas condições.

A constatação da existência de um arranjo hierarquizado dos átomos e interações dentro da estrutura das proteínas, permitiu concluir que a rede de interações não-covalentes subjacente à estrutura 3D das proteínas deve apresentar um arranjo hierárquico dos átomos/resíduos, onde aqueles mais energeticamente conectados devem exercem um papel importante para a estabilidade desta rede. Ao mesmo tempo, espera-se que estas redes apresentem um componente gigante que garanta a percolação, por toda a estrutura, dos sinais percebidos por qualquer um dos átomos/resíduos da estrutura, sendo que os átomos/resíduos mais fortemente conectados devam atuar como os principais difusores destes impulsos por toda a estrutura das proteínas.

Estas inferências, abrem outras perspectivas de trabalhos futuros, onde a estrutura das proteínas de outras famílias topológicas poderiam ser analisadas da mesma forma, o que pode gerar mais resultados que dariam melhor suporte para a generalização destas conclusões iniciais.

Ao mesmo tempo, constatou-se que tanto para as globinas quanto para as serinoproteases, a distância média entre os átomos e os índices de aglomeração apresentados por estas proteínas são muito similares aos observados em outros sistemas observados no mundo real. A constatação da existência desta hierarquia entre os átomos na estrutura das proteínas permite iniciar uma discussão acerca da resilência da rede de interações não-covalentes. Conjectura-se que mutações em uma proteína, que alterassem a conectividade dos átomos/resíduos com

maior número de interações não-covalentes, poderiam gerar estruturas instáveis ou mesmo desnaturadas. Porém, tal não ocorre.

Ao que tudo indica o processo evolutivo das proteínas selecionou outros artifícios que atenuam os impactos destas possíveis mutações.

A análise espectral das proteínas estudadas permitiu observar indícios relacionados aos processos de difusão e arranjo hierárquico dos átomos/resíduos na estrutura das globinas e serinoproteases. O espectro das proteínas analisadas mostrou a existência de um pequeno número de átomos/resíduos que determina os processos de percolação (ou difusão) de impulsos através da estrutura destas proteínas. Tudo leva a crer que a percolação de impulsos pela estrutura das proteínas deve ocorrer ao longo uma estrutura organizada de forma hierárquica, seguindo um percurso similar a uma "árvore", onde estes átomos/resíduos relevantes devem atuar como elementos da raiz desta "árvore". Ao mesmo tempo, o número de passos entre os átomos da "raiz" até os demais átomos seria equivalente ao diâmetro médio em passos – $\langle L \rangle$, apresentado por cada uma das famílias topológicas estudadas.

Os perfis de distribuição dos espectros sugerem que a estrutura das proteínas deve apresentar forte caráter associativo. O pronunciado perfil dos espectros sugere que a densidade de interações não covalente entre átomos, necessária para a emergência do grupo-núcleo ("*core group*") das proteínas é proporcionalmente muito baixa. Se de fato todas as proteínas apresentam estruturas fortemente associativas, é de se esperar que os átomos/resíduos mais energeticamente conectados estejam preferencialmente ligados a outros átomos/resíduos com padrões similares de conectividade.

A análise propedêutica feita neste trabalho, permitiu identificar quais os átomos/resíduos estariam relacionados ao eventual caráter associativo apresentado pelas proteínas. Tanto para as globinas como para as serinoproteases, os resíduos identificados como formadores dos grupos-núcleo, são citados na literatura, pela relevância dos mesmos para a função das proteínas destas famílias. A constatação da existência de grupos núcleo na estrutura das proteínas permite conjecturar que a percolação de informação nas proteínas, deve ocorrer de forma muito rápida. Ao mesmo tempo foi possível observar, para todas as proteínas estudadas, que estes grupos são restritos e que a densidade de interações é grande, sendo que estes grupos núcleo não se estendem pela estrutura das proteínas, estando confinados ao núcleo das redes.

Como estes padrões são típicos de uma rede com elevada associabilidade, é de se esperar que também para estas proteínas, estes núcleos devam prover robustez às estruturas das proteínas, ao concentrar todos os resíduos estruturalmente vitais, em uma região restrita destas proteínas. Assim, é possível que este arranjo estrutural seja aquele que foi evolutivamente selecionado, capaz de compatibilizar a flexibilidade necessária para permitir a evolução destes grupos-núcleo, com a necessidade da manutenção da estabilidade estrutural das proteínas. Isto porque uma mutação destes resíduos não deve ser suficiente para comprometer a conectividade estrutural das proteínas. Pode-se conjecturar que estas mutações não devam comprometer significativamente a capacidade percolação de informação pela estrutura das proteínas.

Ao mesmo tempo, os espectros das proteínas estudadas sugerem que as proteínas devem apresentar forte tendência à instabilidade. Estima-se também que as perturbações ambientais devam surtir pouco efeito na percolação de informação através da estrutura das proteínas, mas que estas devem ser capazes de dar respostas rápidas aos impulsos percebidos pelos grupos-núcleo, saindo com muita facilidade, de uma configuração para outra mais adequada ao novo contexto a que a proteína está exposta.

Os resultados da análise espectral mostram-se particularmente importantes para o entendimento dos fenômenos alostéricos apresentados pelas proteínas. As perturbações causadas por fatores como a entrada de um ligante em um sítio funcional afetam outros sítios distantes, e desta forma regulam a afinidade e a atividade desta proteína. Apesar de métodos experimentais já terem mostrado alguns padrões associados à comunicação alostérica, o entendimento dos princípios gerais de transmissão de informação entre sítios funcionais distantes permanece como um desafio. Os indícios aqui identificados sugerem a existência de resíduos chaves, nas proteínas, responsáveis pela geração e transmissão de tais sinais, ao mesmo tempo em que a topologia destas redes a transmissão destes sinais deve ocorrer seguindo certos atalhos entre as diferentes regiões das estruturas.

Os nodos ("*hub*"), das proteínas, permitiram identificar regiões do espaço topologicamente conservadas e importantes para todas as globinas estudadas. Os "*hubs*" em uma proteína, encontram-se localizados preferencialmente na mesma posição topológica, quando enoveladas, para todas as proteínas de uma mesma família. Entretanto, este posicionamento espacial dos átomos/resíduos ("*hub*") em uma proteína, não ocorre de forma determinística. Ao contrário, cada um destes ("*hubs*") costuma estar localizado nas imediações de um ponto médio no espaço comum a todas as estruturas de uma mesma família, quando alinhadas. Visto que mesmo as proteínas de uma mesma família topológica geralmente apresentam variações no número de resíduos, a localização espacial de uma posição "*hub*" conservada, acaba podendo variar dentro de uma vizinhança estreita. Observa-se ainda a importância do grupo HEME para estrutura das globinas, já que ela também aparece como um "*hub*" em todas as proteínas estudadas.

Por fim, foi feita uma análise estatística, nas globinas, para identificar as propriedades conservadas em outras posições mais conectadas nas globinas, nas quais o resíduo não se mantém conservado entre as diferentes globinas. Desta análise foi possível perceber que, em outras posições da seqüência primária das globinas, diferentes atributos parecem ser, de forma geral, bem conservados nas globinas, sendo que esta conservação de atributos não estão restritos à conservação dos mesmos resíduos em cada posição.

A existência de propriedades pontuais (resíduos "*hub*" e atributos físico-químicos dos aminoácidos) e topológicas (existência de "*locus*" de forte interação intracadeia) significativamente conservadas nas globinas estudadas, deixa acreditar que o processo de evolução das globinas levou à seleção de padrões que permitem coadunar a necessidade de manter a estabilidade e os atributos dinâmicos destas globinas, com a liberdade necessária para promover mutações que permitam a evolução das mesmas. Conjetura-se que estes achados relativos às globinas, sejam extensíveis às outras famílias topológicas existentes. Contudo tal hipótese só

poderá ser confirmada ou não, em futuros trabalhos onde a análise aqui apresentada deverá ser feita para outras famílias topológicas.

Ao final deste trabalho, acreditamos ter mostrado que existe uma rede de interações não-covalentes comum às globinas, e estimamos que tal propriedade deva ser comum às demais famílias topológicas, onde cada uma deve apresentar padrões estruturais próprios. Como conseqüência, abre-se a perspectiva de investir futuramente na determinação de modelos estruturais típicos para cada família topológica. Ao mesmo tempo, futuros estudos podem ser direcionados à simulação dos aspectos dinâmicos das proteínas. O valor destes conhecimentos reside na possibilidade de que os mesmos possam contribuir futuramente para a melhoria do entendimento dos fenômenos funcionais das proteínas, que faculte o projeto e uso racional deste admirável material para os mais diversos fins.

Esperamos que este estudo possa contribuir para o entendimento das características estruturais e dinâmicas das proteínas. Ao observar as proteínas sob a óptica sistêmica e adotando um novo método de análise e identificação das interações não-covalentes entre os átomos no seio das proteínas, identificamos as interações mais plausíveis e fizemos uma estimativa quantitativa da energia potencial inerente a estas interações. Observadas em conjunto, estas interações mostram a existência de uma rede entre os elementos (átomos/resíduos) constituintes das proteínas estudadas, onde esses elementos apresentam uma hierarquia baseada na conectividade e na energia das interações. Tais redes hierárquicas apresentam propriedades notáveis como alto grau de aglomeração, pequena distância média entre vértices, alta resilência a mutações. Ao mesmo tempo, tudo sugere que as redes subjacentes a todas as proteínas apresentem as mesmas propriedades estruturais e a mesma capacidade de transmissão de sinais alostéricos característicos a cada família, onde o fluxo da informação mostra ser direcionado, sendo os sítios de ligação os pontos de deflagração destes sinais.

Visto que estudos desta ordem ainda constituem uma novidade no estudo das proteínas, acreditamos que os problemas relativos à estabilidade estrutural e à dinâmica das mesmas são tópicos que ainda tem muito a ser investigado e para os quais respostas mais satisfatórias devem ser identificadas. Os achados apresentados neste trabalho sugerem que a visão das proteínas como sistemas, aqui proposta, constitui um paradigma pertinente. Em conjunto estes achados demonstram que as proteínas apresentam comportamentos similares aos observados em outros sistemas complexos encontrados no mundo real. Ao mesmo tempo, esses resultados contribuem para demonstrar a pertinência da elaboração, e uso de modelos formais para as proteínas. Potencialmente, estes modelos poderiam auxiliar os avanços futuros para aprimorar o conhecimento das proteínas e no avanço do uso destas para a melhoria da qualidade de vida da humanidade.

Referências Bibliográficas

Aftabuddin, M. e Kundu, S. (2006). Weighted and unweighted network of amino acids within protein. *Physica A-Statistical Mechanics and Its Applications*, 369(2):895–904.

Albert, L. e Barabasi, R. A. (1999). Emergence of scaling in random networks. *SCIENCE*, 286:509–512.

Albert, R. e Barabasi, A. L. (2000). Dynamics of complex systems: Scaling laws for the period of boolean networks. *Physical Review Letters*, 84(24):5660–5663.

Albert, R. e Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

Albert, R.; Jeong, H. e Barabasi, A. L. (1999). Internet diameter of the world-wide web. *Nature*, 401(6749):130–131.

Albet, R.; Jeong, N. e Barabasi, A. L. (2001). Error and attack tolerance of complex networks (vol 406, pg 378, 2000).

Aldana, M. e Cluzel, P. (2003). A natural class of robust networks. *PNAS*, 100(15):8710–8714.

Allen, P. M. (1997). *Cities and Regions as Self-Organizing Systems: Models of Complexity*. Taylor & Franci.

Almaas, E.; Krapivsky, P. L. e Redner, S. (2004). Statistics of weighted tree-like networks. *arXiv.org*, p. 9.

Alves, N. A. e Martinez, A. S. (2006). Topological features of proteins from amino acid residue networks. *arXiv.org*, p. 7.

Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I. e Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, 344(4):1135–1146.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096):223–230.

Ashby, W. (1969). *Systems Thinking*, chapter Self-Regulation and Requisite Variety, pp. 105–124. Penguin Books.

Aste, T.; Saadatfar, M.; Sakellariou, A. e Senden, T. J. (2004). Investigating the geometrical structure of disordered sphere packings. *Physica A: Statistical Mechanics and its Applications*, 339:16–23.

Aste, T.; Saadatfar, M. e Senden, T. J. (2006). Local and global relations between the number of contacts and density in monodisperse sphere packs. *Journal of Statistical Mechanics: Theory and Experiment*.

Aste, T. e Senden, T. (2007). The hierarchical properties of contact networks in granular packings. *arXiv:cond-mat/0504359v1*.

Atilgan, A. R.; Akan, P. e Baysal, C. (2004). Small-world communication of residues and significance for protein dynamics. *Biophys. J.*, 86(1):85–91.

Atilgan, A. R.; Turgut, D. e Atilgan, C. (2007). Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication. *Biophys. J.*, 92(9):3052–3062.

Axelrod, R. (1990). *The Evolution of Cooperation*. Penguin Books.

Axelrod, R. (1997). *The Complexity of Co-operation: Agent-Based Models of Competition and Collaboration*. Princeton University Pre.

Axelrod, R. M. e Cohen, M. D. (2000). *Harnessing Complexity: Organizational Implications of a Scientific Frontier*. Simon and Schuster.

Bagler, G. e Sinha, S. (2005). Network properties of protein structures. *Physica A: Statistical Mechanics and its Applications*, 346(1-2):27–33.

Bak, P.; Tang, C. e Wiesenfeld, K. (1988). Self-organized criticality. *APS Physical Review*, 38(1):364 – 374.

Barabasi, A. L.; Albert, R. e Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A*, 272(1-2):173–187.

Barabasi, A. L.; Albert, R. e Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*, 281(1-4):69–77.

Barkema, G. T. e Mousseau, N. (2001). The activation-relaxation technique: an efficient algorithm for sampling energy landscapes. *Computational Materials Science*, 20:285–292.

Barrat, A.; Barthelemy, M.; Pastor-Satorras, R. e Vespignani, A. (2004a). The architecture of complex weighted networks. *Proceedings of The National Academy of Sciences of The United States of America*, 101(11):3747–3752.

Barrat, A.; Barthelemy, M. e Vespignani, A. (2004b). Modeling the evolution of weighted networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 70(6):066149.

Barrat, A.; Barthelemy, M. e Vespignani, A. (2004c). Weighted evolving networks: Coupling topology and weight dynamics. *Physical Review Letters*, 92(22):228701.

Barthelemy, M. (2004). Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):163–168.

Barthelemy, M.; Barrat, A.; Pastor-Satorras, R. e Vespignani, A. (2005). Characterization and modeling of weighted networks. *Physica A-Statistical Mechanics and Its Applications*, 346(1-2):34–43.

Bashford, D.; Chothia, C. e Lesk, A. M. (1987). Determinants of a protein fold unique features of the globin amino acid sequences. *Journal of Molecular Biology*, 196(1):199–216.

Bastolla, U.; Porto, M.; Roman, H. E. e Vendruscolo, M. (2005). Looking at structure, stability, and evolution of proteins through the principal eigenvector of contact matrices and hydrophobicity profiles. *Gene*, 347(2):219–230.

Bauer, M. e Golinelli, O. (2001). Random incidence matrices: Moments of the spectral density. *Journal of Statistical Physics*, 103(1-2):301–337.

Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. e Bourne, P. (2000). The protein data bank. *Nucleic Acids Research*, 28:235–242.

Bobofchak, K. M.; Pineda, A. O.; Mathews, F. S. e Cera, E. D. (2005). Energetic and structural consequences of perturbing gly-193 in the oxyanion hole of serine proteases. *Journal of Biological Chemistry*, 280(27):25644–25650.

Bonabeau, E.; Dorigo, M. e Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press.

Brede, M. e Sinha, S. (2005). Assortative mixing by degree makes a network more unstable. *arXiv.org*, (arXiv:cond-mat/0507710v1):4.

Brinda, K.; Kannan, N. e Vishveshwara, S. (2002). Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein Eng.*, 15(4):265–277.

Brinda, K. V. e Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophysical Journal*, 89(6):4159–4170.

Callaway, D. S.; Newman, M. E. J.; Strogatz, S. H. e Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85(25):5468–5471.

Casti, J. L. (1998). *Would-Be Worlds: How Simulation is Changing the Frontiers of Science*. Wiley.

Chung, F.; Lu, L. Y. e Vu, V. (2003). Spectra of random graphs with given expected degrees. *Proceedings of The National Academy of Sciences of The United States of America*, 100(11):6313–6318.

Cobern, W. W. (1996). Worldview theory and conceptual change in science education. *Science Education*, 80(5):579 – 610.

Cohen, R.; Erez, K.; ben Avraham, D. e Havlin, S. (2000). Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628.

Cohen, R.; Erez, K.; ben Avraham, D. e Havlin, S. (2001). Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86(16):3682–3685.

Cooper, A. e Dryden, D. T. F. (1984). Allostery without conformational change. *European Biophysics Journal*, 11(2):103–109.

de Aguiar, M. A. M. e Bar-Yam, Y. (2005). Spectral analysis and the dynamic response of complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 71(1):016106.

del Sol, A.; Fujihashi, H.; Amoros, D. e Nussinov, R. (2006a). Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Sci*, 15(9):2120–2128.

del Sol, A.; Fujihashi, H.; Amoros, D. e Nussinov, R. (2006b). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular Systems Biology*, 2:1–12.

del Sol, A. e O'Meara, P. (2005). Small-world network approach to identify key residues in protein-protein interaction. *Proteins-Structure Function and Bioinformatics*, 58(3):672–682.

Desjarlais, J. R. e Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci*, 4(10):2006–2018.

Dhar, P. K.; Zhu, H. e Mishra, S. K. (2004). Computational approach to systems biology: From fraction to integration and beyond. *IEEE Transactions on Nanobioscience*, 3(3):144–152.

Dickerson, R. e Geis, I. (1983). *Hemoglobin: Structure, function, evolution and pathology.* The Benjamin/Cummins Publ. Co.

Dominy, B. e Brooks, C. (1999). Development of a generalized born model parametrization for proteins and nucleic acids. *Journal of Physical Chemistry B*, 103(18):3765–3773.

Dorogovtsev, S. N.; Goltsev, A. V.; Mendes, J. F. F. e Samukhin, A. N. (2003). Spectra of complex networks. *Physical Review E*, 68(4):046109.

Dorogovtsev, S. N.; Goltsev, A. V.; Mendes, J. F. F. e Samukhin, A. N. (2004). Random networks: eigenvalue spectra. *Physica A-Statistical Mechanics and Its Applications*, 338(1-2):76–83.

Durkheim, m. (2003). *Introdução ao pensamento sociológico*. Centauro, 16 edio.

Epstein, J. M. e Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press.

Erdös, P. e Rényi, A. (1959). On random graphs. *Pub. Mathem.*, 6:290–297.

Erdös, P. e Rényi, A. (1960). On the evolution of random graphs. *Bulletin of the International Statistical Institute*, 38(4):343–347.

Faloutsos, M.; Faloutsos, P. e Faloutsos, C. (1999). On power-law relationships of the internet topology. *Comput. Commun. Rev.*, 29:251–262.

Farkas, I.; Derenyi, I.; Jeong, H.; Meda, Z.; Oltvai, Z. N.; Ravasz, E.; Schubert, A.; Barabasi, A. L. e Vicsek, T. (2002). Networks in life: scaling properties and eigenvalue spectra. *Physica A-Statistical Mechanics and Its Applications*, 314(1-4):25–34.

Farkas, I. J.; Derenyi, I.; Barabasi, A. L. e Vicsek, T. (2001). Spectra of "real-world"graphs: Beyond the semicircle law. *Physical Review E*, 6402(2):026704.

Ferber, J. (1999). *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*. Addison Wesley Longman.

Fersht, A. R. e Daggett, V. (2002). Protein folding and unfolding at atomic resolution. *Cell*, 108(4):573–582.

Fox, Thomas, K. P. A. (1998). Application of the resp methodology in the parametrization of organic solvents. *Journal of Physical Chemistry B*, 102(41):8070–8079.

Frauenfelder, H. (1994). From symmetry to complexity. *Chinese Journal of Physics*, 32(6-11):1045–1050.

Freeman, L. C. (1977). Set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.

Gell-Mann, M. (1995). *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Owl Books.

Ghosh, A.; Brinda, K. V. e Vishveshwara, S. (2007). Dynamics of lysozyme structure network: Probing the process of unfolding. *Biophys. J.*

Gleick, J. (1987). *Chaos: making a new science*. Penguin Books.

Goh, K.-I.; Kahng, B. e Kim, D. (2001a). Spectra and eigenvectors of scale-free networks. *Phys. Rev. E*, 64(5):051903.

Goh, K. I.; Kahng, B. e Kim, D. (2001b). Universal behavior of load distribution in scale-free networks. *Physical Review Letters*, 87(27):278701.

Goh, K.-I.; Oh, E.; Jeong, H.; Kahng, B. e Kim, D. (2002). Classification of scale-free networks. *PNAS*, 99(20):12583–12588.

Gol'dshtein, V.; Koganov, G. A. e Surdutovich, G. I. (2004). Vulnerability and hierarchy of complex networks. *arXiv.org*, p. 4.

Golinelli, O. (2003). Statistics of delta peaks in the spectral density of large random trees. *arXiv.org cond-mat cond-mat/0301437*, p. 13.

Goodsell, D. S. e Olson, A. J. (2000). Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure*, 29:105–153.

Goodwin, B. (1995). *How the Leopard Changed Its Spots*. Simon and Schuster.

Greene, L. H. e Higman, V. A. (2003). Uncovering network systems within protein structures. *Journal of Molecular Biology*, 334(4):781–791.

Gujrati, P. D. (2007). Lack of stability in the stillinger-weber analysis, and a stable analysis of the potential energy landscape. *arXiv:cond-mat/0412735v1*, p. 5.

Gunasekaran, K.; Ma, B. e Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins. *Proteins: Structure, Function, and Bioinformatics*, 57(3):433 – 443.

Hardy, J. A. e Wells, J. A. (2004). Searching for new allosteric sites in enzymes. *Current Opinion in Structural Biology*, 14(6):706–715.

Hargrove, M.; Krzywda, S.; Wilkinson, A.; Dou, Y.; Ikeda, S. M. e Olson, J. (1994). Stability of myoglobin: a model for the folding of heme proteins. *Biochemistry*, 33:11767–75.

Higman, V. A. e Greene, L. H. (2006). Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. *Physica A-Statistical Mechanics and Its Applications*, 368(2):595–606.

Hodgson, G. M. (2001). *Darwinism and Evolutionary Economics*, chapter Is Social Evolution Lamarckian or Darwinian, pp. 87–118. Laurent, John and Nightingale.

Holland, J. H. (1995). *Hidden Order: How Adaptation Builds Complexity*. Addison Wesley Publishing Company.

Holland, J. H. (1998). *Emergence From Chaos to Order*. Oxford Univ Press.

Holme, P.; Kim, B. J.; Yoon, C. N. e Han, S. K. (2002). Attack vulnerability of complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 65(5):056109.

Hu, B.; Yan, G.; Wang, W.-X. e Chen, W. (2005). Weighted network models based on local and global rules. *arXiv.org*, p. 9.

Iben, I. E. T.; Braunstein, D.; Doster, W.; Frauenfelder, H.; Hong, M. K.; Johnson, J. B.; Luck, S.; Ormos, P.; Schulte, A.; Steinbach, P. J.; Xie, A. H. e Young, R. D. (1989). Glassy behavior of a protein. *Phys. Rev. Lett.*, 62(16):1916–1919.

Itoh, K. e Sasai, M. (2006). Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein a. *PNAS*, 103(19):7298–7303.

Jeong, H.; Mason, S. P.; Barabasi, A. L. e Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411:41–42.

Jiao, X.; Chang, S.; hua Li, C.; zu Chen, W. e xin Wang, C. (2007). Construction and application of the weighted amino acid network based on energy. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 75(5):051903.

Kalna, G. e Higham, D. J. (2006). Clustering coefficients for weighted networks. *University of Strathclyde Mathematics Research report 3*, p. 7.

Kamp, C. e Christensen, K. (2005). Spectral analysis of protein-protein interactions in drosophila melanogaster. *Physical Review E*, 71(4):041911.

Kapp, O. H.; Moens, L.; Vanfleteren, J.; Trotman, C. M. A.; Suzuki, T. e Vinogradov, S. N. (1995). Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume. *Protein Science*, 4(10):2179–2190.

Kauffman, S. (1996). *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford University Press.

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

Kern, D. e Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, 13(6):748–757.

Krem, M. M. e Cera, E. D. (2001). Molecular markers of serine protease evolution. *The EMBO Journal*, 20(12):3036–3045.

Krem, M. M.; Prasad, S. e Cera, E. D. (2002). Ser214 is crucial for substrate binding to serine proteases. *J. Biol. Chem.*, 277(43):40260–40264.

Krishnadev, O.; Brinda, K. e Vishveshwara, S. (2005). A graph spectral analysis of the structural similarity network of protein chains. *Proteins: Structure, Function, and Bioinformatics*, 61:152–163.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, 1 edio.

Kundu, S. (2005). Amino acid network within protein. *Physica A: Statistical Mechanics and its Applications*, 346(1-2):104–109.

Lehninger, A. L.; Nelson, D. L. e Cox, M. M. (2007). *Princípios de Bioquímica*. 4 edio.

Lesk, A. e Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of molecular Biology*, 3(136):225–70.

Lesk, A. e Chotia, C. (1980). How different amino-acid-sequences determine similar protein structures - structure and evolutionary dynamics of the globins. *Journal Of Molecular Biology*, 136:225.

Levy, Y. e Onuchic, J. N. (2004). Water and proteins: A love-hate relationship. *PNAS*, 101(10):3325–3326.

Li, C. G. e Chen, G. R. (2004). A comprehensive weighted evolving network model. *Physica A-Statistical Mechanics and Its Applications*, 343:288–294.

Li, W. e Cai, X. (2004). Statistical analysis of airport network of china. *Physical Review E*, 69(4):046106.

Liljeros, F.; Edling, C. R.; Amaral, L. A. N.; Stanley, H. E. e Aberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908.

Lochmann, K.; Oger, L. e Stoyan, D. (2006). Statistical analysis of random sphere packings with variable radius distribution. *Solid State Sciences*, 8(12):1397–1413.

Lockless, S. W. e Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299.

Luhman, N. (1990). *Essays on Self Reference*. Columbia University Press.

Martin, P. C.; Siggia, E. D. e Rose, H. A. (1973). Statistical dynamics of classical systems. *Phys. Rev. A*, 8(1):423–437.

Maturana, H. R. e Varela, F. J. (1992). *The tree of knowledge: The biological roots of human understanding*. Shambhala.

Milgotina, E. I.; Voyushina, T. L. e Chestukhina, G. G. (2003). Glutamyl endopeptidases: Structure, function, and practical application. *Russian Journal of Bioorganic Chemistry*, 29(6):511–522.

Milgram, S. (1967). The small-world problem. *Psychology Today*, pp. 60–67.

Morin, E. (1986). *O método: A natureza da natureza*, volume 1. Men Martins, 1 edio.

Mulej, M.; Potocan, V.; Zenko, Z.; Kajzer, S.; Ursic, D. e Knez-Riedl, J. (2004). How to restore bertalanffian systems thinking. *Kybernetes*, 33:48–61.

Newman, M. E. J. (2001a). Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 6401(1):016131.

Newman, M. E. J. (2001b). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 6401(1):016132.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20):208701.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Newman, M. E. J. (2003b). Properties of highly clustered networks. *Physical Review E*, 68(2):026121.

Newman, M. E. J. (2003c). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 70(5):056131.

Newman, M. E. J.; Strogatz, S. H. e Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 6402(2):026118.

Newman, M. E. J. e Watts, D. J. (1999). Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4-6):341–346.

Nicolescu, B. (2000). *Educação e transdisciplinaridade*. UNESCO, USP/Escola do Futuro, CESP.

Nicolis, G. e Prigogine, I. (1989). *Exploring complexity. an introduction*. W.H. Freeman.

Nienaber, V.; Wang, J.; Davidson, D. e Henkin, J. (2000). Re-engineering of human urokinase provides a system for structure-based drug design at high resolution and reveals a novel structural subsite. *J. Biol. Chem.*, 275(10):7239–7248.

Onnela, J. P.; Saramaki, J.; Kertesz, J. e Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103.

Onuchic, José Nelson, W. P. G. (2004). Theory of protein folding. *Current Opinion in Structural Biology*, 14(1):70–75.

Paccanaro, A.; Casbon, J. A. e Saqi, M. A. S. (2006). Spectral clustering of protein sequences. *Nucl. Acids Res.*, 34(5):1571–1580.

Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z. e Wolynes, P. G. (2004). Water in protein structure prediction. *PNAS*, 101(10):3352–3357.

Park, K.; Lai, Y. C. e Ye, N. (2004). Characterization of weighted complex networks. *Physical Review E*, 70(2):026109.

Perutz, M. (1970). Stereochemistry of cooperative effects in haemoglobin. *Nature*, 228:726–739.

Philips, J. C.; Braun, R.; Wei, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L. e Schulten, K. (2005). Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(15):1781–1802.

Prigogine, I. (1990). Time and the problem of the two cultures. In *First International Dialogue on the Transition to Global Society*.

Prigogine, I. e Stengers, I. (1984). *Order out chaos. man's new dialogue with nature.* Bantam Books.

Ptitsyn, O. B. e Ting, K. L. H. (1999). Non-functional conserved residues in globins and their possible role as a folding nucleus. *Journal of Molecular Biology*, 291(3):671–682.

Rao, F. e Caflisch, A. (2004). The protein folding network. *Journal of Molecular Biology*, 342(1):299–306.

Ravasz, E.; Somera, A. L.; Mongru, D. A.; Oltvai, Z. N. e Barbasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

Restrepo, J. G.; Ott, E. e Hunt, B. R. (2006a). Characterizing the dynamical importance of network nodes and links. *Physical Review Letters*, 97(9):094102.

Restrepo, J. G.; Ott, E. e Hunt, B. R. (2006b). Emergence of coherence in complex networks of heterogeneous dynamical systems. *Physical Review Letters*, 96(25):254103.

Rodgers, G. J.; Austin, K.; Kahng, B. e Kim, D. (2005). Eigenvalue spectra of complex networks. *Journal of Physics A-Mathematical and General*, 38(43):9431–9437.

Schrödinger, E. (1944). *What is life? The physical aspect of the living cell.* Cambridge University Press, Cambridge.

Schutz, C. N. e Warshel, A. (2001). What are the dielectric constants of proteins and how to validate electrostatic models? *Proteins: Structure, Function, and Genetics*, 44(4):400 – 417.

Seary, A. J. e Richards, W. D. (2003). Spectral methods for analyzing and visualizing networks: an introduction.

Süel, G. M.; Lockless, S. W.; Wall, M. A. e Ranganathan, R. (2002). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1):59–69.

Silbert, L. E.; Ertas, D.; Grest, G. S.; Halsey, T. C. e Levine, D. (2002). Geometry of frictionless and frictional sphere packings. *Physical Review E*, 65(3):031304.

Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. e Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332.

Sole, R. V.; Kauffman, S. A. e Pastoras, R. (2005). *Scaling and Phase Transitions in Complex Systems*. Oxford University Press.

Stillinger, F. H. e Weber, T. A. (1985). Computer simulation of local order in condensed phases of silicon. *Phys. Rev. B*, 31(8):5262–5271.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–276.

Swain, J. F. e Gierasch, L. M. (2006). The changing landscape of protein allostery. *Current Opinion in Structural Biology*, 16(1):102–108.

Tersoff, J. (1988). New empirical approach for the structure and energy of covalent systems. *Phys. Rev. B*, 37(12):6991–7000.

Toroczkai, Z. e Bassler, K. E. (2004). Jamming is limited in scale-free systems.

Ueda, Y. (1979). Randomly transitional phenomena in the system governed by duffing's equation. *Journal of Statistical Physics*, 20(2):181–196.

Valle, J. G. d. R. (2004). *Dicionário de Latim-Português*, volume 1. IOB/Thomson, 1 edio.

Vazquez, A. e Moreno, Y. (2003). Resilience to damage of graphs with degree correlations. *Physical Review E*, 67(1):015101.

Vazquez, A. e Weigt, M. (2003). Computational complexity arising from degree correlations in networks. *Physical Review E*, 67(2):027101.

Veloso, C.; Silveira, C.; Melo, R.; Ribeiro, C.; Lopes, J.; Santoro, M. e Jr., W. M. (2007). On the characterization of energy networks of proteins. *Genetic and Molecular Research*, 6(4):799–820.

Vendruscolo, M.; Dokholyan, N. V.; Paci, E. e Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 65(6):061910.

Vendruscolo, M.; Paci, E.; Dobson, C. M. e Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409:641–645.

Vera, S. e Waelbroeck, H. (1996). Symmetry breaking and adaptation: The genetic code of retroviral env proteins. *arXiv.org*.

Vinogradov, S. N.; Hoogewijs, D.; Bailly, X.; Arredondo-Peter, R.; Gough, J.; Dewilde, S.; Moens, L. e Vanfleteren, J. R. (2006). A phylogenomic profile of globins. *BMC Evolutionary Biology*, 6:31.

von Bertalanffy, L. (1950). The theory of open systems in physics and biology. *Science*, 111:23–29.

von Bertalanffy, L. (1975). *Teoria geral dos sistemas*. Vozes.

Waldrop, M. M. (1992). *Complexity: The emerging science at the edge of order and chaos*. Simon and Schuster.

Wang, X. F. e Chen, G. (2003). Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, 3(1):6–20.

Wang, Y.; Chakrabarti, D.; Wang, C. e Faloutsos, C. (2003). Epidemic spreading in real networks: An eigenvalue viewpoint.

Wangikar, P. P.; Tendulkar, A. V.; Ramya, S.; Mali, D. N. e Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *Journal of Molecular Biology*, 326(3):955–978.

Watts, D. J. (2002). A simple model of global cascades on random networks. *PNAS*, 99(9):5766–5771.

Watts, D. J. e Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393:440–442.

Webster, G. e Goodwin, B. (1996). *Form and Transformation: Generative and Relational Principles in Biolog*. Cambridge University Press.

Wiener, N. (1948). *Cybernetics: or the Control and Communication in the Animal and the Machine*. The MIT Press, 2 edio.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *The Annals of Mathematics*, 62(3):548–564.

Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proceedings of the National Academy of Sciences of United States of America*, 93:14249–14255.

Worda, J. M.; Lovella, S. C.; LaBeana, T. H.; Taylora, H. C.; Zalisa, M. E.; Presleya, B. K.; Richardson, J. S. e Richardson, D. C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of Molecular Biology*, 285(4):1711–1733.

Wu, Z.-X.; Xu, X.-J. e Wang, Y.-H. (2005). Properties of weighted structured scale-free networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 45(3):385–390.

Yook, S. H.; Jeong, H.; Barabasi, A. L. e Tu, Y. (2001). Weighted evolving networks. *Physical Review Letters*, 86(25):5835–5838.

Zhao, F.; Yang, H. e Wang, B. (2005a). Scaling invariance in spectra of complex networks: A diffusion factorial moment approach. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 72(4):046119.

Zhao, M.; Zhou, T.; Wang, B. H. e Wang, W. X. (2005b). Enhanced synchronizability by structural perturbations. *Physical Review E*, 72(5):057102.

Zhu, L.; Li, P.; Huang, M.; Sage, J. T. e Champion, P. M. (1994). Real time observation of low frequency heme protein vibrations using femtosecond coherence spectroscopy. *Phys. Rev. Lett.*, 72(2):301–304.