RAQUEL CARDOSO DE MELO MINARDI

CLASSIFICAÇÃO ESTRUTURAL DE FAMÍLIAS DE PROTEÍNAS COM BASE EM MAPAS DE CONTATOS

Belo Horizonte
04 de junho de 2008

Universidade Federal de Minas Gerais
Instituto de Cincias Exatas
Programa de Ps-Graduao em Bioinformática

CLASSIFICAÇÃO ESTRUTURAL DE FAMÍLIAS DE PROTEÍNAS COM BASE EM MAPAS DE CONTATOS

Tese apresentada 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.

RAQUEL CARDOSO DE MELO MINARDI

Belo Horizonte
04 de junho de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Classificação Estrutural de Famílias de Proteínas com Base em Mapas de Contatos

RAQUEL CARDOSO DE MELO MINARDI

Tese defendida e aprovada pela banca examinadora constituída por:

Prof. Ph. D. Marcelo Matos Santoro – Orientador
Universidade Federal de Minas Gerais

Prof. Ph. D. Wagner Meira Jr. – Co-orientador
Universidade Federal de Minas Gerais

Prof. Ph. D. Júlio César Dias Lopes – Co-orientador
Universidade Federal de Minas Gerais

Ph. D. Goran Neshich – Co-orientador
Empresa Brasileira de Pesquisa AgropequÃ¡ria

Prof. Ph. D. Júnior Barrera
Universidade de São Paulo

Prof Ph. D. Rodrigo Weber dos Santos
Universidade Federal de Juíz de Fora

Prof. Ph. D. Wálter Filgueira de Azevedo Júnior
Pontifícia Universidade Católica do Rio Grande do Sul

Profa. Ph. D. Glaura da ConceiÇcão Franco
Universidade Federal de Minas Gerais

Belo Horizonte, 04 de junho de 2008

Resumo Estendido

O objetivo deste trabalho é verificar se é possível classificar estruturas de cadeias proteicas utilizando apenas os dados das interações químicas entre os seus resíduos de aminoácidos. Através de mapas de contatos gerados a partir de dados do STING e a utilização de três diferentes métricas baseadas em técnicas de processamento de imagens somos capazes de classificar tais estruturas em famílias de similar estrutura e função.

Fizemos alguns ensaios de variação de atributos no intuito de encontrar possíveis componentes de assinaturas estruturais de cada uma dessas famílias. Verificamos que existem alguns tipos de contatos mais relevantes na discriminação das famílias (pontes de hidrogênio sem intermediação de moléculas de água, contatos hidrofóbicos e ligações íon-íon) e outros menos relevantes (pontes de hidrogênio intermediadas por moléculas de água). Mostramos também que contatos entre resíduos muito próximos na seqüência (menos de 30 resíduos de distância) não são muito úteis na classificação, sendo aparentemente ruídos nesse processo. Além disto, pelos resultados preliminares, nem só os resíduos que formam um grande número de contatos são importantes. Resíduos com poucos contatos aparentemente são imprescindíveis na definição da família estrutural. Mostramos que uma das técnicas de comparação de mapas de contatos desenvolvida pode ser útil, adicionalmente, no alinhamento de contatos. Através destes alinhamentos podemos, por exemplo, verificar as alterações conservativas nos contatos de uma proteína mutante em relação à selvagem. Pode-se também, estudar comparativamente uma mesma proteína de diversas espeécies animais.

Isto gerou ferramentas muito úteis na comparação de proteínas de uma mesma topologia e diferentes espécies e também no entendimento das variações de estabilidade de uma proteína selvagem e seus mutantes.

As técnicas desenvolvidas parecem ser úteis também no estudo de padrões de interações entre diferentes cadeias proteicas. Em ensaios com *serino-proteases* e seus inibidores, os *BPTIs*, mostramos ser possível definir um padrão de contatos potencialmente importantes na complexação do inibidor à protease.

Alguns dos resultados deste trabalho foram implementados e estão disponíveis na

ferramenta STING (http://www.cbi.cnptia.embrapa.br/SMS/). Participamos da concepção e implementação de três diferentes módulos: PCD (*(*Protein Contacts Difference)), TopSiMap (*Topology Similarity Map*) e Topologs (um banco de dados de estruturas similares tomando-se como base apenas contatos).

Abstract

The objective of this work was to verify if it is possible to classify protein chain structures using only the chemical interactions between its residues. Through contact maps and using three different metrics based on image processing techniques we have showed that we are able to classify such structures in families of similar structure and function with precision up to 99%. We have performed some experiments with attributes variation to find possible components of the structural signatures of each of the studied protein families. We have verified that some types of interactions are more discriminator then others (they are hydrogen bonds without water molecules in the middle of residues, hydrophobic contacts and ion-ion linking) and that other are less discriminator (hydrogen bonds intermediated by water molecules). We also have showed that contacts between residues which are sequentially close (less than 30 residues of distance) are not very discriminator attributes for classification, apparently being noises in the process. Moreover, for the preliminary results, the residues that form a great number of contacts are not more important that the less connected ones as one should previously think. Residues with few contacts apparently are essential in the definition of the structural signature of a family. We have showed that one of the techniques for contact maps comparison can additionally be useful as an heuristic for the contact map overlap problem. It can be used to align contact maps and through these alignments we can, for example, study mutations in residues that does not affect the pattern of contacts. We can compare mutant and wild proteins and also, comparatively study a protein of diverse animal species. Another important tested use of the technique is in the discovery of a pattern of interactions between different protein chains in complexes. In assays with serine-proteases and its inhibitors, the BPTIs, we have showed that it is possible to define a set of potentially important contacts in the binding and stabilization of the complexes. Some of the results of this work had been implemented and are available, beyond this site, in the STING (http://www.cbi.cnptia.embrapa.br/SMS). We participate of the conception and implementation of three different modules: PCD (Protein Contacts Difference), TopSiMap (Topology Similarity Map) and Topologs (a data base of similar structures being overcome as base only contacts).

Dedico este trabalho primeiramente a Deus pois sem Ele nada seria possível e não estaríamos aqui desfrutando destes tão importantes momentos.

Dedico, também, ás pessoas mais importantes da minha vida. Estas pessoas que não só me apresentaram os projetos dos sonhos, como desafiaram-me a construí-los e que também foram me ajudando nesta construção dia após dia

- *A minha mãe Maria José, por sempre acreditar em mim mais do que eu mesma, pelo carinho e infinita dedicação.*
- *Ao meu pai Júlio, autodidata e meu maior exemplo de que podemos aprender e fazer muito mais do que imaginam.*
- *Ao meu marido Ângelo por acreditar e compartilhar comigo todos os sonhos e pelo seu enorme amor.*
- *E á minha avó Conceição, meu primeiro modelo de professor. Por sua culpa, vislumbrei um ideal nesta profissão...*

Agradecimentos

A Deus, à minha família e aos professores Marcelo Santoro, Wagner Meira Jr., Júlio César Dias Lopes e ao Dr. Goran Neshich e Dr. Carlos Herique da Silveira.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Diversidade funcional e estrutural de proteínas

A palavra *proteína* vem do grego *protas* que significa "de muita importância". Proteínas são compostos orgânicos complexos que consistem em resíduos de aminoácidos unidos por ligações peptídicas. Foram descobertas em 1.838 por Jöns Jakob Berzelius e são as mais ativamente estudadas moléculas na Bioquímica, sendo essenciais para as estruturas e funções das células vivas e vírus.

Diferentes proteínas desempenham uma ampla variedade de funções biológicas. Algumas proteínas são enzimas (Figura1.1a), catalizadoras de reações químicas. Geralmente aumentam a velocidade de uma reação em pelo menos 1 milhão de vezes. Outras têm papel essencial nos processos de resposta imunológica. Os anticorpos (Figura 1.1b) são proteínas altamente específicas que reconhecem e se combinam com substâncias estranhas como vírus, bactérias e células de outros organismos. Há também aquelas que têm papel estrutural e mecânico como, por exemplo, as proteínas constituintes do citoesqueleto. A alta força de tensão da nossa pele e ossos é devida à presença do *Colágeno* (Figura 1.1c), uma proteína fibrosa. O armazenamento e transporte de substâncias também são feitos por proteínas. A *Hemoglobina* (Figura 1.1d), por exemplo, transporta o oxigênio nas hemácias, enquanto a *Mioglobina* o armazena nos músculos. O ferro é transportado no plasma sangüineo pela *Transferrina* e é armazenado no fígado na forma de um complexo com a *Ferritina*. A *Insulina* (Figura 1.1e) é o hormônio responsável pela redução da taxa de glicose no sangue.

1.2 Aminoácidos

Os *aminoácidos* são as unidades estruturais básicas das proteínas (Figura 1.2). Eles são constituídos por um grupamento *amina* ($-NH_2$), uma *carboxila* ($-COOH$), um

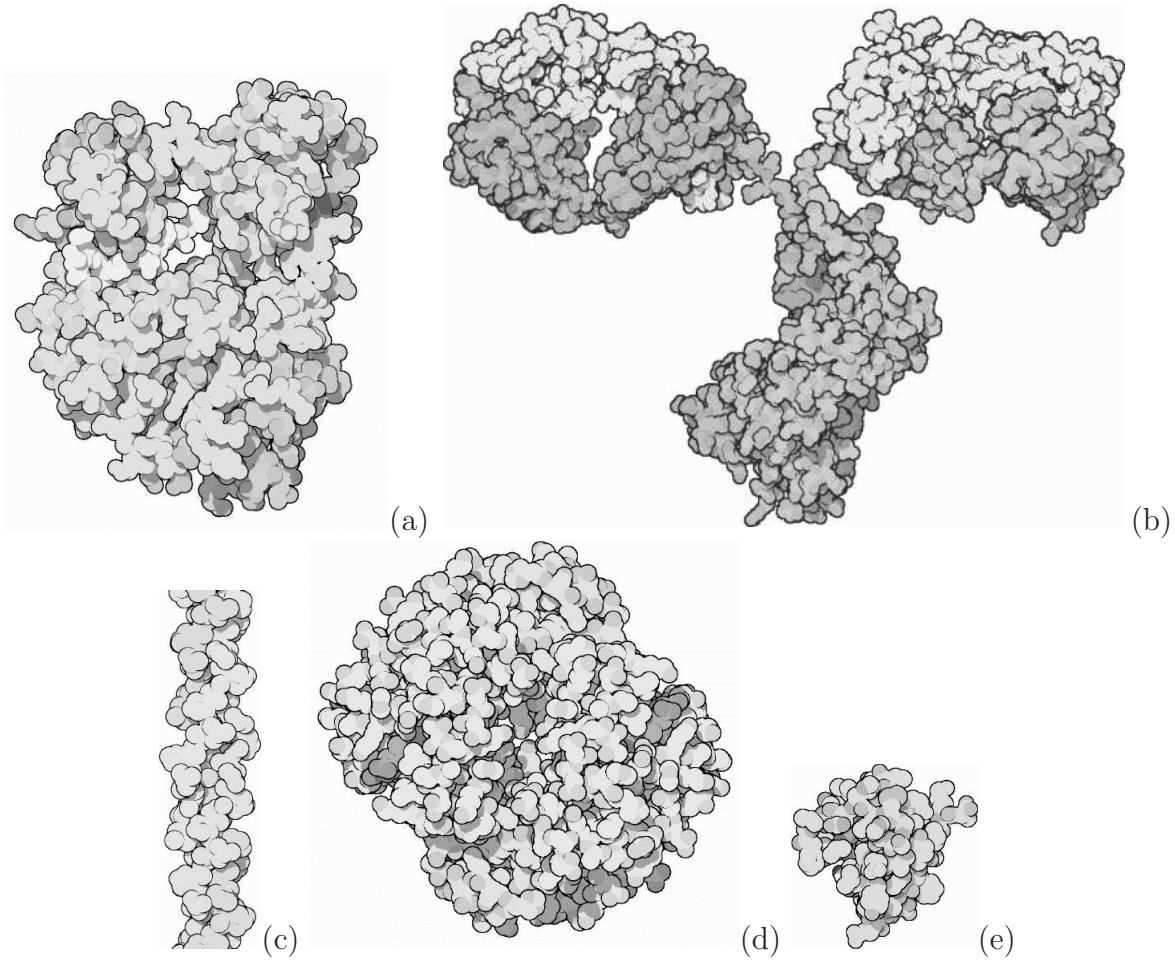


Figura 1.1: Variedade estrutural e funcional das proteínas

(a) *Src Tyrosine Quinase*, enzima de sinalização. Localizada na membrana celular, auxilia na passagem de sinais que regulam a síntese de proteínas e o crescimento celular. **(b)** *Anticorpo IgG1*, um ligante neutralizador do vírus HIV-1. **(c)** *Colágeno*, de papel essencialmente estrutural, é a principal proteína presente em nosso tecido conjuntivo e a mais abundante de nosso organismo. **(d)** *Hemoglobina*, a proteína dos glóbulos vermelhos responsável pelo armazenamento e transporte do oxigênio em nosso organismo. **(e)** *Insulina*, hormônio polipeptídico sintetizado no pâncreas.

átomo de H e um grupamento R diferenciado, todos eles ligados a um átomo de C denominado Cα. O grupamento R é conhecido como *cadeia lateral* (CL).

As proteínas são compostas por um repertório de 20 tipos de aminoácidos mais comumente encontrados nos seres vivos e esse alfabeto é conservado há bilhões de anos. Os nomes destes aminoácidos bem como suas abreviações são apresentados na Tabela 1.1.

O que diferencia estes 20 aminoácidos são suas diversas cadeias laterais (Figura 1.3). Estas variam em tamanho, forma, carga, capacidade de formação de pontes de

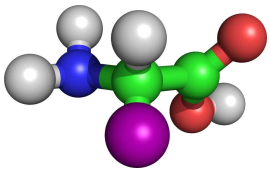


Figura 1.2: Estrutura básica de um aminoácido.

Em azul, o átomo de N da amina; em vermelho, os átomos de O da carboxila; em verde, os átomos de C; em branco, os átomos de H e em violeta o radical variável prasente em todos os aminoácidos.

Tabela 1.1: Nomenclatura e abreviações utilizadas para os aminoácidos comumente encontrados em proteínas.

Nome do aminoácido	**Abreviação de 3 letras**	**Abreviação de 1 letra**
Alanina	ALA	A
Arginina	ARG	R
Asparagina	ASN	N
Aspartato	ASP	D
Cisteína	CYS	C
Glutamato	GLU	E
Glutamina	GLN	Q
Glicina	GLY	G
Histidina	HIS	H
Isoleucina	ILE	I
Leucina	LEU	L
Lisina	LYS	K
Metionina	MET	M
Fenilalanina	PHE	F
Prolina	PRO	P
Serina	SER	S
Treonina	THR	T
Triptofano	TRP	W
Tirosina	TYR	Y
Valine	VAL	V

hidrogênio, caráter hidrofóbico e reatividade química.

1.3 Ligação peptídica

Conforme dito anteriormente, as proteínas são polímeros lineares que se formam pela ligação de grupos carboxila de aminoácidos com os grupos aminas dos aminoácidos

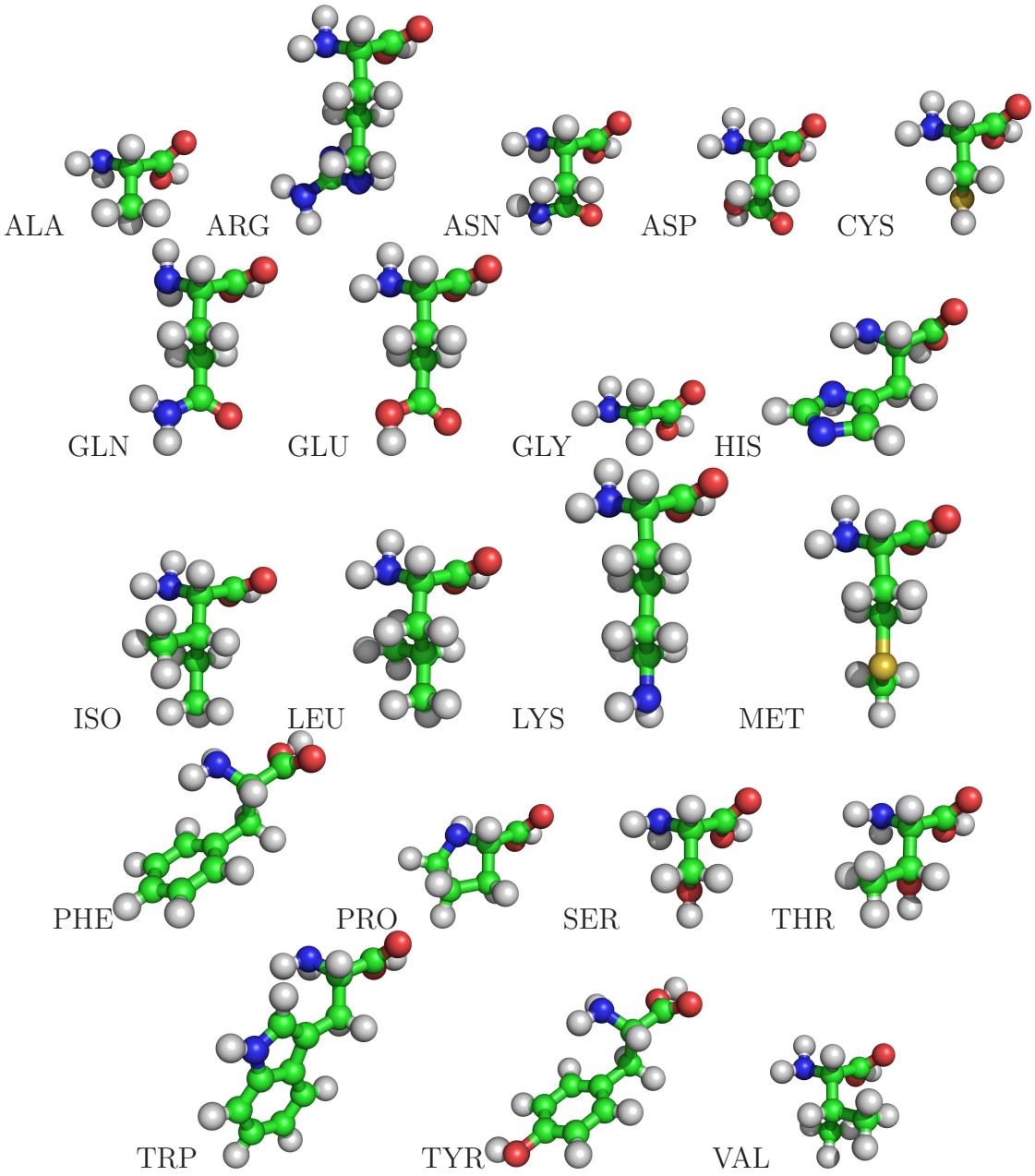


Figura 1.3: 20 aminoácidos mais comumente encontrados nos seres vivos

seguintes. Essa ligação é denominada *ligação peptídica* e ocorre com a liberação de uma molécula de água. Após a ligação de dois aminoácidos (com a perda de átomos de O e H da carboxila que se torna um grupo *carbonila* ($-C=O$) e de um átomo de H da amina originando um grupo *amida* ($-NH$)), estes passam a ser denominados *resíduos de aminoácidos* (Figura 1.4d).

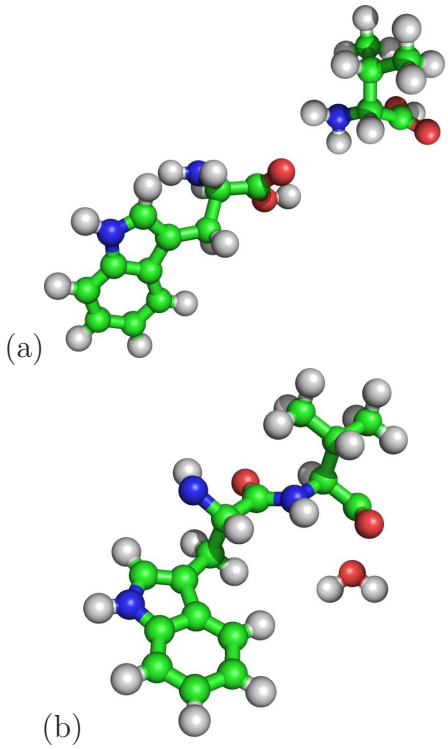


Figura 1.4: Ligação peptídica

Em **(a)**, à esquerda um Triptofano e à direita uma Valina. Em **(b)**, o grupo carboxila do Triptofano se liga ao grupo amina da Valina com a liberação de uma molécula de água. Observe que se forma uma amida entre os resíduos dos 2 aminoácidos da ligação peptídica.

1.4 Estruturas primária, secundária, terciária e quaternária de proteínas

Esta seqüência de resíduos ligados por ligações peptídicas que formam uma cadeia polipeptídica é denominada *estrutura primária* da proteína. Por convenção devido à direção da síntese proteica, o terminal amida da cadeia é tomado como início da seqüência (*N-terminal*) e o carboxila é o fim da cadeia (*C-terminal*). Às partes altamente repetitivas das cadeias polipeptídicas (formadas pelo carbono α e grupos carbonila e amida), damos o nome de *cadeia principal* (CP), sendo as partes variáveis as cadeias laterais (CL).

Existem ainda as denominações *estrutura secundária*, *estrutura terciária* e *estrutura quaternária*. As *estruturas secundárias* são padrões tridimensionais que ocorrem em segmentos de proteínas devido a padrões de pontes de hidrogênio e serão detalhadas posteriormente. A *estrutura terciária* é a estrutura tridimensional da proteína definida pelas coordenadas x, y e z dos seus átomos. A *estrutura quatenária* é um nível adicional de organização molecular que consiste no arranjo de múltiplas cadeias enoveladas em

um complexo com duas ou mais subunidades, iguais ou diferentes.

As estruturas tridimensionais das proteínas são constituídas de domínios. A primeira definição de domínios foi proposta por Wetlaufer em 1973 [Wetlaufer e Ristow, 1973] como unidades estáveis de estruturas de proteínas que podem enovelarse de forma autônoma. Desde então este conceito também tem sido relacionado a unidades de estrutura compacta, com propriedades funcionais e evolutivas.

1.5 Restrições conformacionais da cadeia

1.5.1 Paradoxo de Levinthal

Como pode esta seqüência linear de resíduos de aminoácidos se enovelar formando estruturas tridimensionais extremamente complexas? Em 1968, Cyrus Lenvinthal [Levinthal, 1968] levantou um paradoxo muito importante na teoria da dinâmica de enovelamento de proteínas. Ele provou que a busca de uma cadeia polipeptídica desenovelada por sua conformação nativa não podia ser uma busca aleatória, mas devia ser dirigida.

Considerando uma cadeia polipeptídica hipotética de 100 resíduos de aminoácidos e, com absurda simplificação, considerando ainda que cada resíduo pudesse se apresentar em 3 diferentes conformações, a cadeia teria $3^{100} \approx 5 \times 10^{47}$ configurações. Se esta cadeia pudesse mudar de conformação 10^{13} vezes por segundo, ou 3×10^{20} por ano, levaria 10^{27} anos para gerar todas conformações e todo este tempo é maior que a idade do universo. Como as proteínas se enovelam em escala de segundos ou menos, buscas aleatórias não são efetivamente a forma como as cadeias se enovelam.

1.5.2 Planaridade da ligação peptídica

Existem vários fatores conhecidos que reduzem o astronômico número de possíveis conformações para uma cadeia de resíduos. O primeiro deles é a própria natureza química da ligação peptídica que é, essencialmente, planar de forma que seis átomos dos resíduos ligados estão em um mesmo plano: o Cα e o grupo carbonila do primeiro resíduo e o grupo amida e o Cα do segundo (Figura 1.5).

1.5.3 Ângulos ϕ (phi) e ψ (psi)

A ligação peptídica tem caráter de ligação parcialmente dupla, o que impossibilita a sua rotação e restringe as possíveis conformações da cadeia polipeptídica. Em contraste, as ligações entre o grupo amida e o Cα, assim como entre o grupo carbonila e o Cα, são ligações simples, podendo rotacionar tomando várias orientações. Na Figura 1.6,

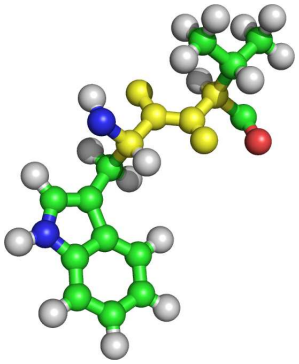


Figura 1.5: Átomos componentes do plano da ligação peptídica

Em amarelo, podemos ver os átomos do grupo carbonila e o Cα do Triptofano e os átomos do grupo amida e o Cα da Valina em um plano.

podemos ver 2 planos consecutivos formados em uma cadeia polipeptídica hipotética (ILE-TRP-VAL) unidos pelo Cα do resíduo do meio (TRP). Devido à possibilidade de rotação das ligações entre o Cα e os grupos amida e carbonila do Triptofano, os planos podem girar com certo grau de liberdade. São esses graus de liberdade que possibilitam que a cadeia polipeptídica tome uma infinidade de conformações.

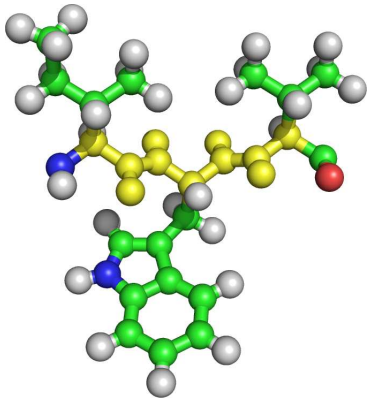


Figura 1.6: Planos consecutivos da cadeia polipeptídica

Nesta figura, acrescentamos outro resíduo a cadeia de polipeptídica hipotética. Observe que temos uma Isoleucina, seguida pelo Triptofano e pela Valina. Em amarelo, podemos ver os átomos formando 2 planos conectados pelo Cα do Triptofano.

As rotações dessas duas ligações são chamadas *ângulos diedros*. O ângulo entre o N da amida e o Cα é chamado ϕ (phi) e o ângulo entre o Cα e o C da carbonila é chamado ψ (psi). Porém, Ramachandran mostrou através de seu mapa que nem todas as combinações de ângulos ϕ e ψ são possíveis devido a conflitos estéricos entre os átomos.

1.5.4 Interações não-Covalentes entre os resíduos de aminoácidos

Conforme explicamos, as proteínas são cadeias de aminoácidos estruturados tridimensionalmente. É essa estrutura que possibilita a execução das mais complexas e diversas funções bioquímicas. A estruturação da cadeia e a sua manutenção neste estado enovelado e funcional deve-se, em grande parte, às interações eletrostáticas não locais entre os resíduos de aminoácidos distantes na seqüência.

A maioria dos processos químicos está relacionada a alterações na distribuição dos elétrons entre os átomos. Todas as interações químicas entre os resíduos de aminoácidos em proteínas envolvem variações nas distribuições de cargas [Lopes, 2006].

É importante considerar que a energia da interação entre átomos varia com a variação da distância entre eles. Obviamente, a grandes distâncias, não existe qualquer interação mas, à medida que a distância diminui, ocorrem interações de crescente intensidade até que o sistema seja estabilizado na mais provável distância de ligação. Neste ponto, temos um mínimo de energia, predominando a atração entre os átomos. Com distâncias mais curtas, e a conseqüente aproximação de suas nuvens eletrônicas, o processo começa a ser repulsivo.

As interações não locais são quase sempre não-covalentes. Uma *ligação covalente* é uma ligação química caracterizada pelo compartilhamento de um ou mais pares de elétrons entre dois componentes, produzindo uma atração que segura a molécula resultante unida. Os átomos tendem a compartilhar estes elétrons para que sua camada de valência seja preenchida. As interações *não-covalentes* são de natureza mais fraca que as covalentes. As covalentes não passam de 40KJ/mol enquanto as não-covalentes podem chegar a 1.000KJ/mol.

Um tipo de interação não covalente e muito importante no entendimento de estruturas de proteínas são as *ligações dipolo-dipolo*. Elas foram inicialmente estudadas e postuladas por Johannes Diderik van der Waals em 1.873, tendo recebido o seu nome. Os *dipolos permamentes* aparecem das ligações químicas entre átomos de diferentes eletronegatividades. Os *dipolos induzidos*, por sua vez, aparecem por indução de campos elétricos nas vizinhanças, em decorrência de interação com cargas elétricas e persistem enquanto persistir a origem do campo elétrico. Elas são também conhecidas como *forças de dispersão de London* em homenagem a Fritz London, seu descobridor. A intensidade das interações entre dipolos permanentes depende da polaridade das ligações, enquanto nos dipolos induzidos ela depende da polarizabilidade dos elétrons, ou seja, da suscetibilidade da nuvem eletrônica à deformação. Átomos maiores e menos eletronegativos são mais polarizáveis e apresentam interações entre dipolos induzidos mais fortes.

As *ligações de hidrogênio*, extremamente importantes na estabilização das estruturas secundárias de proteínas, são também interações dipolo-dipolo, diferenciando-se pela maior intensidade e direcionalidade. A força da ligação de hidrogênio depende do alinhamento entre os átomos que interagem. Flúor, oxigênio e nitrogênio são os mais comuns átomos formadores de pontes de hidrogênio. A exigência para formação de uma ponte de hidrogênio é a ligação polar de um hidrogênio com um átomo eletronegativo, o doador. O átomo aceptor de hidrogênio deve ser um átomo com pares de elétrons livres. Quanto maior a eletronegatividade do átomo doador mais forte a interação. Quanto maior e eletronegatividade do átomo aceptor mais fraca a interação. Apenas oxigênio, nitrogênio e flúor apresentam pares de elétrons não ligados disponíveis. Átomos mais pesados (tais como cloro e enxofre) também podem participar de pontes de hidrogênio, assim com as menos polarizadas (como *C-H* por exemplo).

De grande importância são, adicionalmente, as *ligações íon-íon*. Têm caráter eletrostático como as dipolo-dipolo mas ocorrem entre átomos com cargas formais e são bem mais fortes. Em proteínas existem 3 resíduos carregados positivamente: Argininas, Lisinas e Histidinas (sendo que esta pode ter carga parcial quando desprotonada) e 2 negativamente: Aspartato e Glutamato.

Essenciais no enovelamento proteico são também as *interações hidrofóbicas* uma vez que, nas células, as proteínas estão em meio aquoso. O efeito hidrofóbico está relacionado à tendência das moléculas apolares sofrerem agregação em água. A formação de interações dipolo permanente-dipolo induzido entre as moléculas de água e de ramificações apolares da proteína são mais fortes que as ligações dipolo induzido-dipolo induzido entre trechos da própria proteína. No entanto, ocorre uma reorganização das moléculas de água em torno das partes apolares da proteína imobilizando um grande número de moléculas de água na solvatação. Isto significa perda de entropia das moléculas de água, o que torna o processo desfavorável. Desta forma, trechos apolares tendem a se aglutinar expondo a mínima superfície possível para solvatação.

Apesar de covalentes, é importante mencionar as *pontes dissulfeto*. Elas ocorrem quando dois átomos de enxofre ligam-se pela oxidação dos grupos sulfidrila (*S-H*) dos resíduos de cisteína. São as únicas ligações covalentes e não locais presentes em proteínas sendo também muito importantes no enovelamento e estabilização de algumas proteínas.

1.5.5 Estruturas secundárias

O grupo *CO* (carbonila) é um bom aceptor e o grupo *NH* (amina) é um bom doador. Esses grupos interagem com outros trechos da cadeia sendo muito importantes na estabilização das estruturas de proteínas e reduzindo obviamente o número

de conformações possíveis para esta cadeia.

Em 1.951, Linus Pauling e Robert Corey propuseram a existência de dois tipos de estruturas muito comuns em proteínas: as *α-hélices* [Pauling et al., 1951] e as *folhas-β* [Pauling e Corey, 1951]. Estas descobertas foram feitas com base nos estudos das propensões de formação de pontes de hidrogênio dos átomos da cadeia principal e, posteriormente, comprovadas por difração de raios X.

As α-hélices (Figura 1.7) são estabilizadas por pontes de hidrogênio entre os grupos amida (doador) e carbonila (aceptor) de resíduos da cadeia principal com uma rotação de cerca de 100 graus. Isto significa uma separação de, em média, 3,6 resíduos (≈ 4) e 1,5Å de elevação de cada volta da hélice. Desta forma, a principal característica de uma α-hélice é que entre os resíduos i e $i+4$ existe uma ponte hidrogênio.

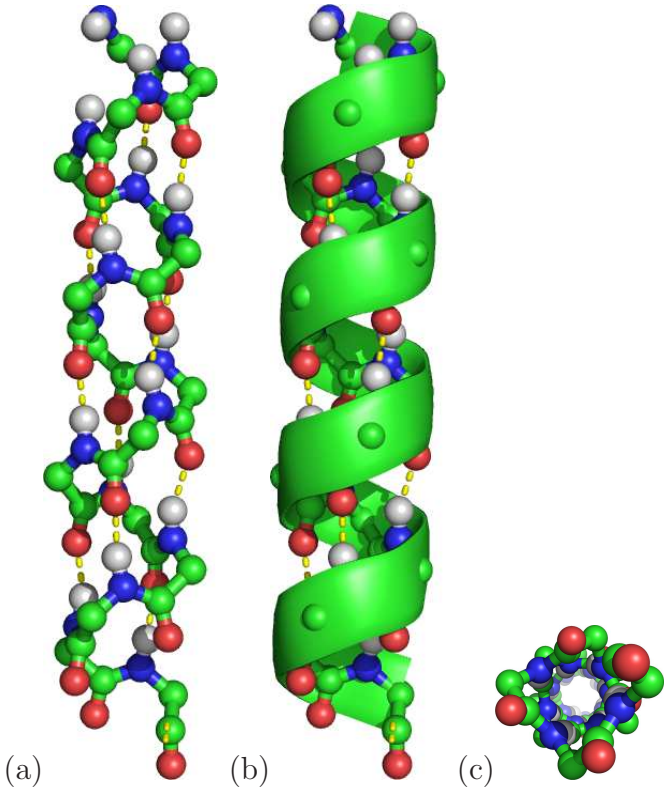


Figura 1.7: α-hélice

(a) Nesta figura, são exibidos apenas os átomos da cadeia principal de uma α-hélice. Note que as pontes de hidrogênio entre os H dos grupos amida e os C dos grupos carbonilas são destacadas com uma linha tracejada. **(b)** A mesma hélice exibida em esquema de *cartoon*. **(c)** Hélice vista de cima.

Existem ainda outros tipos de hélices menos comuns em proteínas: as hélices-3_{10} que apresentam pontes de hidrogênio entre os resíduos i e $i+3$ e as hélices-π, entre os resíduos i e $i+5$.

As α-hélices são bastante compactas não restando espaço em seu interior de modo que as cadeias laterais de seus resíduos ficam sempre apontando para fora da hélice.

Os resíduos com maior propensão de formação de α-hélices são a Metionina, a Alanina, a Leucina, o Glutamato e a Lisina. Por outro lado, a Prolina, a Glicina, a Tirosina e a Serina têm baixa propensão. A Prolina não é um doador de hidrogênio e interfere estericamente uma vez que seu anel restringe o ângulo ϕ da cadeia principal e, por isso, costuma ser uma iniciadora ou finalizadora de hélices. A Glicina apresenta um problema oposto: devido a sua alta flexibilidade conformacional torna cara entropicamente a sua restrição à conformação de hélice.

Como, por formação, todos os dipolos dos grupos carbonil ($C = O$) são posicionados em uma mesma direção e sentido, a hélice tem um momento de dipolo causado por esse efeito agregado. Normalmente, hélices possuem um aminoácido negativo em seu N-terminal. Podem possuir também um positivo em seu C-terminal. O N-terminal de hélices pode ser usado na interação com ligantes carregados negativamente uma vez que a amida de sua cadeia principal pode servir como doadora de H.

As folhas-β (Figura 1.8) são outro tipo de estrutura comum em proteínas e são formadas por pontes de hidrogênio entre grupamentos amida e carbonila em fitas peptídicas. A distância axial entre os resíduos adjacentes é de cerca de 3,5Å.

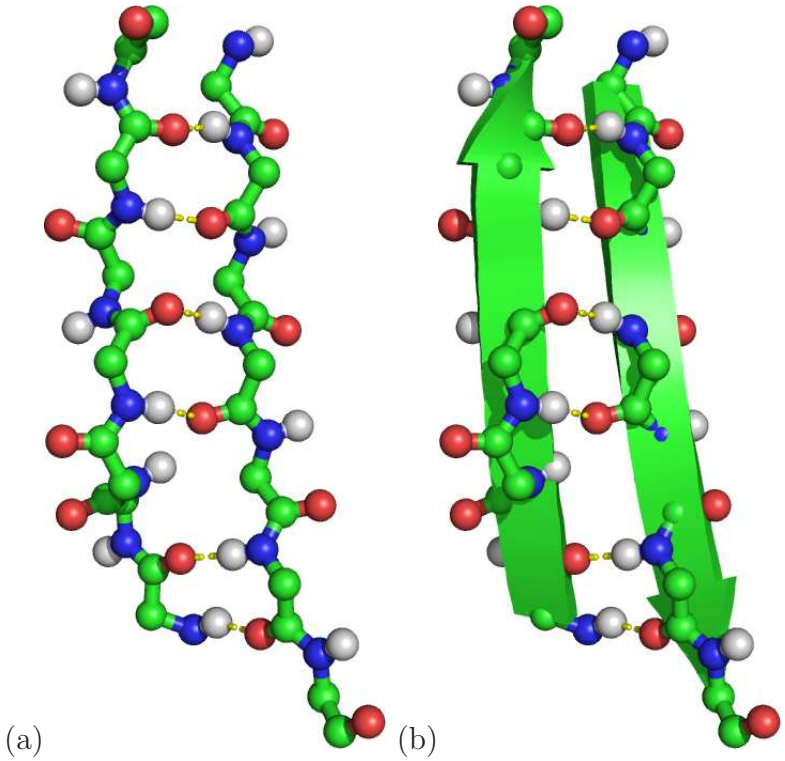
Folhas-β podem aparecer em *paralelo* ou *antiparalelo* de acordo com as direções (em termos de N-terminal e C-terminal) das fitas em contato. Veja o exemplo de folhas-β retirado da *Carboxipeptidase A* na Figura 1.9.

Note que quando vários segmentos da cadeia principal se emparelham e formam uma rede de pontes de hidrogênio, as cadeias laterais (que não foram exibidas na Figura 1.8) apontam uma para cima outra para baixo da rede sucessivamente, conforme Figura 1.10.

1.6 Especificidades dos resíduos de aminoácidos no enovelamento e atividade de proteínas

A *Alanina* é um aminoácido apolar, ou seja, hidrofóbico. É um dos aminoácidos mais freqüentes nas proteínas dos seres vivos.

A *Arginina* é uma cadeia alifática de 4 carbonos finalizada por um grupo *guanidina* (CH_5N_3). Este grupamento é formado pela oxidação do grupo *guanina*. Em condições fisiológicas, com um pK_a de aproximadamente $12,5$, é encontrado protonado ($CH_6N_3^+$), portanto com carga $+1$. Devido à sua geometria, sua distribuição de cargas e sua habilidade de formar pontes de hidrogênio, este aminoácido é usualmente encontrado interagindo com grupamentos negativos. Por este motivo é, geralmente, encontrada



(a) (b)

Figura 1.8: Folha-β

(a) Nesta figura, são exibidos apenas os átomos da cadeia principal de folhas-β. As pontes de hidrogênio que estabilizam esta estrutura são apresentadas em linha tracejada. **(b)** As mesmas folhas-β vistas em esquema de *cartoon*.

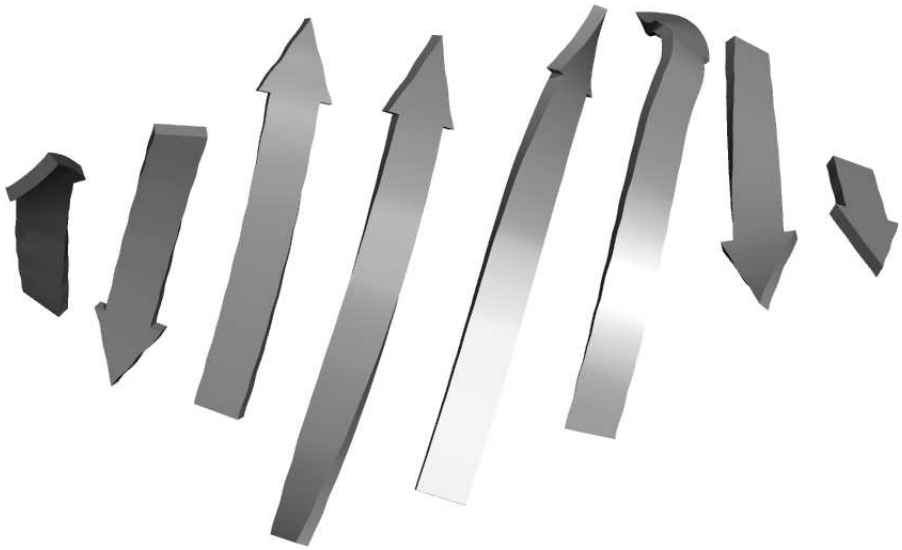


Figura 1.9: Folhas-β paralelas e anti-paralelas

exposta ao solvente onde pode interagir com as moléculas polares da água.

A *Asparagina* tem um grupamento carboxi-amida ($R - CO - NH_2$) em sua cadeia

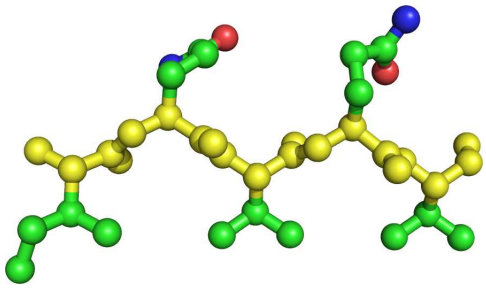


Figura 1.10: Posicionamento das cadeias laterais em folhas-β

Nesta figura, apresentamos um segmento da cadeia que forma a folha-*beta* da Figura 1.8. Os átomos de H foram removidos para melhorar a clareza e os átomos da cadeia principal (que forma a rede de pontes de hidrogênio) são exibidos em amarelo. Perceba o posiocionamente alternando para cima e para baixo das cadeias laterais. As pontes de hidrogênio, netes caso, estão perpendiculares ao plano deste papel.

lateral. Devido ao seu alto potencial de formação de pontes de hidrogênio com a cadeia principal de proteínas, é freqüentemente encontrada em inícios e términos de α-hélices, além de voltas de folhas-β.

O *Aspartato* é o ânion carboxilato do *ácido aspártico*, apresentando carga -1 no grupamento COO da cadeia lateral em pH fisiológico.

A *Cisteína* possui um grupamento *tiol* em sua cadeia lateral, o que lhe dá características hidrofílicas. Devido à alta reatividade química (nucleofílico e facilmente oxidado) deste grupamento, este resíduo é de muita importância estrutural e funcional em muitas proteínas.

O *Glutamato* é o ânion carboxilato do *ácido glutâmico*. Como o nome indica, ele possui um *ácido carboxílico* ($-C(=O)OH$) em sua cadeia lateral e, em pH fisiológico é encontrado desprotonado com carga -1.

A *Glutamina* é um aminoácido formado pela substitição de um *hidroxil* do Ácido Glutâmico por um grupo funcional *amina*.

A *Glicina* é o aminoácido mais simples. Sua cadeia lateral é formada por apenas um átomo de H e seu $C\alpha$ não é quiral.

A *Histidina* possui um grupo *imidazole* em sua cadeia lateral. Este grupamento possui 2 átomos de N: um deles é ligado a um H e, portanto, é ácido; o outro é básico. Estas propriedades são exploradas de formas diferentes. Em tríades catalíticas, o N básico pode abstrair um próton de Serinas, Treoninas e Cisteínas para ativá-las como um nucleófilo. Ela também pode ser útil na transferência de próton de uma molécula para outra através da abstração de um próton da molécula origem por seu N básico e da posterior doação do próton do seu N ácido para a molécula destino. A Histidina tem grande afinidade por metais.

A *Isoleucina* é um aminoácido, cuja cadeia lateral é composta apenas de átomos de C e H sendo, portanto, bastante hidrofóbica.

A *Leucina* também possui sua cadeia lateral composta apenas por átomos de C e H e é hidrofóbica.

A *Lisina* é um resíduo de aminoácido de cadeia alifática e, em *pH* fisiológico, é encontrada com carga +1.

A *Metionina* é um resíduo de aminoácido apolar e contém um átomo de S.

A *Fenilalanina* possui um grupamento *benzil* em sua cadeia lateral de forma que é um resíduo hidrofóbico.

A *Prolina* é um dos resíduos mais rígidos devido ao seu anel ser formado com a inclusão de átomos da cadeia principal. Este resíduo não favorece a formação de estrutiras secundárias sendo muito comuns no início de α-hélices e folhas-β. Também é frequentemente encontrada em voltas e exposta ao solvente. Como não tem o hidrogênio do grupo amida, não serve como doador de H mas apenas aceptor.

A *Serina* é um resíduo polar sendo muito importante para a função catalílitica de algumas enzimas.

A *Treonina* é um resíduo polar, semelhante à Serina.

O *Triptofano* se diferencia dos demais resíduos, pois sua cadeia lateral é composta por um grupo *indol*. Este grupamento é um composto aromático bicíclico consistindo de um anel de benzeno com 6 carbonos e um anel pirrólico com 5 membros sendo um nitrogênio. É um resíduo apolar e bastante volumoso.

A *Tirosina* possui sua cadeia lateral formada por um grupo *fenol* que lhe confere função especial como transportadora de grupos *fosfato*. É um resíduo polar.

A *Valina* é um resíduo bastante hidrofóbico.

Entender como esse alfabeto é usado na criação das mais complexas estruturas tridimensionais (Figura 1.1) que possibilitam a essas moléculas desempenharem as mais variadas funções biológicas é uma questão em aberto na bioquímica.

1.7 Famílias de proteínas modelo

1.7.1 Globinas

Nos trabalhos desenvolvidos ao longo desta tese, usaremos como principal família experimental as *Globinas*. Elas foram as primeiras proteínas a terem sua estrutura elucidada, sendo as mais bem estudadas. Proteínas deste enovelamento podem ser encontradas como monômeros ou em complexos. São extremamente compactas e compostas por cerca de 153 resíduos de aminoácidos, tendo um tamanho aproximado de $45 \times 35 \times 25$Å. Para funcionar, dependem da presença do grupo prostético *heme* que

coordena o oxigênio através de um átomo de ferro. Cerca de 70% de sua cadeia é enovelada em forma de, em média, 8 hélices. Seu interior é composto basicamente por resíduos apolares como leucina, valina, metionina e fenilalanina. Os resíduos carregados, aspartato, glutamato, lisina e arginina, estão quase sempre expostos ao solvente. Os únicos resíduos polares no interior da molécula são duas histidinas que são essenciais na ligação de ferro e oxigênio.

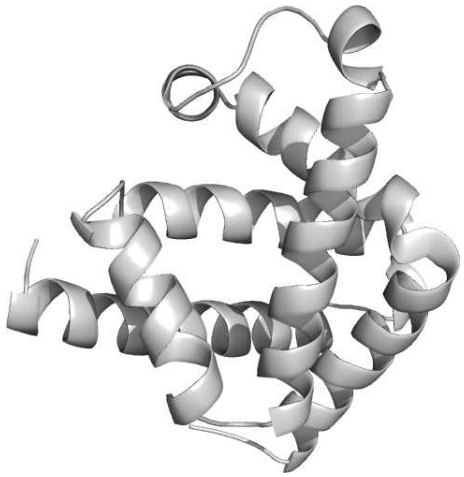


Figura 1.11: Mioglobina de Baleia (PDB id 1a6m)

1.7.2 Outras famílias

Adicionalmente, utilizamos nos nossos experimentos outras famílias de proteínas de enovelamentos diveros:

- *Apolipoproteínas*, proteínas compostas por um feixe de 4 α-hélices;
- *Plastocianinas*, proteínas constituídas por um barril de 6 fitas β;
- *Retinol-binding proteins*, proteínas consituídas por um barril de 8 fitas β acompanhado por pequenas α-hélices;
- *Tioredoxinas* proteínas compostas por folha α / β aberta e torcida.

1.7.3 Complexos Serino-protease - BPTI

Durante o desenvolvimento desta tese, optamos por aplicar as técnicas desenvolvidas para classificação de estruturas na tentativa de se buscar padrões de interações entre cadeias de proteínas. Para estes experimentos, o complexo modelo foi o de *Serino-proteases* com seu principal inibidor, o *Bovine Pancratic Tryipson Inhibtor* (BPTI).

As Serino-proteases são peptidases, ou seja, enzimas responsáveis pela quebra de ligações peptídicas e são caracterizadas pela presença de um resíduo de serina em seu sítio catalítico (tríade catalítica, uma vez que é constituída por 3 resíduos). Participam de inúmeras funções vitais nos seres vivos como, por exemplo, coagulação, imunização e digestão.

Estas enzimas podem ser inibidas por um grande conjunto de outras proteínas. Uma delas é o BPTI que é uma pequena proteína globular composta de 53 resíduos e estabilizada por 3 pontes dissulfeto. Esta molécula foi uma das primeiras a terem sua estrutura resolvida por NMR (*Ressonância Nuclear Magnética*) e é administrada como medicação para reduzir o sangramento principalmente em cirurgias de coração e fígado.

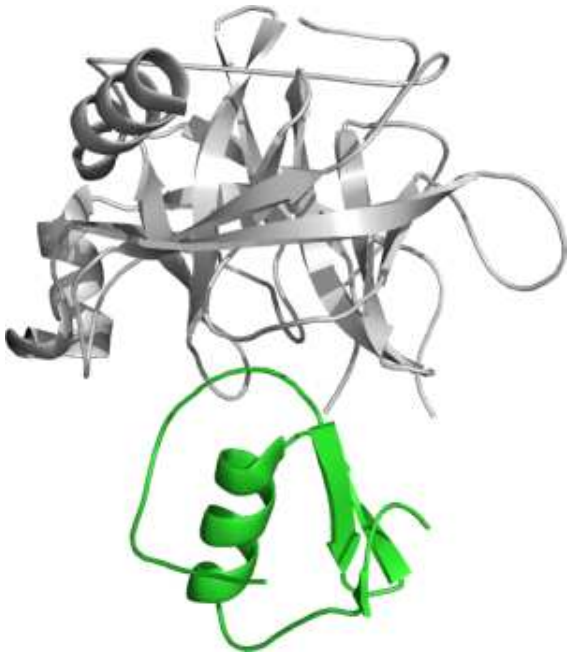


Figura 1.12: Complexo Serino-protease - BPTI (Quimotripsina (PDB id 1cho))

A Serino-protease é apresentada em ciza e o BPTI em verde.

1.8 Dados disponíveis sobre proteínas

O *Uniprot* (Universal Protein Resource) [Bairoch et al., 2004] do *European Bioinformatics Institute* (EBI) é o maior catálogo de informações sobre seqüências de proteínas. Na versão atual, estão disponíveis cerca de 350.000 seqüências das mais variadas famílias de proteínas.

O EBI provê ainda outros 16 bancos de dados com informações sobre seqüências anotadas de proteínas. Apresentam uma classificação das seqüências de acordo com

sua similaridade, das interações entre diferentes proteínas, de seus sítios funcionais, de proteínas que são enzimas e seus sítios catalíticos, entre outras.

Dentre as milhões de seqüências disponíveis nos bancos de dados públicos, apenas cerca de 50.000 estruturas de proteínas e seus complexos foram resolvidas e estão depositadas no *Protein Data Bank* (PDB) [Berman et al., 2000]. Cada arquivo no PDB possui várias informações das quais destacamos a posição no espaço tridimensional de cada átomo das moléculas de proteínas. Neste trabalho, utilizamos apenas proteínas e seus complexos com estrutura resolvida, ou seja, as coordenadas de seus átomos.

1.9 Seqüência × estrutura × função de proteínas

Por volta de 1.955, Christian Anfinsen publicou seus primeiros trabalhos [Anfinsen et al., 1954, Anfinsen et al., 1955] e duas décadas depois ganhou o Premio Nobel em Química [Anfinsen, 1973] com a demonstração, em experimentos com a *Ribonuclease*, da relação entre a seqüência e a estrutura de proteínas. A *Ribonuclease* é uma enzima constituída por uma única cadeia de 124 resíduos com a formação de 4 pontes dissulfeto. Ele desnaturou a proteína na pretensão de verificar em quais condições a mesma poderia ser renaturada.

Agentes como *uréia* ou *cloreto de guanidina* rompem as ligações não covalentes. Pontes dissulfeto podem ser desfeitas reversivelmente através do tratamento com β-*mercaptoetanol*. Anfinsen tratou a *Ribonuclease* com essas substâncias, desenovelando completamente as proteínas. Com a posterior redução na concentração destes compostos, verificou que a enzima pouco a pouco recuperava sua atividade enzimática perdida com a desnaturação. Todas as propriedades físicas e químicas da enzima renaturada eram idênticas às da enzima nativa. Estes experimentos mostraram que toda a informação necessária para especificar a estrutura cataliticamente ativa da *Ribonuclease* estava contida na seqüência de resíduos de aminoácidos que a compõem.

Estudos posteriores mostraram a generalidade desse achado que é um dos postulados centrais da Bioquímica: a seqüência especifica a conformação, ou a estrutura. Esta dependência é muito importante devido à intima relação entre estrutura e função. A função que uma proteína desempenha em um organismo é completamente dependente de sua estrutura tridimensional uma vez que é essa quem confere a especificidade à molécula.

1.10 Importância de se classificar estruturas

Estruturas de proteínas podem ser classificadas de formas variadas por:

```
1a6mA VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED  60
1dlwA ___________SLFEQLGGQAA___________VQAVTAQFYANIQADATVATFFNGID  37
                  :: :: .:.*             :. * : : .::   * * :  . *

1a6mA LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP 120
1dlwA MPNQTNKTAAFLCAALG__GPNAWTGRNLKEVHAN___MGVSNAQFT_TVIGHLRSALTG  91
      : ::  .. : * * *   * :    : * : **.   : :.  :*   .* *:  :

1a6mA GDFGADAQGAMNKALELFRKDIAAKYKELGY                              151
1dlwA AGVAAALVEQTVAVAETVRGDVVTV______                             116
      ....*        . * .* *:.:
```

Figura 1.13: Alinhamento das seqüências das Mioglobinas de baleia (PDB id 1a6m) e de ciliado (PDB id 1dlw).

Asteríscos indicam resíduos conservados em ambas as seqüências; dois pontos, mutações conservativas e ponto, mutações semi-conservativas.

- similaridade funcional
- similaridade evolucionária da seqüência de resíduos de aminoácidos
- similaridade de enovelamento.

A comparação de seqüências é um método bastante simples de se obter informações sobre a relação estrutural e evolucionária de proteínas. Duas proteínas com cerca de 40% de identidade entre os aminoácidos de sua seqüência terão, com altíssima probabilidade, estruturas similares [Leach, 2001]. Quando uma seqüência de estrutura desconhecida têm alta similaridade com uma de estrutura resolvida, podemos deduzir a nova estrutura através de modelos computacionais feitos a partir da estrutura modelo.

Porém, considere a comparação entre duas *Mioglobinas*: a primeira de baleia e a outra de ciliado (Figura 1.13). Apesar da alta similaridade estrutural e identidade funcional, conforme pode ser comprovado no alinhamento abaixo, existe apenas 12,58% de identidade entre seus aminácidos no alinhamento de suas seqüências. Mesmo se relaxarmos essa comparação considerando as mutações conservativas e semi-conservativas, obtemos índices de 36,42% e 47,68% respectivamente. Isto nos mostra que existem seqüências pouco relacionadas mesmo para proteínas muito similares o que enfraquece a abordagem apenas por seqüências.

É preciso comparar as proteínas estruturalmente. As estruturas das proteínas podem elucidar sua função e sua história evolucionária. Qual é a origem da semelhança estrutural de proteínas, cujas seqüências não apresentam similaridade seqüencial significativa? Para elucidar essa questão estudos de classificação de estruturas de proteínas são muito importantes. Eles têm definido famílias de proteínas que compartilham

um núcleo estrutural similar, ou seja, os mesmos elementos de estrutura secundária conectados na mesma topologia de forma independente da variabilidade seqüencial. Proteínas de enovelamento similar, geralmente, são relacionadas evolutivamente e desempenham funções similares [Brenner et al., 1995].

Em [Murzin et al., 1995], os autores apresentam o *Structural Classification of Proteins (SCOP)*, um banco de dados de classificação estrutural de domínios de proteínas que foi contruído basicamente por inspeção visual e comparação de estruturas através de métodos automáticos. Os domínios são classificados hierárquicamente contemplando relacionamentos evolucionários e estruturais nos seguintes níveis: famílias, superfamílias, enovelamento e classe conforme será detalhado na Seção 2.1.2.

Posteriormente, outros autores em [Pearl et al., 2003] apresentam um novo banco de dados de estruturas de domínios de proteínas. Nesta base, cada domínio é classificado em super-famílias e famílias de seqüência. Os mesmos autores produziram também um *software* denominado CATHEDRAL para comparação de estruturas de proteínas. Este sistema é totalmente baseado no casamento de estruturas secundárias e tenta classificar uma estrutura de família desconhecida em uma das famílias do CATH.

1.11 Assinaturas estruturais

Assinaturas estruturais são representações, possivelmente multidimensionais e concisas, das características das proteínas de mesmo enovelamento. São um conjunto de características inerentes às seqüências que são determinantes do seu enovelamento e atividade.

1.12 Mapas de contatos e sua relação com a estrutura

A conformação tridimensional de uma proteína pode ser representada de forma bastante compacta como uma matriz esparsa, quadrada, simétrica e binária de contatos inter-resíduos, ou mapa de contatos. Um mapa de contatos é uma representação particularmente útil da estrutura de proteínas provendo informações sobre suas estruturas secundárias e capturando aspectos de sua estrutura tridimensional.

Uma proteína de n resíduos tem um mapa de contato $n \times n$. Se dois resíduos de aminoácidos a_i e a_j estiverem em contato, a posição (i, j) terá um ponto, caso contrário, ficará em branco.

Dizemos que dois resíduos de aminoácidos estão em contato se fazem uma ligação não-covalente (exceto as pontes dissulfeto). Existem várias metodologias propostas

para definição destes contatos. A mais simples delas consiste em utilizar um valor de corte para a distância de separação no espaço tridimensional entre os átomos dos resíduos (seja considerando todos os seus átomos ou apenas os carbonos α). Em [Hu et al., 2002], os autores utilizam uma distância de corte de 7Å. [Sobolev et al., 1999] descrevem uma metodologia muito mais apurada para detecção dos contatos. Ela considera não só as distâncias inter-atômicas como também a natureza dos átomos próximos e suas ligações. A Figura 1.14 a seguir mostra um mapa de contatos de uma *Mioglobina*.

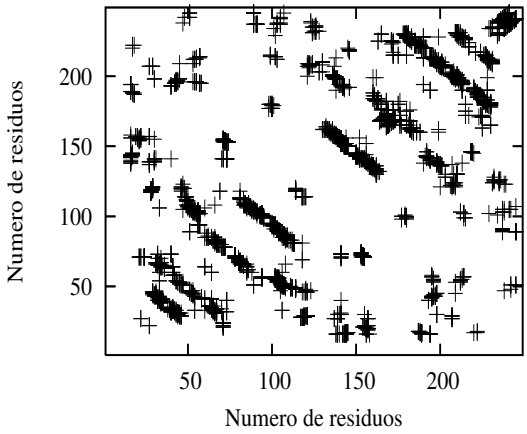


Figura 1.14: Um exemplo de mapa de contatos.

Mapa de contatos de uma *Mioglobina* de baleia (PDB id 1a6m).

Para mostrar como os mapas de contatos são uma boa e robusta representação da estrutura de proteínas, vamos detalhar este mesmo mapa de *Mioglobina* de baleia, associando alguns trechos à estrutura.

Observe que existe um grande número de contatos próximos à diagonal do mapa (Figura 1.15). Estes são contatos entre resíduos bastante próximos na seqüência. Geralmente, são pontes de hidrogênio responsáveis pela formação das α-hélices. É possível perceber claramente interrupções nestes contatos da diagonal. Estas interrupções indicam as regiões de cadeia não estruturada em hélices. Podemos observar no mapa de contatos as 8 hélices comumente encontradas nas Globinas (denominadas na literatura pelas letras de A a H).

Os agrupamentos de contatos distantes da diagonal indicam contatos não locais. Observando na estrutura da *Mioglobina* as hélices que estão próximas (obviamente fazendo contato umas com as outras), vamos verificar no mapa que existem contatos entre elas. As hélices G e H, por exemplo, estão ligeiramente cruzadas e em contato, de forma que no quadrante do mapa relativo a estas hélices, é possível ver grande número de interações (em destaque na Figura 1.16). Por outro lado, as hélices C e H estão

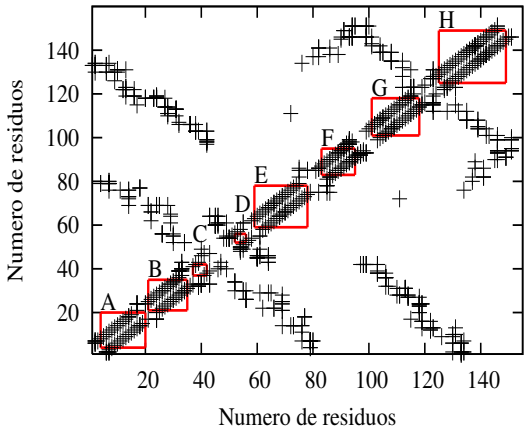


Figura 1.15: Contatos responsáveis pela formação de α-hélices.

extremamente afastadas estruturalmente de modo que é natural não encontrar nenhum contato relativo a estes trechos no mapa.

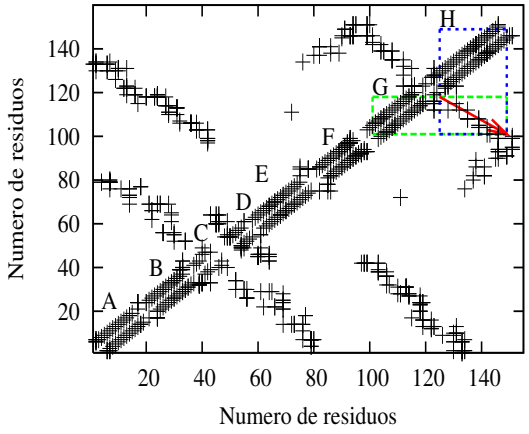
Observe ainda que os grupamentos de contatos não locais podem aparecer como retas crescentes ou descrescentes. Esta é uma característica interessante por mostrar se os trechos da cadeia em contato têm ou não a mesma orientação na seqüência. Agrupamentos crescentes indicam que as partes estão em contato paralelamente, ou seja, seus N-terminais e C-terminais estão na mesma orientação (como aproximadamente acontece com as hélices F e H). No caso desta *Mioglobina*, a maioria dos agrupamentos são descrescentes indicando contatos antiparalelos (como por exemplo as hélices G e H).

1.13 Motivação

As proteínas são macromoléculas essenciais não só na estruturação como em processos químicos das células vivas e vírus. O entendimento de como um repertório de 20 aminoácidos é usado na composição dessas moléculas com tão diferenciadas e complexas estruturas e funções biológicas é uma questão em aberto na Bioquímica moderna. Apesar das restrições estruturais impostas pelas ligações peptídicas, os ângulos diedrais dão à cadeia de aminoácidos tamanha liberdade que é, atualmente, impossível prever a estrutura de uma proteína partindo apenas de sua seqüência de aminoácidos. Entender profundamente a relação entre a seqüência de aminoácidos, a estrutura e a função de proteínas é de capital importância no entendimento do processo de enovelamento destas e conseqüentemente, na elucidação de patologias provenientes da sua má-formação e possível desenvolvimento de terapias.

O estabelecimento de assinaturas estruturais para famílias de proteínas é um passo

(a)



(b)

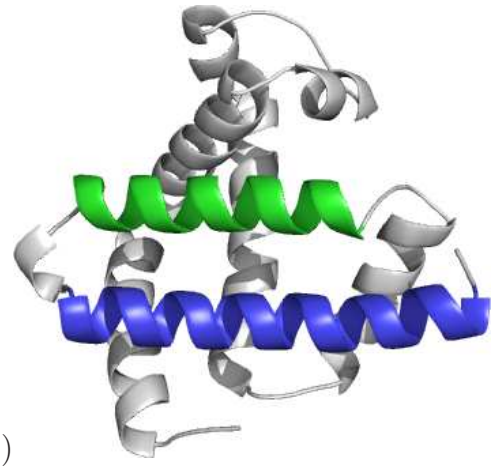


Figura 1.16: Um exemplo da associação entre os contatos de um mapa e uma estrutura.

(a) Mapa de contato de uma *Mioglobina* de baleia (PDB id 1a6m) e **(b)** a respectiva estrutura da proteína.

essencial nesse processo de busca e conhecimento dos aspectos necessários para que um grupo de proteínas, com seqüências potencialmente bastante diversas, enovelem-se em semelhantes estruturas e desempenhem funções idênticas.

Acreditamos que existe um padrão de ligações não-covalentes que seja preservado para cada família de proteínas funcionalmente equivalentes. É objetivo deste trabalho estabelecer e desenvolver metodologias para obter esse padrão de contatos que deve ser mantido mesmo com alta variabilidade na dimensão seqüencial. Acreditamos que, mesmo com a variação do alfabeto que compõe um dado conjunto de proteínas de mesma função, os contatos mais preservados são responsáveis pela estruturação similar das proteínas, o que lhes confere a mesma semântica ou funcionalidade.

1.13.1 Trabalhos relacionados

Ao iniciar este projeto, não foram encontrados no nosso levantamento bibliográfico sistemas de classificação de estruturas de proteínas com base em mapas de contatos, mas apenas alguns métodos de comparação e análise desses mapas. Em [Holm e Sander, 1991], os autores apresentam uma metodologia para encontrar subestruturas comuns a um conjunto de proteínas através da análise de suas matrizes de distâncias. As matrizes de distâncias são matrizes quadradas e simétricas assim como os mapas de contatos mas em cada posição (i, j) é apresentada a distância euclidiana 3D em Å do resiíduo i para o j. Em [Lancia et al., 2001], os autores mostram que o problema da sobreposição de mapas de contatos (*contact map overlap*) é NP [1] provando a sua alta complexidade computacional e apresentam um algoritmo para solução ótima para apenas alguns mapas com restrições específicas.

[Caprara et al., 2004] dá continuidade ao trabalho apresentando nova abordagem para solução que inclui outros tipos de mapas mas ainda com restrições. Em [Krasnogor e Pelta, 2004], encontramos a primeira métrica de similaridade baseada em mapas de contatos entre duas proteínas.

Em 2007, foram publicados os dois primeiros servidores *web* para comparação estrutural de proteínas e mapas de contatos. O primeiro deles [Chung et al., 2007] é uma ferramenta que detecta contatos potencialmente conservados em um conjunto de proteínas através de seu alinhamento estrutural. Dessa forma, ele parte de um alinhamento estrutural para alinhar mapas de contatos e buscar contatos preservados. O outro [Barthel et al., 2007] fez um trabalho de integração de várias métricas para comparação estrutural e definição de uma métrica consenso para os casos em que as várias métricas utilizadas divergem muito. Fomos pioneiros nesta área uma vez que o STING, em sua versão Star lançada em 2006 [Neshich et al., 2006b] já apresentava os módulos TopSiMap, Topologs e PCD que são resultados deste projeto e possibilitam ao usuário a comparação de mapas contato visualmente e através de algoritmos, a recuperação de proteínas de mapas de contatos semelhantes.

Os algoritmos de comparação de mapas de contatos desenvolvidos ao longo deste trabalho baseiam-se em algoritmos de processamento digital de imagens e visão computacional. Até o momento, não encontramos outros trabalhos que os utilizem na comparação de mapas de contatos.

[1] Na teoria de complexidade computacional, a classe de complexidade NP (de não-polinomial) é composta por problemas que são decidíveis por uma máquina de Turing não-determinística. [Cormen et al., 2001] Na prática, problemas deste tipo são aqueles cujo trabalho computacional envolvido em sua resolução podem ser descritos como funções não-polinomiais, ou seja, problemas de alta complexidade e para os quais o poder computacional existente não é suficiente para solucionar de forma ótima o problema principalmente para grandes entradas.

1.14 Objetivo geral

Desenvolver um classificador de estruturas de proteínas com base nos contatos intramoleculares entre os resíduos de aminoácidos da cadeia polipeptídica.

1.15 Objetivos específicos

1. Determinação de atributos que sejam componentes essenciais de assinaturas estruturais de proteínas funcionalmente idênticas;

2. Desenvolver um algoritmo que permita a compilação de assinaturas estruturais para cada família de proteínas depositadas no PDB;

3. Construção de uma ferramenta, que será disponibilizada publicamente, para análise e comparação de padrões de contatos entre duas proteínas relacionadas.

Capítulo 2

Materiais e métodos

Neste capítulo, apresentamos um resumo dos materiais e métodos apresentados ao longo das publicações desta tese. Finalizamos este capítulo com explicações dos procedimentos realizados na seleção das bases de dados utilizadas nos experimentos apresentados no capítulo de resultados e discussões que ainda não foram publicados.

2.1 Repositórios públicos de dados

2.1.1 PDB

O PDB (*Protein Data Bank*) [Berman et al., 2000] é atualmente o maior e mais completo repositório de estruturas de proteínas existente e vem experimentando um crescimento exponencial. Ele traz mais de 46.000 arquivos com coordenadas de moléculas e / ou complexos protéicos. Segundo estatísticas do próprio repositório, existe alta redundância de dados sendo aproximadamente 17.000 cadeias com menos de 90% de homologia seqüencial. Para cada cadeia, podem existir dados de diversos mutantes simples ou múltiplos além da existência de múltiplos cenários experimentais nos quais a estrutura foi resolvida.

As principais técnicas utilizadas na resolução de estruturas são a difração de raios-X, a ressonância nuclear magnética (NMR) e a microscopia eletrônica. A grande maioria das estruturas depositadas no PDB foram resolvidas por difração de raios-X. Em média, a resolução é de 2,18Å com desvio padrão de 1,31Å.

2.1.2 SCOP

Muito esforço tem sido feito no intuito de organizar o catálogo de estruturas do PDB. Uma das iniciativas de classificação das cadeias do PDB foi feita pelo SCOP (*Structural Classification of Proteins*) [Brenner et al., 1995]. Na versão atual (1.71) do

SCOP, 27.599 das cerca de 46.000 entradas do PDB foram anotadas o que significa 75.930 cadeias de 1.160 diferentes enovelamentos. Este trabalho foi realizado não só através de softwares mas também de inspeção manual. A classificação deste banco de dados se dá em termos de famílias, super-famílias, enovelamentos e classes. Segundo os autores, proteínas são de uma mesma *família* se tem alta similaridade seqüêncial e estrutural. Proteínas da mesma *super-família* são provavelmente relacionadas evolutivamente compartilhando o mesmo enovelamento e desempenhando funções bastante similares. Proteínas compartilham o mesmo *enovelamento* se possuem o mesmo arranjo arquitetural, ou seja, são estruturalmente muito próximas. As *classes* do SCOP são definidas com base na composição das cadeias em termos de estruturas secundárias: se a maioria é α (formadas, na maioria, por α-hélices) ou β (formadas, na maioria, por folhas β) ou uma junção delas.

O SCOP é muito útil na validação dos resultados deste trabalho uma vez que é uma excelente anotação das cadeias depositadas no PDB. Adicionalmente, são disponibilizados arquivos texto facilmente legíveis por *scripts* nos quais pode-se obter, não só a classificação em termos de classes, enovelamentos, famílias e super-famílias mas também a descrição da cadeia e do organismo (nomenclatura científica e comum) do qual a proteína foi extraída. Neste trabalho, utilizamos a sua classificação com base no enovelamento.

2.1.3 ASTRAL

O PDB é um repositório de dados muito completo e útil para diversas áreas de pesquisa o que também faz com que ele seja muito redundante. Para este trabalho, muitas vezes foi necessário trabalhar com um conjunto não redundante de proteínas. Essa seleção é bastante trabalhosa e deveria excluir seqüências muito similares, estruturas muito redundantes, considerar o organismo da qual ela foi extraída, entre outros aspectos a avaliar. Quando precisamos diminuir a redundância no conjunto de dados recorremos à seleção do ASTRAL [Brenner et al., 2000, Chandonia et al., 2002, Chandonia et al., 2004]. Este banco de dados é parcialmente derivado do SCOP e provê proteínas não redundantes com base em um valor de corte para a similaridade seqüencial das cadeias.

2.1.4 STING

O STING [Neshich et al., 2006b, Neshich et al., 2005, Neshich et al., 2003] é um completo banco de dados acompanhado de várias ferramentas para análise estrutural de proteínas. Seu módulo de contatos [Mancini et al., 2004] possibilita a definição e

análise de interações não covalentes (considerando adicionalmente as pontes dissulfeto). Os autores dividiram as possíveis interações em 14 tipos:

- Contatos hidrofóbicos;
- Contatos carregados atrativos (interações íon-íon);
- Contatos carregados repulsivos (interações íon-íon);
- Pontes de hidrogênio entre cadeia principal e cadeia principal (sem ou com uma ou duas moléculas de água);
- Pontes de hidrogênio entre cadeia principal e cadeia lateral (sem ou com uma ou duas moléculas de água);
- Pontes de hidrogênio entre cadeia lateral e cadeia lateral (sem ou com uma ou duas moléculas de água);
- Empilhamento aromático (interações dipolo induzido-dipolo induzido entre anéis aromáticos);
- Pontes dissulfeto

O STING utiliza a definição de contatos proposta em [Sobolev et al., 1999]. Ele considera pontes de hidrogênio os contatos entre 2,0 e 3,2Å atribuindo a elas 2,6kcal/mol de energia, contatos hidrofóbicos de 2,0 a 3,8Å e 0,6kcal/mol, carregados entre 2,0 e 6,0Å e 10,0kcal/mol, pontes dissulfeto entre 1,5 e 2,8Å e 85,0kcal/mol. Para os empilhamentos aromáticos a energia é 0,5kcal/mol e a distância não foi encontrada na literatura.

2.2 Metodologia para cálculo dos contatos

Nossa metodologia para cálculo dos contatos foi parcialmente baseada em [Sobolev et al., 1999, Neshich et al., 2006b, Neshich et al., 2005, Neshich et al., 2003]. Todos os átomos de cada um dos 20 resíduos de aminoácidos mais comumente encontrados em proteínas foram classificados em uma ou mais das seguintes classes:

- Hidrofóbicos
- Positivos
- Negativos

- Aceptores de ponte de hidrogênio
- Doadores de ponte de hidrogênio
- Aromáticos
- Enxofres

Seguem as classes dos átomos:

- Hidrofóbicos: ALA(CB), ARG(CB, CG, CD), ASN(CB), ASP(CB), CYS(CB), GLN(CB, CG), GLU(CB, CG), HIS(CB, CG, CD2, CE1), ILE(CB, CG1, CG2, CD1), LEU(CB, CG, CD1, CD2), LYS(CB, CG, CD), MET(CB, CG, CE), PHE(CB, CG, CD1, CD2, CE1, CE2, CZ), PRO(CB, CG, CD), THR(CG2), TRP(CB, CG, CD1, CD2, CE2, CE3, CH2, CZ, CZ2, CZ3), TYR(CB, CG, CD1, CD2, CE1, CE2, CZ), VAL(CB, CG1, CG2)
- Positivos: ARG(NH1, NH2), HIS(ND1, NE2), LYS(NZ)
- Negativos: ASP(OD1, OD2), GLU(OE1, OE2)
- Aceptores: ALA(O), ARG(O), ASN(O, OD1), ASP(O, OD1, OD2), CYS(O), GLN(O, OE1), GLU(O, OE1, OE2), GLY(O), HIS(O), ILE(O), LEU(O), LYS(O), MET(O), PHE(O), PRO(O), SER(O), THR(O), TRP(O), TYR(O), VAL(O)
- Doadores: ALA(N), ARG(N, NE, NH1, NH2), ASN(N, ND2, OD1), ASP(N), CYS(N), GLN(N, NE2), GLU(N), GLY(N), HIS(N, ND1, NE2), ILE(N), LEU(N), LYS(N, NZ), MET(N), PHE(N), PRO(N), SER(N, OG), THR(N, OG1), TRP(N, NE1), TYR(N, OH), VAL(N)
- Aromáticos: HIS(CG, ND1, CD2, CE1, NE2), PHE(CG, CD1, CD2, CE1, CE2, CZ), TRP(CG, CD1, CD2, NE1, CE2, CE3, CZ2, CZ3, CH2), TYR(CD1, CD2, CE1, CE2, CG, CZ)
- Enxofre: CYS(S), MET(SD)

Consideramos que dois resíduos de aminoácidos fazem algum tipo de contato se, e somente se:

1. A distância seqüencial entre eles for de, no mínimo, 3 resíduos;
2. Algum dos átomos de um dos resíduos estiver a uma distância tridimensional dentro dos intervalos de corte pré-definidos para suas classes de algum átomo do outro resíduo;
3. Os ângulos entre os átomos não são considerados no cômputo dos contatos.

Definimos entre átomos dessas classes os seguintes tipos de contatos:

Tipo de contato	Classes de átomos	Valor de corte (Å)
Hidrofóbicos	ambos hidrofóbicos	entre 2 e 3,8
Carregados atrativos	positivos e negativos	entre 2 e 6
Carregados repulsivos	ambos positivos ou negativos	entre 2 e 6
Pontes de hidrogênio	aceptores e doadores	entre 2 e 3,2
Empilhamentos aromáticos	ambos aromáticos	entre 3 e 8
Pontes dissulfeto	ambos enxofre	entre 1,5 e 2,8

Tabela 2.1: Tipos de contatos e seus valores de corte.

2.3 Seleção das bases de dados para os experimentos

Para verificar a precisão dos classificadores propostos foi necessário selecionar um conjunto de proteínas de um enovelamento específico e outro conjunto de enovelamentos diferentes e variados. O objetivo dos experimentos foi calcular a precisão dos classificadores na recuperação de elementos da família específica misturados com outras de enovelamentos diferentes. Utilizamos o banco de dados SCOP na seleção das proteínas uma vez que ele as divide de acordo com o enovelamento.

Selecionamos as Globinas como enovelamento modelo e, adicionalmente, verificamos a precisão dos classificadores com outras famílias diferentes. Seguem as famílias trabalhadas:

- Globinas
- Apolipoproteínas
- Plastocianinas
- RBPs (*Retinol binding proteins*)
- Tioredoxinas

As Globinas (Figura 2.1(a)) são as proteínas responsáveis pelo transporte de moléculas de oxigênio nos músculos e no sangue e estão entre as mais bem estudadas proteínas. São compostas exclusivamente por α-hélices. As Apolipoproteínas (Figura 2.1(b)), também compostas exclusivamente por α-hélices, são proteínas que ligam lipídios e constituem as Lipoproteínas do plasma. São importantes no transporte dos lipídios ingeridos através do fluxo sangüineo do intestino para o fígado e de lipídios sintetizados pelo organismo para os tecidos que os armazenam, metabolizam e secretam. As Plastocianinas (Figura 2.1(c)) são proteínas envolvidas no transporte de elétrons na

fotossíntese. Contêm um átomo de cobre e são compostas basicamente por folhas-β em um arranjo em forma de barril. As RBPs (Figura 2.1(d)), também proteínas predominantemente compostas por folhas-β, têm função relacionada com o transporte de Retinol e são responsáveis por solubilizar e estabilizar ligantes hidrofóbicos em solução aquosa. Tioredoxinas (Figura 2.1(e)) são proteínas compostas por uma mistura de α-hélices e folhas-β. Atuam como anti-oxidantes facilitando a redução de outras proteínas.

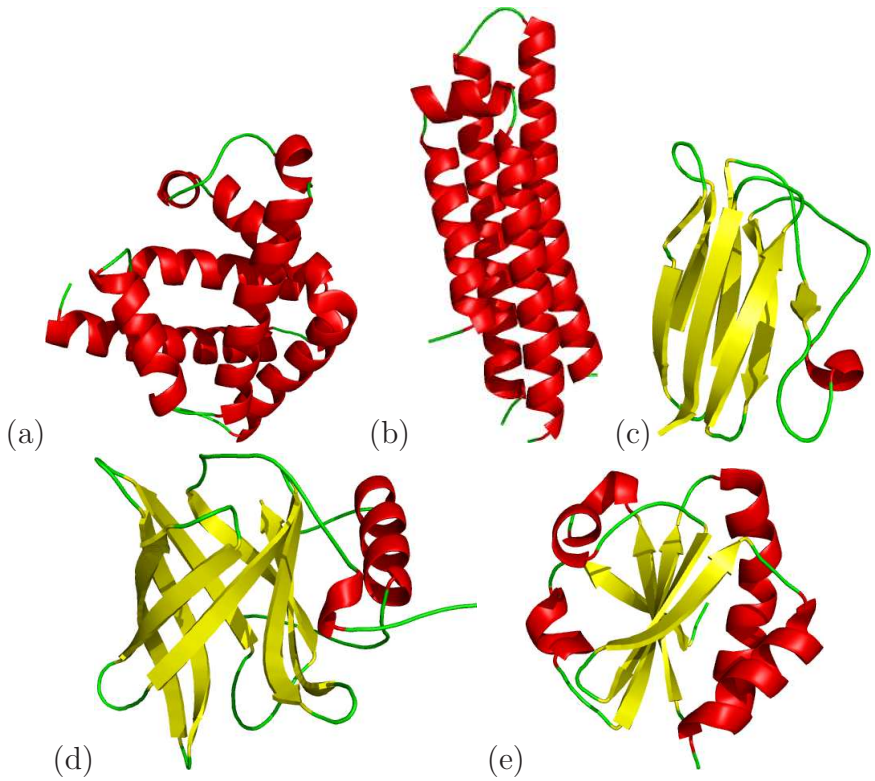


Figura 2.1: Tipos de enovelamentos utilizados nos testes deste trabalho: (a) Globina (PDB id 1a6mA) (b) Apolipoproteína (PDB id 1nfnA) (c) Plastocianina (PDB id 1plcA) (d) RBP (PDB id 1rbpA) (e) Tioredoxina (PDB id 2trxA).

2.3.1 Seleção das Globinas

A consulta pelo enovelamento Globina na versão atual do banco de dados SCOP retornou 1.356 exemplares de Globinas. Percebemos que algumas dessas cadeias possuíam domínios Globina juntamente com outros tipos de domínios, como é o caso da Flavo-hemoglobina ilustrada na Figura 2.2. Por esse motivo, fizemos uma verificação manual verificando se cada cadeia de Globina indicada representava mesmo apenas o domínio Globina.

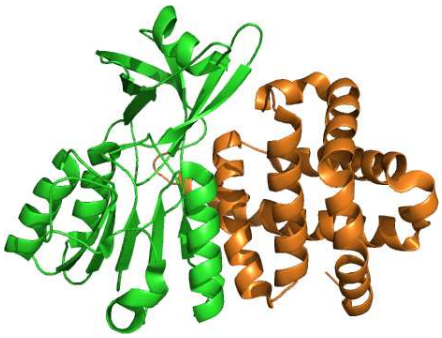


Figura 2.2: Flavohemoglobina: exemplo de cadeia de proteína com domínio Globina jutamente com outro domínio. Proteínas multi-domínio, tais como esta, foram excluídas da nossa base de dados.

Do conjunto curado de Globinas foram selecionados 50 exemplares que foram alinhados utilizando o software PriSM [Yang e Honig, 1999] e são apresentados na Figura 2.3. O PriSM é um software para análise e modelagem de proteínas que tem duas vantagens em relação a outros pacotes: suporta o alinhamento de um grande número de cadeias e não utiliza nenhum parâmetro para realizar os alinhamentos.

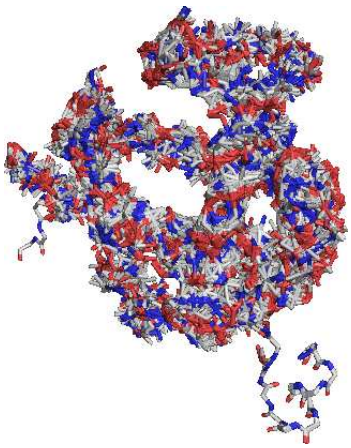


Figura 2.3: Alinhamento estrutural dos 50 exemplares de Globinas utilizados neste trabalho. Para obter maior clareza, exibimos apenas os átomos da cadeia principal das proteínas.

Exibimos, no Anexo A, os alinhamentos das seqüências dos 50 exemplares de Globinas utilizados neste trabalho.

2.3.1.1 Seleção das Mioglobinas

Além de selecionar proteínas variadas do enovelamento Globina, optamos por selecionar um subconjunto bastante homogêneo deste enovelamento. Selecionamos outra

base de dados composta pelas Mioglobinas. Na versão atual do SCOP (1.71), há 217 cadeias destas proteínas. São 151 provenientes de baleia, 7 de cavalo marinho, 1 de foca, 33 de porco, 20 de cavalo, 1 humana, 1 de elefante, 2 de tartaruga e 1 de atum. Selecionamos mais uma vez 50 exemplares de Mioglobinas de forma a manter os exemplares de espécies menos comuns no PDB e balanceando a escolha de espécies mais comuns, eliminando alguns deles.

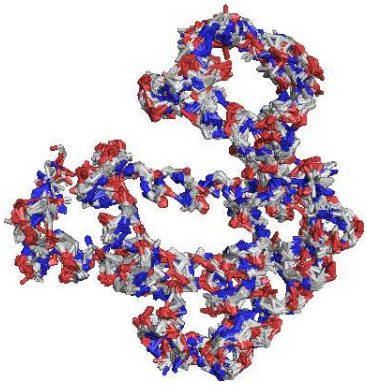


Figura 2.4: Alinhamento estrutural dos 50 exemplares de Mioglobinas utilizados neste trabalho.

No Anexo A, apresentamos o alinhamentos das seqüências destas Mioglobinas.

2.3.2 Seleção das proteínas de enovelamentos variados

Como as Globinas têm cerca de 150 resíduos de aminoácidos, as Apolipoproteínas 190, as Plastocianinas 100, as RPBS 180 e as Tioredoxinas 110, selecionamos do SCOP 50 cadeias aleatóriamente dentre aquelas cujo número de resíduos de aminoácidos estava dentro do intervalo [100,200]. Nesse conjunto temos proteínas α, β, $\alpha/\ \beta$ e $\alpha+\beta$. Acreditamos que proteínas com números de resíduos muito diferentes dificilmente seriam confundidas uma vez que o número de contatos a comparar seria também muito diferente.

2.4 Métricas para comparação dos mapas de contatos

Nesta seção, mostraremos como a abordagem de casamento de imagens é utilizada para medir a similaridade estrutural de duas proteínas com base em seus mapas de contato. Em particular, exploramos 2 diferentes paradigmas no tratamento deste problema:

- O paradigma de *recuperação de imagens com base no conteúdo* (RIBC) resolvido com uma métrica baseada nas caraterísticas das imagens, o *correlogramo de cores* (CC);

- O paradigma de *registro de imagens* (RI) que solucionamos com duas técnicas baseadas na similaridade das imagens: *raio médio de dispersão* (RMD) e *earth mover's distance* (EMD).

A RIBC é uma disciplina científica amplamente baseada na noção de que é possível comprimir imagens preservando sua semântica [Pentland et al., 1994]. As imagens são comprimidas em um vetor assinatura de menor tamanho possível, visando a eficiência de possíveis consultas às bases de assinaturas. Usualmente, esses vetores assinatura são computados com base em atributos de baixo nível extraídos diretamente das imagens tais como cores, texturas ou primitivas geométricas e seus relacionamentos espaciais na imagem que provêm informações semânticas de alto nível [Mojsilovic et al., 2004].

Uma forte motivação para aplicação deste tipo de técnica é o crescimento das bases de proteínas como o próprio PDB. A indexação dessas bases de dados é uma operação computacionalmente cara mas, uma vez criados os vetores assinatura, a pesquisa é bastante eficiente.

O paradigma de RI [Brown, 1992] é usualmente utilizado na comparação de imagens de um mesmo objeto que sofre transformações não rígidas
[Maintz e VIergever, 1998]. Um custo é atribuído para cada deformação que o objeto precisa sofrer e a dissimilaridade entre as imagens é computada como sendo o mínimo custo para deformar uma imagem na outra.

A motivação pela qual aplicamos este tipo de técnica é que proteínas de seres distintos evoluíram de moléculas ancestrais e suas distâncias filogenéticas devem estar fortemente correlacionadas com a dissimilaridade estrutural. Assim, se pudéssemos, de alguma forma, modelar as deformações necessárias para transformar um mapa de contatos de uma primeira proteína em um mapa de uma outra proteína como uma seqüência de transformações que imitariam os efeitos da evolução na sua estrutura, a similaridade estrutural entre essas proteínas poderia ser calculada como a seqüência de transformações de custo mínimo.

Existe um compromisso na escolha desses diferentes paradigmas. As técnicas de RIBC tendem a ser mais eficientes em grandes conjuntos de dados mas, por outro lado, as técnicas de RI tendem a ser mais acuradas, pelo menos na comparação de imagens próximas.

2.4.1 A abordagem de recuperação de imagens com base no conteúdo

Para especificar completamente o funcionamento do algoritmo de RIBC, é necessário definir como o vetor assinatura de cada possível imagem é gerado e como a similaridade entre tais vetores é computada [Del-Bimbo, 1999].

O CC [Huang et al., 1997] expressa como a correlação de pares de cores se altera com a distância. Especifica a probalidade de se encontrar um pixel de cor j a uma distância k de outro pixel de cor i. Seja I uma imagem $n \times n$ com espaço de cores quantizado em m cores $c_1, ..., c_m$. Seja a distância $d \leq n$ um parâmetro de entrada para o sistema. Assim, o correlogramo de I é definido para $i, j \in [m], k \in [d]$ como

$$\gamma_{c_i,c_j}^{(k)}(I) \triangleq \underset{p_1 \in I_{c_i}, p_2 \in I}{Prob} \left[p_2 \in I_{c_j} \mid |p_1 - p_2| = k \right], \tag{2.1}$$

onde a notação $p_1 \in I_{c_i}$ significa que a cor do pixel p_1 na imagem I é c_i, isto é, que $p_1 \in I, I(p_1) = c_i$.

Para computar o correlogramo, temos que avaliar a seguinte equação:

$$\gamma_{c_i,c_j}^{(k)}(I) = \frac{\Gamma_{c_i,c_j}^{(k)}(I)}{h_{c_i} \cdot 8k}, \tag{2.2}$$

onde h_{c_i} é o valor do histograma de cores de c_i e

$$\Gamma_{c_i,c_j}^{(k)} \triangleq \left| \left\{ p_1 \in I_{c_i}, p_2 \in I_{c_j} | \ |p_1 - p_2| = k \right\} \right|. \tag{2.3}$$

O algoritmo mais ingênuo para calcular esta expressão é de $O(n^2 d^2)$. Porém, usando a versão com programação dinâmica, também proposta em [Huang et al., 1997] o algoritmo seria $O(n^2 d)$. Note que, como o número de cores em nossas imagens é muito reduzido, não avaliamos o custo do algoritmo com base no número de cores.

A métrica do correlogramo é relativamente insensível a elementos individuais do vetor. Ela corresponde, entretanto, a uma média ponderada das discrepâncias de todo o conjunto de características das assinaturas das imagens. No caso de dois correlogramos das imagens I e I', estes pesos são inversamente proporcionais a $\gamma_{c_i,c_j}^{(k)}(I) + \gamma_{c_i,c_j}^{(k)}(I')$, isto é, quanto maior este termo é, menor a influência do par de cores (c_i, c_j) na medida final. Mais especificamente, a métrica d para os correlogramos das imagens I e I' é:

$$|I - I'|_{\gamma,d_1} \triangleq \sum_{\substack{i,j \in [m], \\ k \in [d]}} \frac{|\gamma_{c_i,c_j}^{(k)}(I) - \gamma_{c_i,c_j}^{(k)}(I')|}{1 + \gamma_{c_i,c_j}^{(k)}(I) + \gamma_{c_i,c_j}^{(k)}(I')}, \tag{2.4}$$

onde o 1 no denominador evita a divisões por zero. Note que, depois de construídos

os correlogramos, o cálculo da métrica é $O(n)$, o que garante a eficiência na resposta a consultas mesmo em grandes bases de dados.

Mostraremos um exemplo de aplicação da técnica com a utilização de dois mapas de contatos hipotéticos. Na Figura 2.5, apresentamos 2 mapas de contatos 5×5 e contendo 3 tipos de contatos: vermelhos, verdes e azuis. Queremos computar a dissimilaridade entre eles através do CC de forma bastante simplificada.

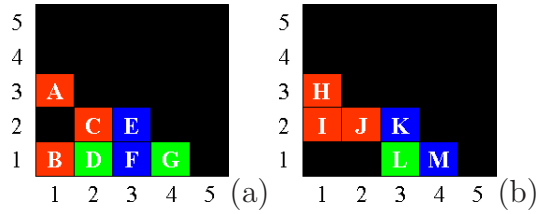


Figura 2.5: Mapas de contatos hipotéticos a serem comparados nos exemplos.

Para computar a dissimilaridade entre os mapas de contato é necessário, primeiramente, computar os histogramas de distribuição espacial das cores. Para tal, medimos a distância de todos os pixels coloridos a todos os outros pixels da mesma cor (conforme Tabelas 2.2, 2.3 e 2.4). As tabelas de distâncias são, obviamente, simétricas de forma que consideremos apenas uma das metades. Como a imagem tem tamanho 5x5, a maior distância possível seria $\sqrt{18}$ ou 4,24, uma vez que não consideramos a diagonal que é sempre 0. O histograma vai ter então 4 posições sendo que a primeira significa o número de pixels que distam de 1 a 2 (exclusive), a segunda de 2 a 3 (exclusive) e assim por diante.

Tabela 2.2: Distâncias entre os pixels vermelhos de cada imagem no exemplo.

	A	B	C
A	0	2	1
B	2	0	1
C	1	1	0

	H	I	J
H	0	1	1
I	1	0	1
J	1	1	0

	D	G
D	0	2
G	2	0

	L
L	0

Tabela 2.3: Distâncias entre os pixels verdes de cada imagem no exemplo.

Para a cor vermelha, temos o seguinte vetor de freqüências $F_{A_{vermelho}} = (2; 1; 0; 0)$ que resulta nas seguintes probabilidades $P_{A_{vermelho}} \approx (0,66; 0,34; 0; 0)$ e $F_{B_{vermelho}} =$

	D	G
D	0	2
G	2	0

	L
L	0

	E	F
E	0	1
F	1	0

	K	M
K	0	1
M	1	0

Tabela 2.4: Distâncias entre os pixels azuis de cada imagem no exemplo.

$(3; 0; 0; 0)$ que resulta em $P_{B_{vermelho}} = (1; 0; 0; 0)$. Somando os módulos das diferenças entre cada posição dos vetores obtemos $0,34+0,34 = 0,68$. Para normalizar, dividimos este valor pelo número de pixels vermelhos nos dois mapas obtendo $0,68/6 \approx 0,11$. De forma similar teremos $F_{A_{verde}} = (0; 1; 0; 0)$ e $F_{B_{verde}} = (0; 0; 0; 0)$ uma vez que não existem pares de contatos verdes no mapa B. Teremos $P_{A_{verde}} = (0; 1; 0; 0)$ e $P_{B_{verde}} = (0; 0; 0; 0)$ resultando em dissimilaridade 1. Teremos também $F_{A_{azul}} = (1; 0; 0; 0)$ e $F_{B_{azul}} = (1; 0; 0; 0)$, resultando em vetores de probabilidade idênticos e dissimilaridade 0. Dividindo pelo número de contatos verdes $1/3 \approx 0,33$. O resultado final é a soma das dissimilaridades para todas as cores e, nesse caso, seria $0,11 + 0 + 0,33 = 0,44$.

2.4.2 A abordagem de registro de imagens

2.4.2.1 O raio médio de dispersão

Esta técnica é baseada em [Kutulakos, 2000], onde é introduzido o conceito de *transformações de embaralhamento*. Estas são transformações geométricas onde embaralha-se pixels por no máximo um raio de dispersão r.

O uso deste tipo de transformação na análise da dissimilaridade estrutural de proteínas é atraente porque sua natureza espacialmente localizada preserva características geométricas de alto nível, assim como as transformações evolucionárias na estrutura primária das proteínas fazem na estrutura.

Neste trabalho, fizemos uma adaptação desta ideia e definimos o conceito de raio médio de dispersão, $\hat{r}_{disp}$, entre duas imagens como a distância Euclidiana entre pixels em uma imagem e o pixel da mesma cor mais próximo na outra imagem. Mais formalmente, o raio médio de dispersão entre duas imagens $n \times n$ é dado por:

$$\hat{r}_{disp}\,(I, I') \triangleq \frac{1}{2\,n^2} \sum_{i,j \in [n]} r(I, I', i, j) + r(I', I, i, j), \tag{2.5}$$

onde

$$r(I, I', i, j) \triangleq \min_{\substack{x,y \in [n], \\ I(i,j) = I'(x,y)}} \left[\sqrt{(x-i)^2 + (y-j)^2} \right]. \tag{2.6}$$

O algoritmo ingênuo para esta computação tem custo $O(n^4)$. Entretanto, pré-computando, para cada cor $c_i, i \in [m]$, a transformada de distância relativa aos pixels

da imagem I de cor c_i usando o algoritmo de *Chamfer* (que é $O(n^2)$) e repetindo esse procedimento para a imagem I', reduzimos este custo para $O(n^2)$. Após essa pré-computação, cada termo $r(I, I', i, j)$ na Equação (2.5) é processado em $O(1)$, apenas pela busca na posição (i, j) na transformada de distância relativa aos pixels de I' que têm a cor $I(i, j)$.

Na prática, todos os pixels brancos foram excluídos dos cálculos uma vez que representam ausência de contatos. Como os mapas de contatos são matrizes bastante esparsas, criamos listas auxiliares de $O(n)$ elementos de forma a responder as consultas em tempo $O(n)$.

Finalmente, observe que dois mapas de contatos a serem comparados tem na grande maioria das vezes tamanhos diferentes. Para superar este problema, reescalamos todos os mapas de contatos para o tamanho 1000×1000.

Mostraremos um exemplo de aplicação do RMD com os mapas da Figura 2.5. Para computar a dissimilaridade entre dois mapas devemos encontrar pixels de cada cor nos mais próximos na segunda imagem (conforme Tabelas 2.5, 2.6 e 2.7).

	H	I	J
A	0	1	1
B	2	1	1
C	1	1	0

Tabela 2.5: Distâncias entre os pixels vermelhos entre o par de imagens no exemplo.

	L
D	1
G	1

Tabela 2.6: Distâncias entre os pixels verdes entre o par de imagens no exemplo.

	K	M
E	0	1
F	1	1

Tabela 2.7: Distâncias entre os pixels azuis entre o par de imagens no exemplo.

Os custos computados serão dados pelas distâncias entre os pixels casados. Assim, teremos $A \to H$ com custo 0, $B \to I$ com custo 1, $C \to J$ com custo 0. Como o índice deve ser simétrico, fazemos na ordem inversa e obtemos os seguintes mapeamentos $H \to A$ com custo 0, $I \to A$ com custo 1 e $J \to C$ com custo 0. Note que quando

existem duas opções de mesmo custo, escolhemos arbitrariamente entre as opções. Somando todos estes custos e dividindo pelo número de contatos vermelhos nos dois mapas obtemos $(1 + 1)/6 \approx 0,33$. Para o tipo verde, teremos $D \rightarrow L$ com custo 1 e $G \rightarrow L$ com custo 1. No sentido inverso, $L \rightarrow D$ com custo 1. Normalizando, teremos $(1 + 1 + 1)/3 = 1$. Os mapeamentos do tipo azul serão $E \rightarrow K$ com custo 0, $F \rightarrow K$ com custo 1 e no sentido inverso $K \rightarrow E$ com custo 0 e $M \rightarrow E$ com custo 1. Normalizando, teremos $(1 + 1)/4 = 0,5$. Totalizando, $0,33 + 0,5 + 1 = 1,83$.

2.4.2.2 O *earth mover's distance*

Uma possível limitação da métrica descrita na subseção anterior é que ela permite que múltiplos contatos em um mapa casem com o mesmo contato do outro. Assim, a métrica não é capaz de diferenciar entre grupamentos densos e esparços de contatos. Esta limitação pode ser evitada com o uso da métrica *earth mover's distance* (EMD).

A utilização desta métrica em bases de imagens foi inicialmente proposta em [Rubner et al., 1998]. Especificamente, o trabalho sugere o uso da métrica em assinaturas de images com base em intensidade ou histograma de cores, por exemplo. Neste trabalho, aplicamos a técnica diretamente nos mapas de contato o que faz com que a técnica seja baseada em similaridade e não característica.

A ideia por trás do EMD é tratar cada pixel colorido em uma mapa de contato como uma unidade de terra espalhada por um espaço de tamanho conhecido e os pixels em um segundo mapa de contato como buracos com capacidade para uma unidade de terra no mesmo espaço. A cor de cada unidade de terra ou buraco é dada de acordo com a cor dos pixels. O EMD mede a quantidade de trabalho necessário para preencher os buracos com terra, com a restrição de que buracos de uma cor podem ser apenas preenchidos com terra da mesma cor.

Como proposto em [Rubner et al., 1998], a computação do EMD é equivalente a resolver o famoso *problema do transporte*. Mais especificamente, o EMD é obtido encontrando o conjunto de fluxos não-negativos $f_{i,j,x,y}, g_{x,y}$ que minimize o trabalho total do carregador de terra, w, definido como:

$$w\left(I, I'\right) \triangleq \sum_{i,j,x,y \in [n]} f_{i,j,x,y}\, d(i,j,x,y) \; + \sum_{x,y \in [n]} g_{x,y}\, d_{max}, \tag{2.7}$$

onde

$$d(i,j,x,y) \triangleq \begin{cases} \sqrt{(x-i)^2 + (y-j)^2}, \text{ if } I(i,j) = I'(x,y), \\ \infty, \text{ caso contrário,} \end{cases} \tag{2.8}$$

sujeito às seguintes restrições:

$$\forall_{x,y\in[n]} \left[\sum_{i,j\in[n]} f_{i,j,x,y} + g_{x,y} = 1 \right], \tag{2.9}$$

$$\forall_{i,j\in[n]} \left[\sum_{x,y\in[n]} f_{i,j,x,y} = 1 \right]. \tag{2.10}$$

Na Equação (2.7), o fator $d(i, j, x, y)$ corresponde ao custo de mover uma unidade de massa do local (i, j) na imagem I para a posição (x, y) na imagem I'. Na mesma equação, d_{max} é uma penalidade para cada buraco deixado vazio devido ao número de pixels daquela cor na imagem I ser menor que na imagem I'. Este é um parâmetro de entrada para o algoritmo. A Equação (2.9) garante que todo buraco será preenchido com uma unidade de massa ou uma penalidade d_{max} será aplicada. Finalmente, a Equação (2.10) garante que cada pixel na imagem I será fornecedor de apenas uma unidade de terra.

A métrica final é normalizada em relação ao fluxo total:

$$d_{em}(I, I') \triangleq \frac{1}{n^2} w_{em}(I, I'). \tag{2.11}$$

A solução padrão para o problema do transporte envolve o uso do método *simplex* [Dantzig, 1951] no qual, no pior caso, o custo computacional é exponencial. Felizmente, este caso é extremamente raro e, no caso médio, o custo é proporcional ao número de restrições [Wagner, 1986]. Se considerássemos todos os pixels de cada mapa de contato, o custo seria $O(n^6)$. Desconsiderando novamente os pixels brancos, o custo médio seria $O(n^3)$.

Mostraremos, agora, o exemplo da aplicação do EMD para os mesmos mapas de contatos da Figura 2.5. Como nossos mapas tem 3 tipos de contatos, devemos resolver 3 modelos do problema do transporte separadamente.

Façamos os cálculos para os pixels vermelhos. Considerando que o custo de pontos não casados é 3, teremos que minimizar a seguinte equação: $w_{vermelho}(I, I') = 0F_{AH} + 1f_{AI} + 1f_{AJ} + 2f_{BH} + 1f_{BI} + 1f_{BJ} + 1f_{CH} + 1f_{CI} + 0f_{CJ} + 3g_A + 3g_B + 3g_C$. Os coeficientes são os custos de se mapear um pixel no outro, ou seja, as distâncias entre eles. A minimização é sujeita às seguintes restrições:

$$\begin{gathered} f_{AH} + f_{AI} + f_{AJ} + g_A = 1 \\ f_{BH} + f_{BI} + f_{BJ} + g_B = 1 \\ f_{CH} + f_{CI} + f_{CJ} + g_C = 1 \\ f_{AH} + f_{BH} + f_{CH} = 1 \end{gathered}$$

$$f_{AI} + f_{BI} + f_{CI} = 1$$
$$f_{AJ} + f_{BJ} + f_{CJ} = 1$$

Estas restrições indicam que cada ponto da imagem (a) pode cair em, no máximo, um ponto da imagem (b). Caso não exista ponto em (b) para receber um ponto de (4), um custo adicional é aplicado. Além disto, cada ponto da imagem (b) pode receber, no máximo, um ponto de (a). Minimizando a expressão, verificamos as seguintes correspondências: $A \rightarrow H$ com custo 0, $B \rightarrow I$ com custo 1 e $C \rightarrow J$ com custo 0. Observe que $w_{vermelho}(I, I') = 1/6 \approx 0,16$.

Para os pixels verdes minimizamos $W_{verde}(I, I') = 1f_{DL} + 1f_{GL} + 3g_D + 3g_G$ com as seguintes restrições:

$$f_{DL} + g_D = 1$$
$$f_{GL} + g_G = 1$$
$$f_{DL} + f_{GL} = 1$$

Obtemos $G \rightarrow L$ com custo 1 e D fica sem mapeamento gerando um custo 3. Logo, $w_{verde}(I, I') = 4/3 \approx 1,33$.

Para os pixels azuis minimizamos $w_{azul}(I, I') = 0f_{EK} + 1f_{EM} + 1f_{FK} + 1f_{FM} + 3g_E + 3g_F$ com as seguintes restrições:

$$f_{EK} + f_{EM} + g_E = 1$$
$$f_{FK} + f_{FM} + g_F = 1$$
$$f_{EK} + f_{FK} = 1$$
$$f_{EM} + f_{FM} = 1$$

Obtemos $E \rightarrow K$ com custo 0 e $F \rightarrow M$ com custo 1, logo $w_{azul}(I, I') = 1/4 = 0,25$.

A dissimilaridade final será dada por $w(I, I') = w_{vermelho}(I, I') + w_{verde}(I, I') + w_{azul}(I, I') = 0,16 + 1,33 + 0,25 = 1,74$.

2.5 Algoritmo para definição de assinaturas estruturais

2.5.1 Determinação dos agrupamentos de contatos

De acordo com [Guting, 1994], as informações sobre os contatos com as quais trabalhamos nos mapas de contatos são dados espaciais. No intuito de definir as assinaturas estruturais da famílias de proteínas, precisamos ser capazes de identificar automaticamente agrupamentos de contatos em cada mapa.

Para tal tarefa, existem inúmeros algoritmos descritos na literatura de mineração de dados. Há basicamente dois tipos de algoritmos [Kaufman e Rousseeuw, 1990]: os de particionamento e os hierárquicos. Os algoritmos de particionamento constroem partições da base de dados D que possui n objetos em um conjunto de k agrupamentos. Normalmente k é um parâmetro de entrada para estes algoritmos o que é indesejável no nosso caso. O algoritmo começa com uma partição arbitrária e vai refinando esta de forma a otimizar a função objetivo. Os algoritmos hierárquicos criam uma decomposição hierárquica de D. Esta decomposição é representada por um dendograma, uma árvore resultante da divisão iterativa de D. Neste caso, não existe o parâmetro de entrada k mas é necessário definir a condição de parada nas divisões da árvore.

Optamos por utilizar o DBSCAN [Ester et al., 1996] que é um algoritmo de particionamento baseado em densidade. A vantagem deste método é a capacidade de identificar não somente agrupamentos tipicamente esféricos mas sim de qualquer forma. A idéia principal do método consiste no cálculo da densidade que implica que cada ponto de um cluster precisa ter um número mínimo de pontos a um raio r definido arbitrariamente, ou seja, sua densidade precisa superar um determinado valor de corte.

Assim, o algoritmo implementado consiste em sortear um contato aleatoriamente no mapa e, dado o raio r, incluir os contatos que se encontram a uma distância euclidiana menor ou igual a este raio. O processo segue iterativamente com a adição dos pontos que estão dentro do raio r dos pontos recém-adicionados até que não restem pontos a adicionar. Neste caso, um novo contato não pertencente ao agrupamento definido é sorteado para iniciar um novo agrupamento. O processo se repete até que não existam pontos fora dos agrupamentos. Obviamente, há que se definir uma densidade mínima para definição dos agrupamentos.

2.5.2 Separação dos clusters definidos incorretamente

A transformada de Hough [Hough, 1962] foi desenvolvida em 1962 para detectar características analiticamente representáveis em imagens binarizadas, assim como linhas, círculos e elipses. Para detectar uma linha, Hough utilizou a equação decilive-intercepto definida por $y = ax + b$. Usando uma matriz acumuladora, examina-se cada ponto e calcula-se os parâmetros da equação a e b. Incrementa-se, então, o acumulador referente aos parâmetros ($A[a, b]$). Após o processamento de todos os pontos, procura-se os picos da matriz acumuladora sendo estes os indicadores de possíveis linhas na imagem.

Neste trabalho, utilizamos esta transformada para dividir agrupamentos que são unidos pelo DBSCAN, mas na verdade são linhas perependiculares entre si. Neste caso, através dos picos, somos capazes de verificar se um agrupamento contém apenas

uma ou se é a união de várias linhas. Sendo a união, fazemos a separação dos pontos com base nas suas distâncias às possíveis retas reveladas pela transformada.

2.5.3 Definição dos vetores característicos dos agrupamentos

Uma vez definidos os agrupamentos e sendo eles lineares, nomeamos cada cluster por um vetor que o caracteriza. Os vetores são definidos de forma simplificada por um ponto origem e um ponto destino. O ponto origem é o ponto de menor x e o de destino, o de maior x.

2.5.4 Métrica para comparação das assinaturas

Para comparar os conjuntos de vetores característicos de um mapa (assinatura) com os de outros utilizamos a mesma métrica EMD definida na seção 2.4.2.2 porém ao invés de usar os pontos referentes aos contatos utilizamos os pontos representativos dos vetores da assinatura.

2.6 Estratégia de avaliação dos classificadores utilizando curvas ROC

Nesta seção, apresentamos os conceitos necessários para o entendimento de nossa estratégia de avaliação das métricas propostas.

Matrizes de confusão [Kohavi, 2004] contêm informação sobre as classes reais e preditas dos objetos e possibilitam avaliar o desempenho de sistemas de classificação.

As curvas ROC (*Receiver Operating Characteristics*) [Fawcett, 2006] são uma outra forma de avaliação destes sistemas. Em uma curva ROC, plotamos no eixo x a taxa de falsos positivos e, no eixo y a taxa de verdadeiros positivos. A taxa de falsos positivos consiste no número de instâncias negativas preditas como positivas dividido pelo número de instâncias negativas, a taxa de verdadeiros positivos o número de instaâncias positivas preditas como positivas dividido pelo número de instâncias positivas.

No espaço da curva, o ponto $(0, 1)$ indica números de um classificador perfeito: classifica todas as instâncias positivas e negativas corretamente. Neste ponto a taxa de falsos positivos é 0 e a de verdadeiros positivos é 1. O ponto $(0, 0)$ representa o classificador que prediz todas as instâncias como negativas e o ponto $(1, 1)$, positivas. Já o ponto $(1, 0)$ é o classificador que erra todas as predições.

Em muitos casos, os classificadores possuem parâmetros que precisam ser estimados para elevar a taxa de verdadeiros positivos (às vezes com o custo de se elevar também a taxa de falsos positivos) ou diminuir a taxa de falsos negativos (possivelmente reduzindo

também a taxa de verdadeiros positivos). Cada conjunto de valores selecionados para os parâmetros geram um ponto (*taxa de falsos positivos*, *taxa de verdadeiros positivos*) e uma série destes pontos é usada para plotar a curva ROC. Neste trabalho, o parâmetro que precisa ser estimado é o valor de corte usado na decisão se uma instância pertence ou não a uma família de proteínas.

Uma vantagem desta abordagem é que as curvas ROC são independentes da distribuição das classes e encapsulam toda a informação contida nas matrizes de confusão uma vez que a taxa de falsos negativos é complementar à taxa de verdadeiros positivos e a de verdadeiros negativos à de falsos positivos. Estas curvas provêm uma ferramenta visual para avaliação do compromisso entre a identificação correta de todas as instâncias positivas e as instâncias negativas incorretamente classificadas. Outra característica muito interessante é que a área sob a curva pode ser usada como uma medida de precisão dos sistemas de classificação. Outra métrica de precisão muito utilizada é a distância de um ponto ao ponto $(0, 1)$ (representativo do classificador perfeito).

Neste trabalho, todas as medidas de precisão dos classificadores com as famílias estudadas baseiam-se na área sob a curva ROC média entre todas as curvas para proteínas da família.

Capítulo 3

Publicações

Neste capítulo, apresentamos as publicações geradas com resultados desta tese. Uma cópia dos artigos é apresentada no Anexo B.

3.1 *An image-matching approach to protein similarity analysis*

O artigo [Fernandes-Jr. et al., 2004] é o primeiro trabalho integrante desta tese. Foi apresentado em 2004 no XVII Simpósio Brasileiro de Processamento de Imagens e Computação Gráfica que aconteceu em Curitiba.

Neste trabalho, apresentamos a idéia de modelar o problema de comparação estrutural de proteínas como um problema de comparação entre imagens coloridas. Para cada proteína, produzimos o mapa de contatos utilizando os cálculos de interações não-covalentes do STING [Neshich et al., 2003]. Estes mapas de contatos são compostos por pontes de hidrogênio, interações hidrofóbicas e contatos carregados atrativos.

Inicialmente, implementamos um algoritmo de processamento de imagens baseado no paradigma de *recuperação de imagens com base no conteúdo*. Segundo este paradigma, é possível comprimir imagens e uma base de dados preservando sua semântica. Para cada imagem, uma assinatura é construída de forma que a base resultante indexada é pesquisada de forma bastante eficiente. Esta compressão é feita através da extração de características como cores, texturas e primitivas geométricas (linhas, segmentos, curvas, fronteiras, junções, etc.). Na modelagem proposta, cada tipo de interação não covalente é modelada como uma cor na imagem de forma que analisamos a distribuição espacial das cores da imagem. Este algoritmo é denominado *Correlogramo de Cores* e foi considerado bastante interessante dado o tamanho das bases de dados de estruturas de proteínas existentes atualmente.

Em seguida, implementamos outro algoritmo baseado no paradigma de *registro de*

imagens. Ele mede quão similares duas proteínas são calculando o custo de se deformar os mapas de contatos de uma convertendo-a no mapa da outra. Chamamos esta métrica de *Raio Médio de Dispersão*. Este paradigma é muito utilizado no casamento de um mesmo objeto que sofre deformações não-rígidas em diversas imagens. Uma forte motivação para a aplicação deste idéia é que proteínas evoluíram de ancestrais comuns e a sua distância filogenética é fortemente correlacionada com a sua dissimilaridade estrutural. Dessa forma tentamos modelar as alterações necessárias para transformar uma proteína em outra pelas deformações necessárias para ajustar um mapa de contato a outro.

Para testar esta metodologia utilizamos um conjunto de 28 proteínas de diferentes enovelamentos entre proteínas α, β e $\alpha\beta$. Usamos como família modelo as Mioglobinas, coletadas de 9 diferentes espécies: baleia, cavalo, elefante, tartaruga, cavalo marinho, foca, porco, ser humano e atum. Comparando todas as proteínas da base com a Mioglobina humana, verificamos que a métrica baseada no Correlogramo de Cores recuperou 6 das 8 Mioglobinas (dentre as 8 proteínas consideradas mais parecidas com a *query*) enquanto a baseada no Raio Médio de Dispersão recuperou todos os exemplares.

Este trabalho apresentou como principal resultado a possibilidade de se comparar estruturas de proteínas através de seus mapas de contatos. Tivemos uma primeira indicação de que existe um padrão de contatos em cadeias de proteínas de uma família e que este deve ser um importante componente da assinatura estrutural desta família.

3.2 *A contact-map matching approach to protein structure similarity analysis*

No artigo anterior [Fernandes-Jr. et al., 2004], propusemos uma modelagem baseada em casamento de imagens para analisar a similaridade entre estruturas de proteínas através de seus mapas de contatos. Os resultados foram promissores apesar de os experimentos terem sido feitos com poucos exemplares de Mioglobinas e de proteínas de outras famílias diversas.

Neste trabalho [Melo et al., 2006], montamos uma base de dados mais apropriadas para confirmar os resultados do artigo anterior. Selecionamos todos os monômeros de proteínas de enovelamentos diversos:

- 224 *Globinas*, as proteínas responsáveis pelo transporte de oxigênio no sangue e músculos;
- 13 *Apolipoproteínas*, lipoproteínas compostas por um conjunto de 4 α-hélices;

- 15 *Plastocianinas*, proteínas transportadoras de elétrons compostas, na maior parte, por folhas-β;
- 18 *Retinol-Binding Proteins* (R.B.P.s), composta por um barril de folhas-β;
- 8 *Tioredoxinas*, compostas por uma mistura de α-hélices e folhas-β.

Nosso objetivo foi tentar recuperar proteínas de cada uma destas cinco famílias misturadas a uma base de 187 outros monômeros selecionados do PDB.

O classificador baseado no Correlogramo de Cores apresentou precisões entre 89,12% e 98,44% enquanto o baseado no Raio Médio de Dispersão, entre 81,69% e 99,84%.

Além destas análises de precisão na recuperação de proteínas de uma mesma família dentre outras de famílias diversas, analisamos a habilidade dos classificadores em ordenar as proteínas da mesma família em termos de dissimilaridade de estruturas. Alinhamentos estruturais entre as proteínas *query* e outras proteínas da família mostraram que os índices de dissimilaridade calculados pelas métricas propostas possuem alta correlação com o R.M.S.D. dos alinhamentos estruturais.

Com este trabalho, mostramos que as métricas propostas apresentaram excelentes resultados na recuperação de proteínas de diversas famílias e composições em termos de estruturas secundárias assim como na ordenação de proteínas de mesmo enovelamento em termos da similaridade estrutural.

3.3 *Similarity-based versus feature-based analysis of structural protein similarity*

Neste manuscrito [Melo et al., 2008], introduzimos uma nova técnica que acreditamos poder elevar as precisões dos nossos classificadores. A técnica de registro de imagens apresentada em [Fernandes-Jr. et al., 2004] possibilita que mais de um contato de um primeiro mapa seja casado com um contato do segundo mapa. Por acreditar que isto poderia causar algum problema na medição da dissimilaridade entre os mapas, propusemos neste trabalho uma métrica baseada no *Earth Mover's Distance*.

Esta métrica modela o primeiro mapa como um conjunto de montes de terra a ser movido para buracos, que são os contatos do segundo mapa. A dissimilaridade dos mapas é dada pelo trabalho de se mover os montes de terra do primeiro mapa para o segundo. O trabalho é medido pela distância entre os pontos onde se localizar os contatos nos dois mapas. Cada monte de terra pode ser movido para um, e somente um, buraco. Cada buraco, por sua vez, pode receber um, e somente um, monte de terra. Este é um famoso problema de otimização que consiste em escolher quais montes serão movidos para buraco de forma a realizar o mínimo de trabalho possível.

Para nossa surpresa, observamos que os resultados da nova métrica proposta foram pouco superiores que as da métrica do Raio Médio de Dispersão. De fato, para famílias mais conservadas estruturalmente, a métrica anterior já tinha excelentes resultados na recuperação das Apolipoproteínas e R.B.P.s. Para as outras famílias, conseguimos uma melhoria com a nova métrica.

3.4 *Mining structural signatures of proteins*

Neste trabalho [Melo et al., 2007a], apresentamos uma metodologia para busca de assinaturas estruturais em proteínas baseada no padrão de contatos em cada cadeia. Utilizando técnicas de mineração de dados, exploramos uma base de mapas de contatos no aspecto de localização espacial dos contatos no intuito de evidenciar uma assinatura estrutural que defina a família de proteínas.

Nos experimentos, foram usados exemplares de Mioglobinas, Apolipoproteínas, Plastocianinas, R.B.P.s e Tioredoxinas. Visualizando os mapas de contatos de proteínas de uma mesma família, verificamos que os padrões de contatos apresentados por cada família, são agrupamentos de contatos hidrofóbicos (os grupos são formados por contatos não-locais) ou pontes de hidrogênio (os grupos são formados por contatos locais). Optamos assim por testar nossa abordagem com estes dois tipos de contatos inicialmente.

Para detectar automaticamente os agrupamentos presentes nos mapas de contatos de nossa base, utilizamos um algoritmo de *clustering* baseado em densidade, o *DBSCAN*. Este algoritmo é capaz de tratar uma importante característica dos mapas de contatos que outros algoritmos deste tipo não são capazes: mapas de contatos possuem agrupamentos de formato linear que são sempre paralelos ou anti-paralelos à diagonal do mapa.

A intenção deste trabalho foi identificar segmentos de reta representativos de cada agrupamento de um mapa de contato e, finalmente, verificar se estes segmentos de reta estão ou não presentes em todos os exemplares de um família de proteínas. De fato, esta representação facilita o reconhecimento de padrões relevantes. Todavia, muitos dos agrupamentos identificados pelo DBSCAN apresentavam forma de "L". Isto ocorre sempre que dois agrupamentos se tocam. Nestes casos, o segmento de reta identificado fica totalmente distorcido. Para solucionar este problema, usamos a *transformada de Hought*, que ajuda a identificar se um cluster encontrado pelo DBSCAN é realmente um segmento de reta ou vários.

Finalmente, obtivemos através desta metodologia assinaturas para cada mapa de contato. Essas assinaturas consistem de um conjunto de vetores. Estes vetores têm

sempre direção paralela ou perpendicular à diagonal do mapa e a direção foi arbitrária de forma que a origem esta sempre à esquerda e o destino à direira.

Além de caracterizar cada mapa de contato com uma assinatura, propusemos uma metodologia de classificação de estruturas baseada nestas. Fomos capazes de recuperar Mioglobinas de um conjunto de Mioglobinas e não-Mioglobinas com uma precisão de 95%, o que mostra que cada assinatura realmente apresenta um padrão para a família.

3.5 *Finding protein-protein interaction patterns by contact map matching*

Neste trabalho [Melo et al., 2007b], apresentamos uma nova possível aplicação para as metodologias desenvolvidas de comparação e classificação de mapas de contatos. Ela consiste na definição de padrões de interações entre cadeias, ou seja, na interface entre cadeias proteicas de um complexo.

Para tal, propomos um novo tipo de mapas de contatos. Neste mapa, o eixo x representa uma cadeia e o y, a outra. Dessa forma, os mapas representam os contatos entre 2 cadeias, não mais sendo quadrados e simétricos como acontece com os mapas de contatos tradicionais.

Para os experimentos, foram selecionadas cadeias de *Serino-Proteases* por serem umas das mais estudadas proteínas que se apresentam complexadas com outras cadeias. Encontramos no banco de dados SCOP essa molécula complexada com 12 diferentes tipos de inibidores. Escolhemos trabalhar com o *Bovine Pancreatic Trypsin Inhibtor* (B.P.T.I.) por ser o inibidor com mais exemplares no PDB. As Serino-Proteases que encontramos complexadas com o B.P.T.I foram *Tripsinas*, *Quimotripsinas*, *Trombinas*, *Matriptases* e *Kalikreínas*.

Utilizamos o algoritmo de comparação entre mapas de contatos para gerar os índices de dissimilaridade entre as moléculas e posteriormente utilizamos os índices para gerar uma árvore na qual cada complexo Serino-Protease - B.P.T.I. é ligado ao complexo mais parecido em termos de contatos de interface. Verificamos que os complexos com o mesmo tipo de Serino-Protease tenderam a se agrupar, conforme esperado, o que nos dá indícios de que a metodologia utilizada para classificar cadeias também pode ser utilizada com sucesso para classificar mapas de interação proteína-proteína.

Adicionalmente, neste trabalho propusemos uma nova utilização para o algoritmo baseado no *Earth Mover's Distance*: fazer o alinhamento dos mapas de contatos. A idéia consiste em considerar como alinhados os contatos que forem casados pelo algoritmo de otimização. Verificamos que os alinhamentos foram corretos e obtivemos contatos conservados em todos os complexos. O algoritmo foi capaz de identificar

contatos conservados entre resíduos bem descritos na literatura por estarem no sítio catalítico da proteína ou no trecho conhecido como "*oxianion hole*".

3.6 *The STAR sting server: a multiplatform environment for protein structure analysis*

Finalmente, apresentamos o artigo da versão STAR do pacote de programas de análise estrutural de proteínas Sting [Neshich et al., 2006b]. Alguns dos resultados desta tese foram incorporados à esta versão do programa na forma dos módulos: P.C.D., TopSiMap e Topologs.

O *Protein Contacts Difference* (P.C.D.) é um módulo que oferece um relatório comparativo entre os contatos de duas cadeias proteicas. Ele apresenta os contatos conservados, novos e extintos de uma cadeia para outra. Através de seu código de cores, é possível identificar os tipos de contatos. É uma ferramenta muito útil na análise dos contatos conservados e modificados no caso de mutações na seqüência de resíduos, apresentando no relatório a distância tridimensional dos contatos ao resíduos mutantes.

O TopSiMap (*Topological Similarity Map*) é uma ferramenta de análise comparativa entre a topologia de proteínas através de mapas de contatos. Neste programa, é possível ver duas cadeias proteicas alinhadas bem como comparar seus mapas de contatos que podem ser visualizados de forma interativa. O usuário pode selecionar apenas os contatos preservados entre dois mapas, os contatos que existem em apenas um dos mapas, fazer uma filtragem por contatos de cada tipo, por contatos com o intermédio de moléculas de água, podem aproximar o mapa e pode visualizar os contatos selecionados na estrutura da proteína através do plug-in JMol ou Chime.

O Topologs ASTRAL 40 é um banco de dados de classificação estrutural de proteínas com base em seus padrões de contatos. O subconjunto do PDB apresentado no banco de dados ASTRAL 40 teve seus mapas de contatos computados e processados pelos nossos algoritmos de comparação de mapas de contatos. Isto torna possível, para cada cadeia desta base, selecionar as 100 cadeias de mapas de contatos mais parecidos. Além disto, é possível verificar os alinhamentos estruturais assim como analisar interativamente os mapas de contatos entre uma cadeia e as 100 mais similares.

Estes sistemas foram implementados utilizando perl para os scripts de tratamento de dados de coordenadas atômicas provenientes do e Java e jsp para a implementaÃ§ão do servidor web.

Capítulo 4

Resultados e discussões

4.1 Calibração dos classificadores

Dois dos classificadores propostos neste trabalho (Correlogramo de cores e *Earth movers distance*) são paramétricos. Por esse motivo, utilizamos a base de Mioglobinas para calibrar estes classificadores, ou seja, obter o melhor valor aproximado para estes parâmetros.

4.1.1 Correlogramo de cores

O parâmetro a ser calibrado no Correlogramo de cores é a distância d. Este é o valor máximo de distância entre dois contatos do mesmo tipo que terão a sua freqüência computada no vetor assinatura. Na Figura 4.1, plotamos as curvas ROC para $5 \leq d \leq 100$. A precisão de cada configuração é especificada no gráfico.

Observamos que a precisão do classificador cresce a medida que o valor d aumenta. Isto já era esperado uma vez que quanto maior o raio de varredura mais informação acrescentamos ao classificador sob pena de aumentar o tempo de execução, obviamente. Como, por definição $d \leq n$, continuamos aumentando o valor do raio até 200 que é o maior tamanho de cadeia da nossa base de mapas de contatos. Apresentamos na Figura 4.2 a precisão dos classificadores com o aumento do valor d. Observe que enquanto $d \leq 100$, a precisão é crescente (sendo a taxa de crescimento dessa precisão descrescente). Para $d > 100$, não verificamos aumento expressivo da precisão. Portanto, optamos por utilizar $d = 100$ em todos os experimentos deste trabalho.

4.1.2 *Earth mover's distance*

A métrica EMD possui o parâmetro de entrada d_{max}. Todas as vezes que comparamos dois mapas de contatos que tem números de contatos de um mesmo tipo diferentes,

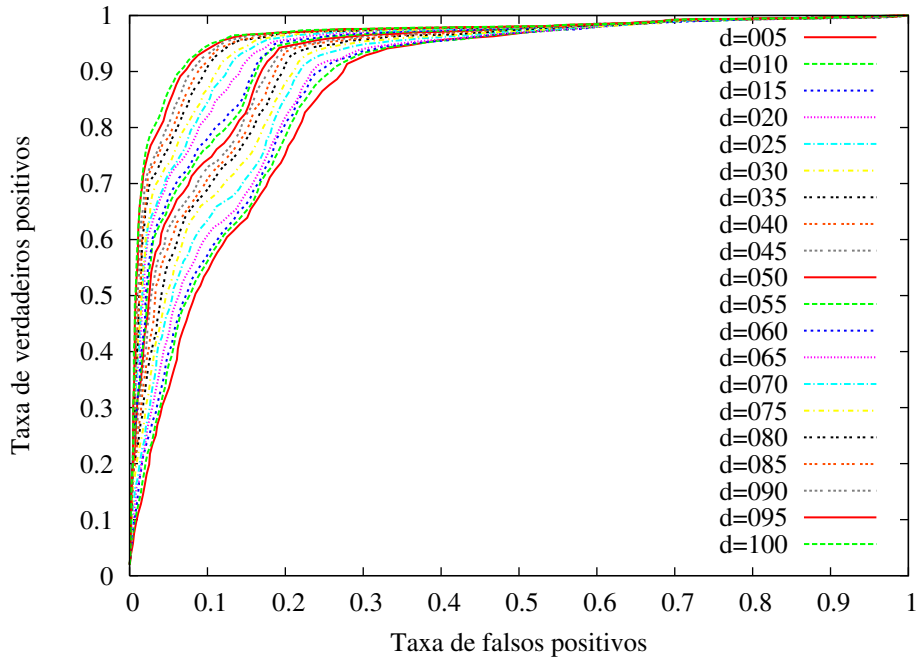


Figura 4.1: Curvas ROC do Correlogramo de cores com a variação do parâmetro de raio máximo de varredura d.

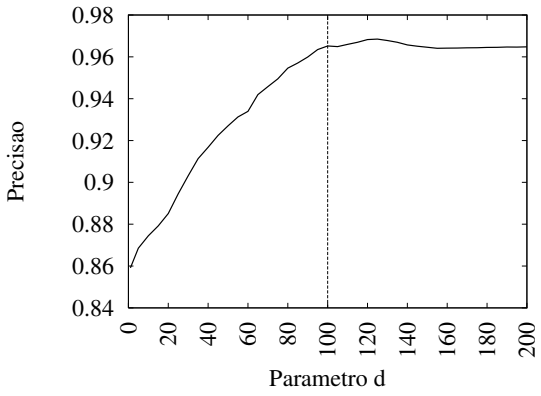


Figura 4.2: Variação da precisão do classificador baseado no CC com o aumento do parâmetro d.

a penalidade d_{max} será somada ao custo de transformar um mapa no outro, ou seja, à dissimilaridade entre os mapas. Este valor foi calibrado, de forma idêntica ao procedimento aplicado para calibrar o parâmetro da métrica anterior, através de curvas ROC. Apresentamos na Figura 4.3 a variação da precisão deste classificador com o aumento do parâmetro d_{max}. O ponto $d_{max} = 35$ é o ponto onde obtemos maior precisão na classificação.

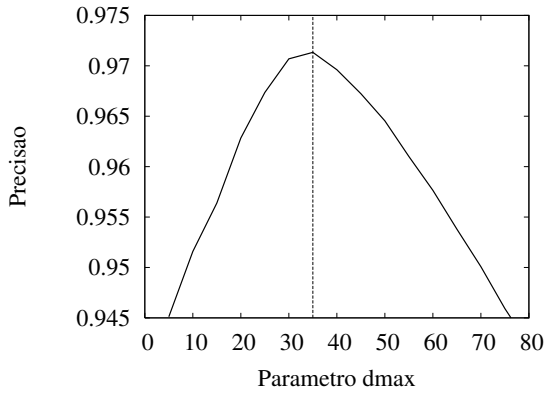


Figura 4.3: Variação da precisão do classificador baseado na métrica com o aumento do parâmetro d_{max}.

4.2 Análise dos atributos dos contatos usados na classificação

4.2.1 Tipos de contatos

Mostramos que é possível classificar estruturas de proteínas através dos padrões de interações hidrofóbicas, pontes de hidrogênio (sem água) e contatos carregados atrativos. Posteriormente, decidimos verificar se os três tipos de contatos eram igualmente conservados e portanto importantes como atributos para classificação estrutural de cadeias protéicas. Tentamos, então recuperar Mioglobinas dentre as proteínas de enovelamentos variados utilizando-nos separadamente de cada um dos três tipos iniciais trabalhados (contatos hidrofóbicos, pontes de hidrogênio sem água e contatos carregados atrativos). Conforme podemos ver na Figura 4.4, a precisão foi maior utilizando apenas pontes de hidrogênio (99,17%) ou contatos hidrofóbicos (98,80%) do que com a configuração com os três tipos de contatos da configuração proposta inicialmente. A classificação teve sua precisão reduzida em 19,5%, em comparação com a configuração inicial, quando utilizamos apenas os contatos carregados atrativos. Portanto, este tipo

de interação mostra-se menos conservado que as interações hidrofóbicas e pontes de hidrogênio, em *Mioglobinas*.

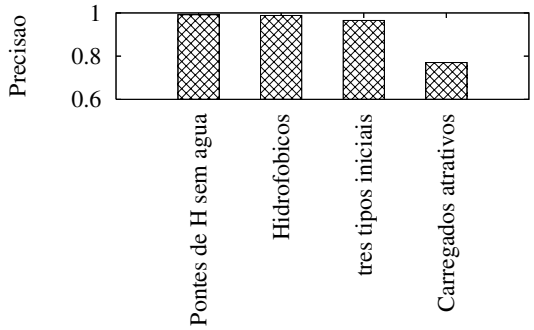


Figura 4.4: Análise comparativa da precisão da classificação de Mioglobinas utilizando a métrica CC com a configuração inicial e com os contatos hidrofóbicos, pontes de hidrogênio (sem moléculas de água) e contatos carregados atrativos separadamente.

Posteriormente, adicionamos os outros tipos de interações: carregados repulsivos, empilhamentos aromáticos e pontes dissulfeto. A Figura 4.5 mostra que os resultados com estes tipos de interações alcançaram precisões abaixo das obtidas pelos tipos de contatos iniciais. Uma observação importante é a baixíssima precisão das pontes dissulfeto. Este tipo de interação é inexistente em Globinas de forma que não pode ser utilizado para recuperação de cadeias dessas proteínas. O que ocorre neste caso é que toda cadeia que não possua ponte dissulfeto, e com qualquer enovelamento, é considerada idêntica a uma Globina. As precisões obtidas foram 93,56%, 69,92% e 33,69% com empilhamentos aromáticos, contatos carregados repulsivos e pontes dissulfeto, respectivamente.

Em relação às pontes de hidrogênio, sabemos que estas possuem diferentes papéis na estruturação das proteínas. Pontes de hidrogênio têm papel fundamental na formação das estruturas secundárias. Nas α-hélices, por exemplo, átomos da cadeia principal de resíduos i compartilham hidrogênios com átomos da cadeia principal de resíduos $i + 4$. Folhas-β também são formadas com pontes de hidrogênio entre resíduos distantes na seqüência. O STING computa pontes de hidrogênio e as disponibiliza aos seus usuários separadamente de acordo com os átomos que participam da interação: se são átomos da cadeia principal ou da cadeia lateral. Nos experimentos discutidos até o momento utilizamos as pontes de hidrogênio indistintamente, ou seja, tratamos pontes de hidrogêno entre átomos da cadeia principal (MC-MC), átomo da cadeia principal e átomo da cadeia lateral (MC-SC) e átomos das cadeias laterais (SC-SC) como se fossem o mesmo tipo de interação. A Figura 4.6 mostra o que acontece com a precisão

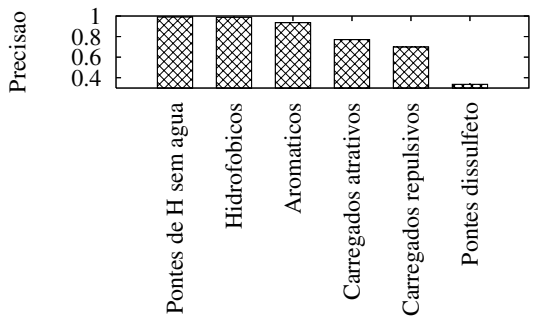


Figura 4.5: Análise comparativa da precisão da classificação de Mioglobinas utilizando a métrica CC com pontes de hidrogênio (sem moléculas de água), contatos hidrofóbicos, contatos carregados atrativos e repulsivos, empilhamentos aromáticos e pontes dissulfeto.

dos classificadores se separamos as pontes de hidrogênios em diferentes qualidades e as tratamos como se fossem diferentes atributos. Neste gráfico podemos observar que a melhor configuração para as pontes de hidrogênio é quando as consideramos indistintamente. Isto indica que este tipo de contato é altamente conservado espacialmente em proteínas mas não é muito específico em termos de localização atômica. Isto é, dois resíduos podem fazer pontes de hidrogênio entre diferentes átomos (sendo eles de cadeia principal ou lateral) e esta variação da localicação atômica não parece ser tão relevante para estruturação da proteína. Observamos também que as pontes envolvendo átomos da cadeia principal são bem mais conservados que aqueles envolvendo átomos da cadeia lateral. Possivelmente isto é explicado pelo fato de a cadeia principal ter bem menos graus de liberdade que a cadeia lateral.

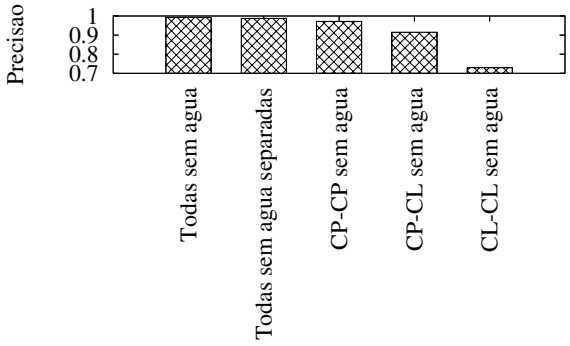


Figura 4.6: Análise comparativa da precisão da classificação de Mioglobinas utilizando a métrica CC com diferenres tratamentos de pontes de hidrogênio.

Finalmente, calculamos a precisão do classificador utilizando pontes de hidrogênio com intermédio de uma molécula de água, conforme pode ser verificado na Figura 4.7. Observamos que a precisão caiu em 24,48%. Isto mostra que provavelmente as moléculas de água aprisionadas nos cristais de proteínas não são muito conservadas na família das Globinas.

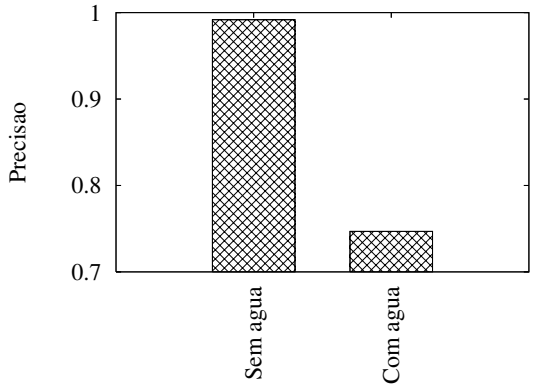


Figura 4.7: Análise comparativa da precisão da classificação de Mioglobinas utilizando a métrica CC com pontes de hidrogênio com e sem intermédio de moléculas de água.

Finalmente, apresentamos na Figura 4.8 as precisões da classificação de Mioglobinas com todas as variações nos tipos de contatos.

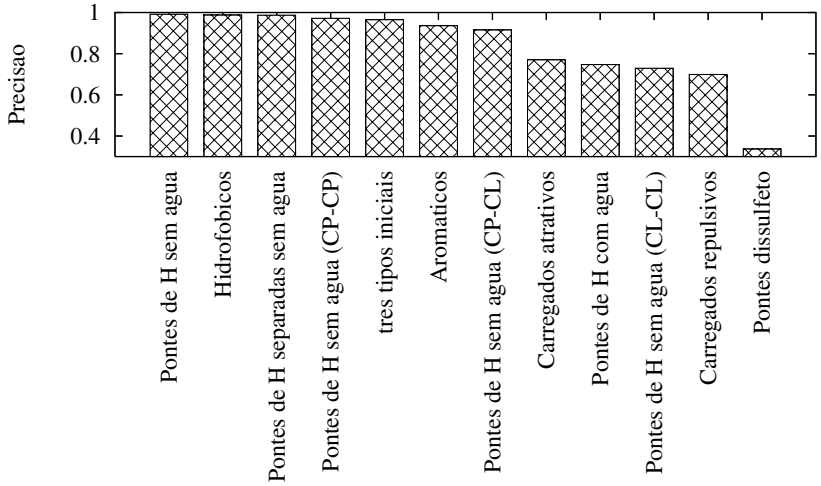


Figura 4.8: Análise comparativa da precisão da classificação de Mioglobinas utilizando a métrica CC com todas as variações de tipos de contatos.

Em relação aos contatos hidrofóbicos, utilizamos primeiramente o valor de corte

padrão sugerido pelo STING. Posteriormente, verificamos que este valor não possibilitava a seleção de todos os contatos hidrofóbicos [Silveira et al., 2008]. Como pode ser observado na Figura 4.9, o valor de corte para definição de contatos hidrofóbicos que maximiza a precisão da classificação é em torno de 7Å.

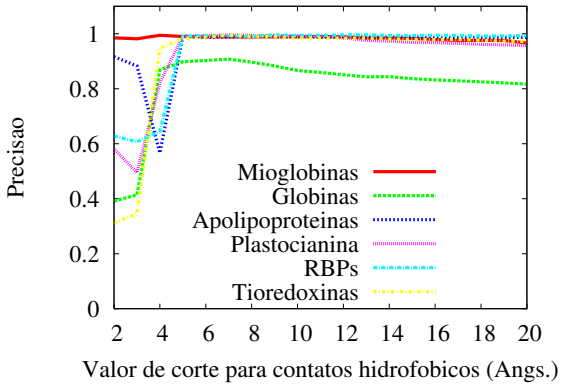


Figura 4.9: Variação da precisão da classificação utilizando interações hidrofóbicas com a variação do valor de corte para definição dos contatos hidrofóbicos.

4.2.2 Eliminação dos contatos de curta distância seqüencial

A Figura 4.10(a) mostra um histograma no qual apresentamos as freqüencias das distâncias seqüenciais entre resíduos que fazem qualquer tipo de contato em todas as cadeias presentes no PDB. Em (b), exibimos os mesmos dados, porém para valores de distância seqüencial menor que 100 resíduos. Observe que a grande maioria dos contatos são locais, ou seja, ocorrem entre resíduos com 10 ou menos resíduos de separação na cadeia polipeptídica. Verificamos neste experimento a variação da precisão com a eliminação de contatos próximos seqüencialmente. Observamos na Figura 4.11 que quando desconsideramos estes contatos a precisão decresce progressivamente o que indica que os contatos locais são conservados e, portanto, importantes na definição do enovelamento e da assinatura estrutural de famílias de proteínas.

4.2.3 Eliminação dos contatos com resíduos pouco conectados

Um resíduo de aminoácido pode fazer interações químicas não covalentes com vários outros resíduos da cadeia. Verificamos neste experimento se resíduos muito conectados são mais conservados que resíduos pouco conectados. A Figura 4.12 mostra a freqüencia do número de contatos por resíduo em todo o PDB. A grande maioria dos resíduos faz contatos com menos de 5 outros resíduos.

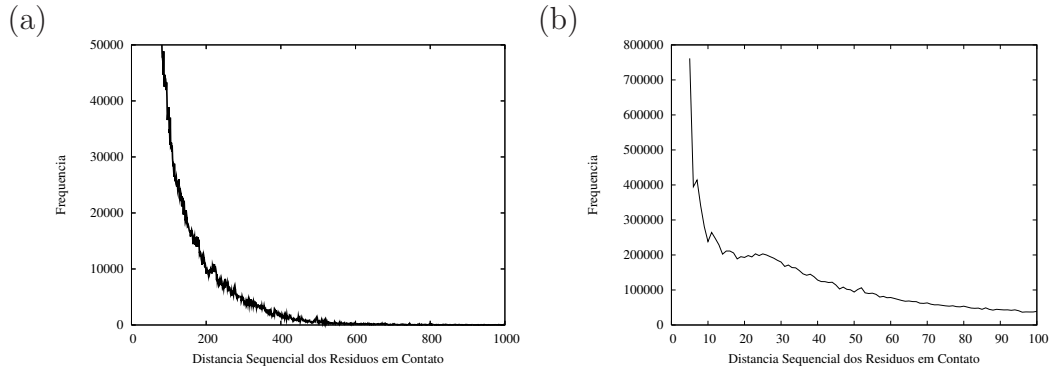


Figura 4.10: Freqüência dos valores de distância seqüencial de resíduos em contato em todo o PDB.

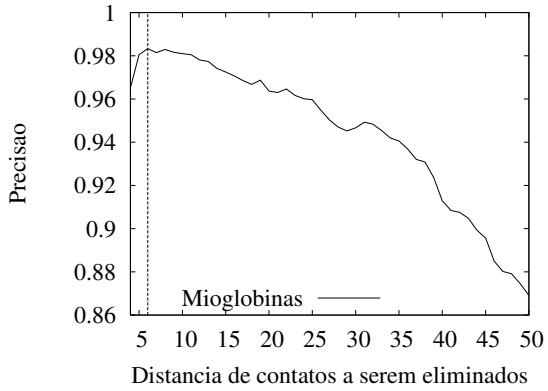


Figura 4.11: Variação da precisão com a eliminação de contatos próximos seqüencialmente.

Na Figura 4.13, constatamos que mesmo os contatos entre resíduos pouco conectados parecem ser importantes na definição do enovelamento de uma proteína e que, ao considerar apenas contatos entre resíduos que atuam como *hubs* em proteínas, estamos perdendo informação. Portanto, neste trabalho, não detectamos conservação suficiente para classificar proteínas apenas usando resíduos muito conectados.

4.3 Resultados finais com a melhor configuração dos sistemas de classificação

Os melhores resultados obtidos foram com a utilização de contatos hidrofóbicos e pontes de hidrogênio. Os contatos hidrofóbicos mostraram-se mais conservados no valor de corte 7Å. Já com as pontes de hidrogênio, verificamos que há um aumento na precisão quando consideramos indistintamente contatos de cadeia principal e lateral e sem intermédio de moléculas de água. Testamos o classificador com Globinas e

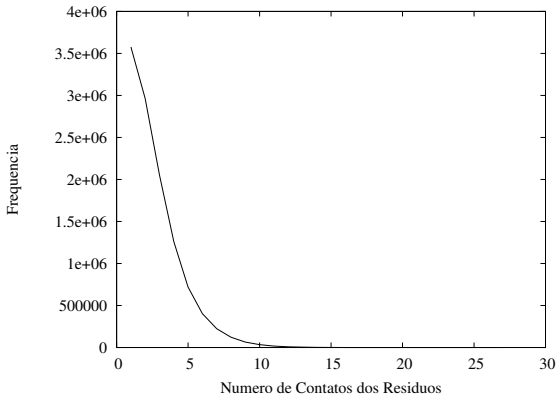


Figura 4.12: Freqüencia dos números de contatos de um resíduo com outros resíduos em todo o PDB.

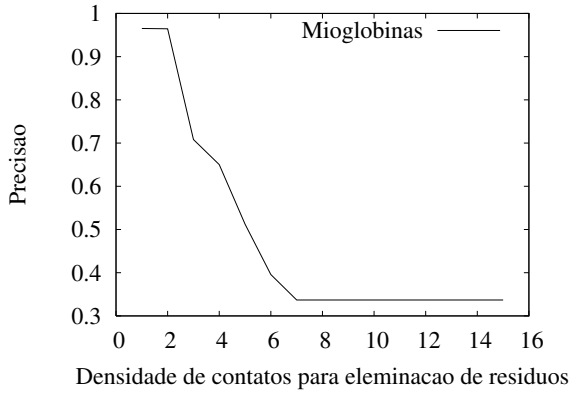


Figura 4.13: Variação da precisão com a eliminação de contatos com resíduos que fazem contatos com poucos resíduos.

Mioglobinas além de outras famílias de tamanhos parecidos mas enovelamentos bastante variados: Apolipoproteínas, Plastocianincas, RBPs e Tioredoxinas. Para todas as famílias obtivemos uma precisão média de 94,04% com contatos hidrofóbicos e de 97,89% com as pontes de hidrogênio. A menor precião obtida foi de 79,10% na recuperação de RBPs por contatos hidrofóbicos e a maior foi de 99,20% na recuperação de Plastocianinas utilizando pontes de hidrogênio.

4.4 Contribuições deste trabalho no software STING

Nesta subseção, mostramos alguns softwares que foram desenvolvidos com resultados desta pesquisa em parceria com o Dr. Goran Neshich, do CNPTIA/EMBRAPA de

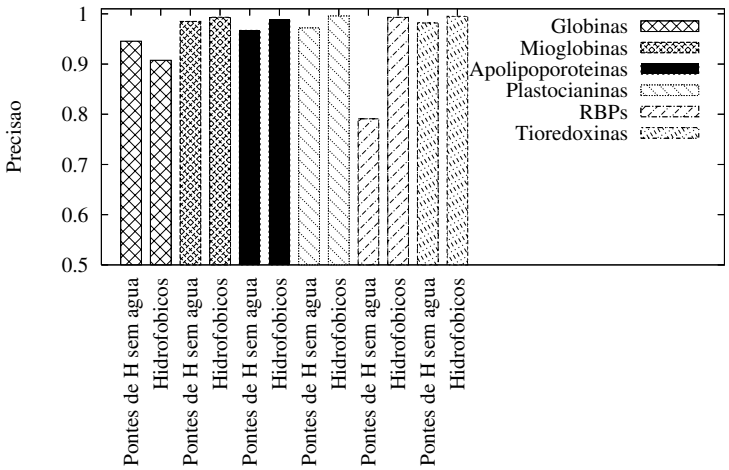


Figura 4.14: Precisão dos classificadores com a melhor configuração utilizando contatos hidrofóbicos e pontes de hidrogênio sem água para variadas famílias de proteínas.

Campinas, co-orientador deste trabalho. Eles estão hoje incorporados ao pacote Blue Star STING [Neshich et al., 2006a].

4.4.1 PCD

No PCD ou *Protein Contacts Difference* os usuários obtêm um relatório completo comparativo das interações intra-cadeia para quaisquer duas cadeias no formato PDB. O programa retorna uma lista de interações que foram preservadas nas duas cadeias assim como uma lista daquelas que constam em apenas uma delas. O sistema também possibilita a comparação de uma cadeia selvagem e sua mutante simples analisando os contatos alterados e sua distância em relação ao resíduo mutado.

4.4.2 TopSiMap

O TopSiMap é um módulo que também possibilita a comparação entre os contatos de duas cadeias PDB. Ele plota as figuras dos mapas de contatos de cada cadeia e é bastante interativo possibilitando a seleção de tipos de contatos, variação das distâncias dos contatos e seleção daqueles que são preservados ou não. Este módulo também possibilita a visualização dos contatos selecionados nas duas moléculas alinhadas através do *plugin* chime ou JMol. Existe também um relatório das energias envolvidas nos contatos.

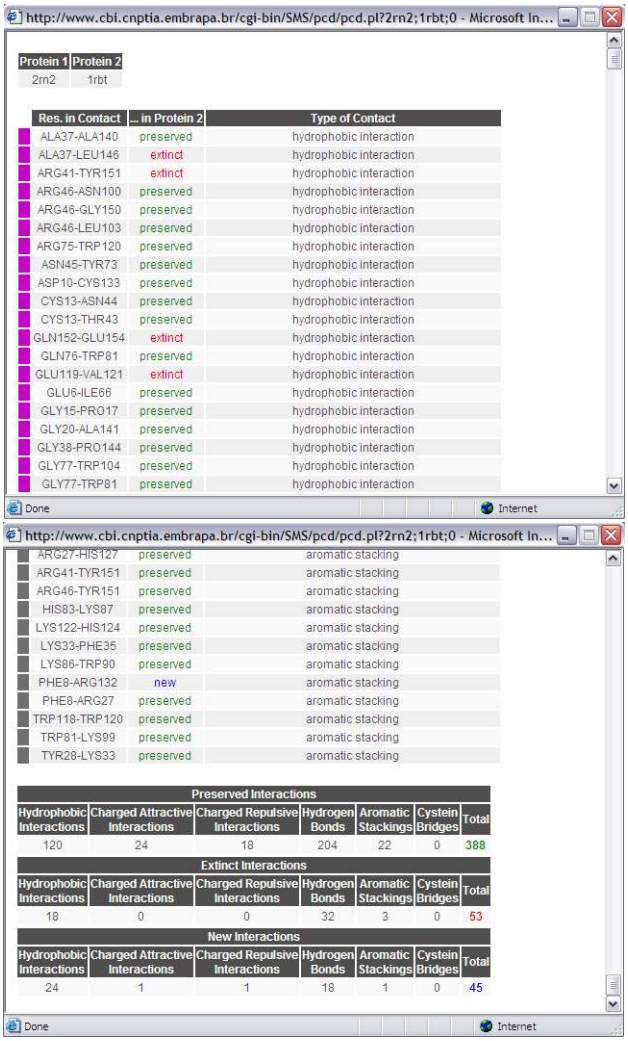


Figura 4.15: Relatório da diferença de contatos entre duas cadeias do módulo PCD do STING.

4.4.3 Topologs ASTRAL 40

É um banco de dados de cadeias PDBs homólogas com base nas interações intra-cadeia. Para todo o ASTRAL 40, computamos uma lista das cadeias mais parecidas com base em seus mapas de contatos. O banco pode ser consultado por cadeia específica, mas também possibilitamos a navegação pela lista de todas as 4.911 cadeias representativas do PDB (Figura 4.20).

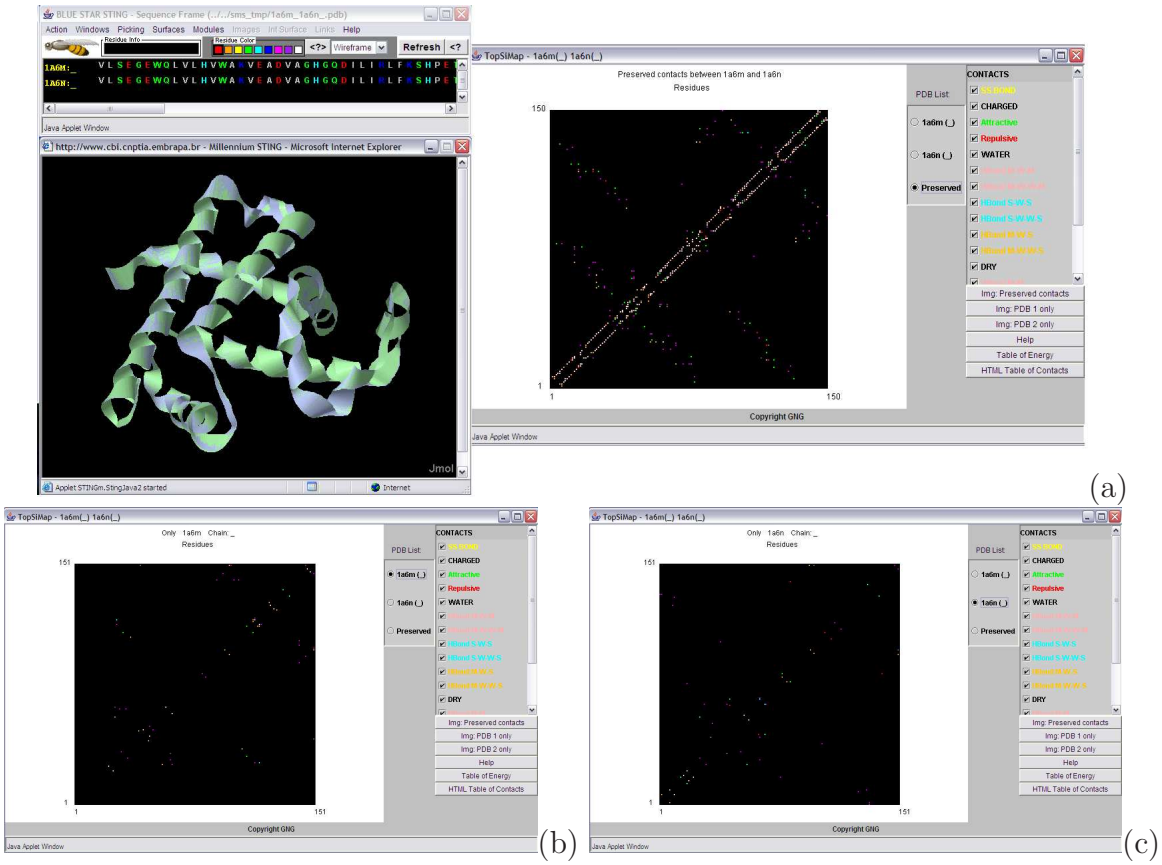


Figura 4.16: Interface do módulo TopSiMap do STING. **(a)** Telas de alinhamento de seqüência e de estruturas e mapa de contatos preservados nas duas cadeias comparadas. **(b)** Contatos presentes apenas na primeira cadeia. **(c)** Contatos presentes apenas na segunda cadeia.

4.5 Sistema de comparação de mapas de contatos disponível na internet

Projetamos e implementamos um banco de dados relacional utilizando o MySQL para armazenar todos os resultados dos experimentos. Além disto, para facilitar e publicar os resultados deste projeto, modelamos e implementamos com o uso de jsp um *web site* (bioinfo.speed.dcc.ufmg.br/3dbio/raquelcm) com os resultados dos experimentos apresentados nesta tese.

Neste site, atualmente é possível visualizar os resultados das bases utilizadas nesta tese, mas pretendemos englobar todo o PDB. O usuário pode, depois de selecionar uma das bases de dados, buscar por proteínas de mapas de contatos semelhantes à uma cadeia de consulta. É possível visualizar as estruturas e comparar os mapas de contatos.

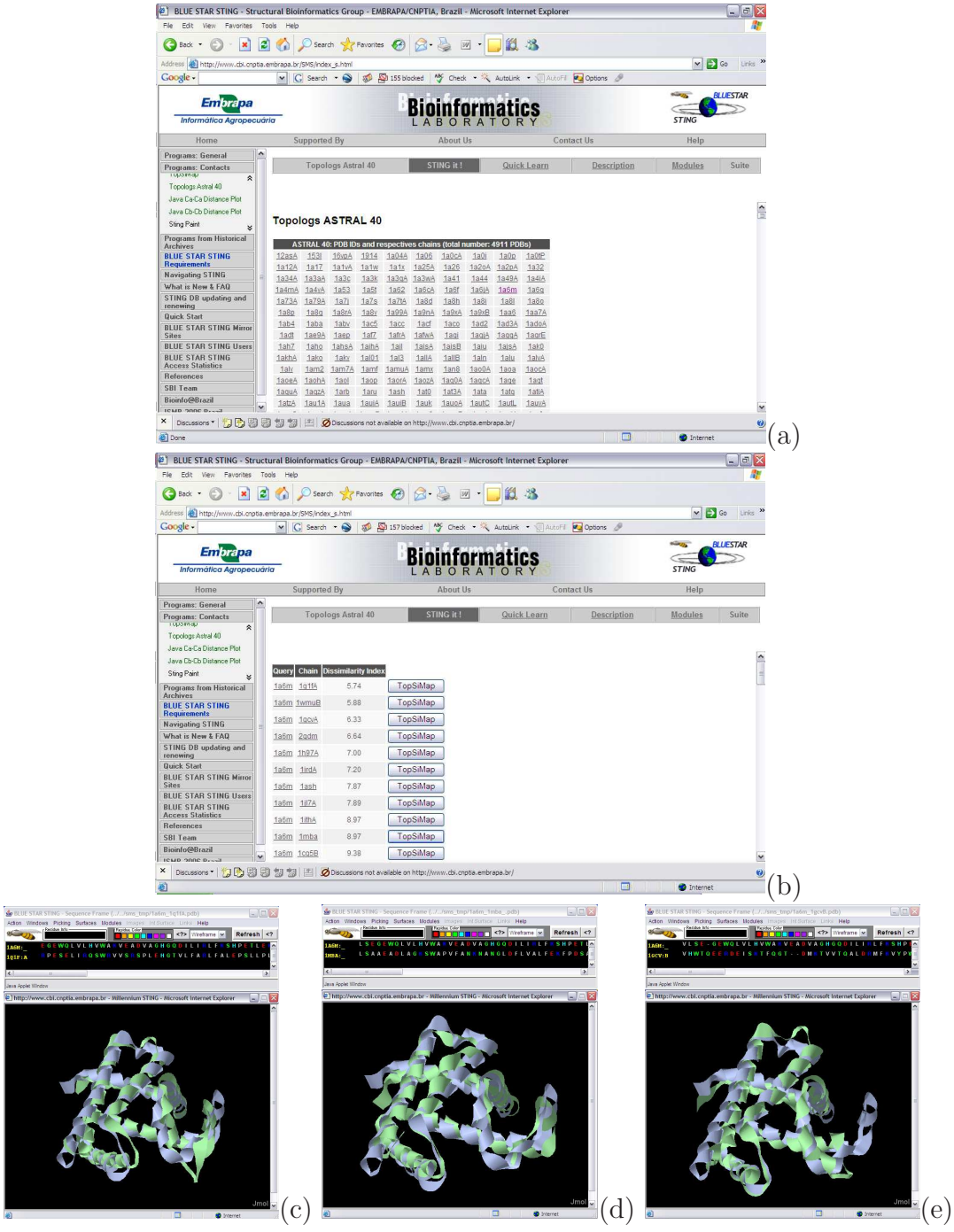


Figura 4.17: Banco de dados Topologs do STING. *(a)* Tela de ids PDB de cerca de 4.000 cadeias do ASTRAL 40. **(b)** Lista de homólogos da cadeia com base nos contatos com *links* para análise comparativa das seqüências, estruturas e mapas de contatos. São exibidas as 100 cadeias mais parecidas dentre as cerca de 4.000 da base. **(c), (d) e (e)** Primeira, décima e vigésima estruturas mais parecidas com a *mioglobina* usada no exemplo.

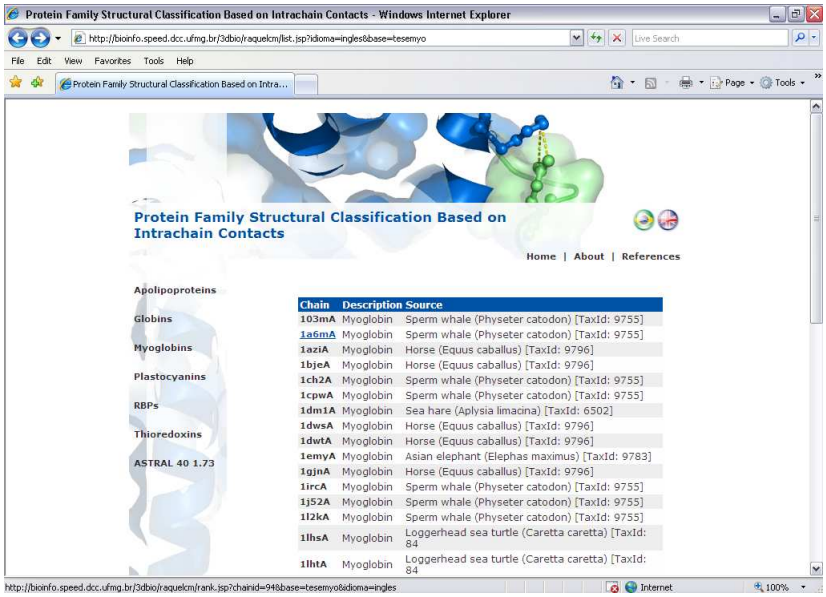


Figura 4.18: *Web site* com os resultados deste trabalho. Tela de visualização de base de dados.

Nesta tela, os usuários podem visualizar as cadeias de proteínas de cada uma das cinco famílias que fizeram parte dos nossos experimentos. Cada cadeia possui um *link* que leva ao *rank* de todas as proteínas da base ordenadas pela dissimilaridade entre os seus mapas de contatos.

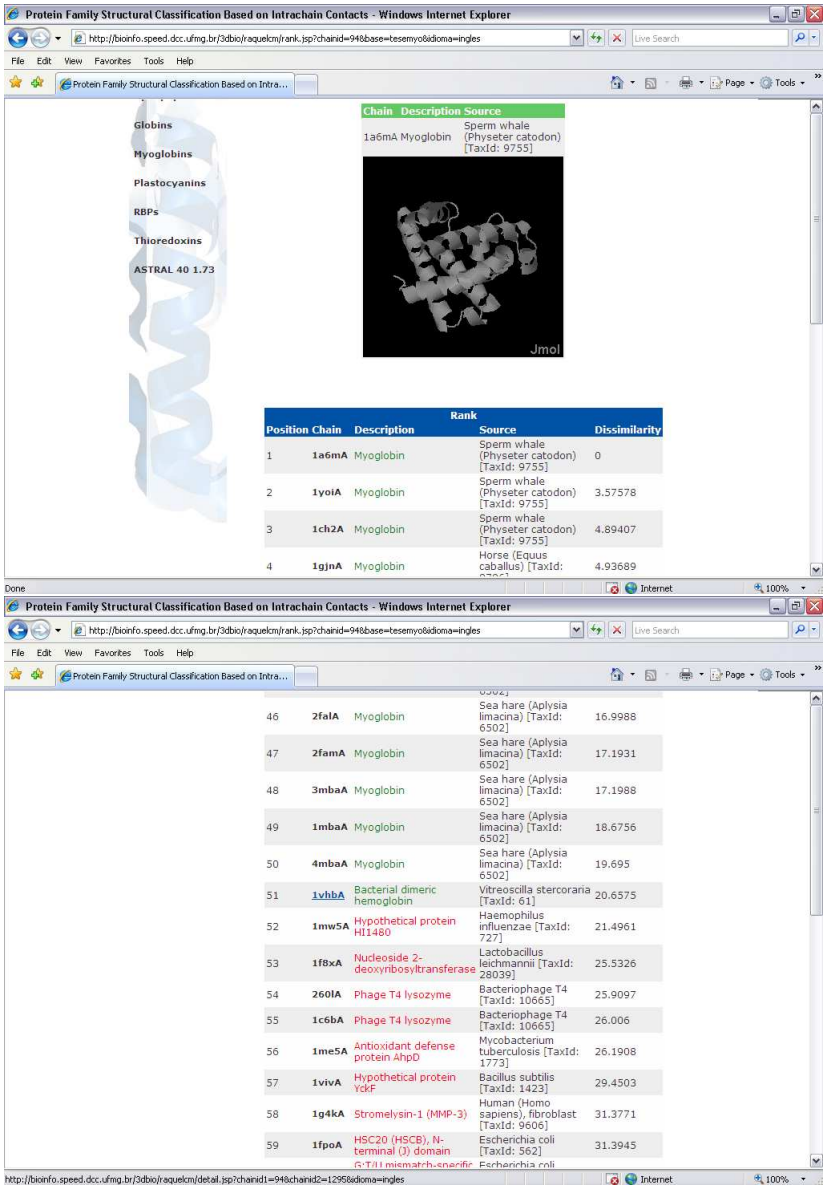


Figura 4.19: *Web site* com os resultados deste trabalho. Tela de visualização de *rank* de cadeias ordenadas por similaridade em relação à uma cadeia consultada.

Uma vez selecionada a cadeia da base de dados, o usuário pode visualizar nesta tela o *rank* de todas as proteínas da base experimental ordenadas pela dissimilaridade entre os seus mapas de contatos. Nesta tela, cada cadeia possui um *link* que leva a visualização da cadeia da consulta e a cadeia selecionada do *rank*. É possível ver os detalhes sobre cada cadeia, visualizar e interagir com as estruturas, além das figuras dos mapas de contatos.

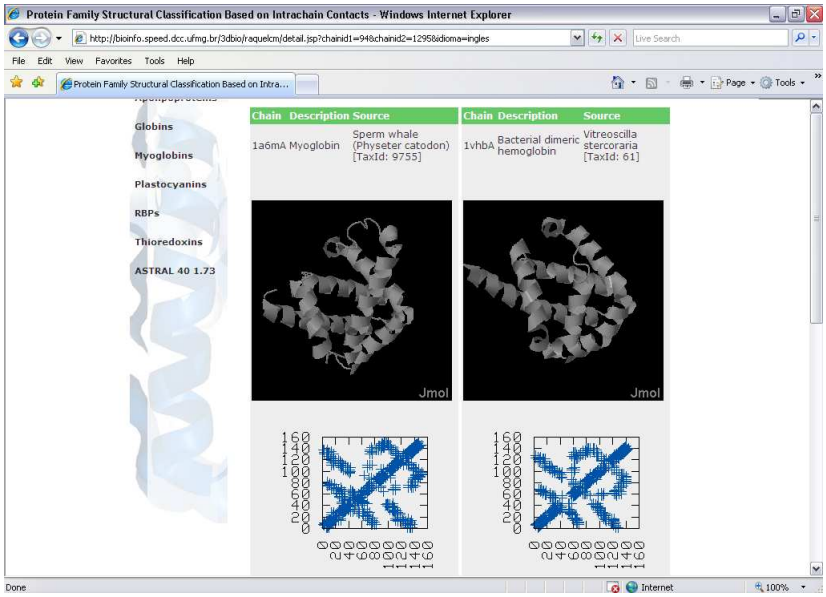


Figura 4.20: *Web site* com os resultados deste trabalho. Tela de visualização dos detalhes e comparação entre cadeia da consulta e cadeia do *rank*.

Uma vez feita uma consulta e tendo-se selecionado uma cadeia do *rank*, o usuário pode visualizar nesta tela a cadeia da consulta e a cadeia selecionada. É possível ver os detalhes sobre a cadeia, visualizar e interagir com a estrutura através de um *plug-in* do software JMol mais a Máquina Virtual Java, além dos mapas de contatos.

Capítulo 5

Conclusões

Neste trabalho, modelamos o problema de comparar estruturalmente duas cadeias proteicas como o problema de comparação entre seus mapas de contatos.

Inicialmente, propusemos uma metodologia de comparação estrutural de proteínas baseada em técnicas de processamento digital de imagens. Propusemos uma métrica baseada no paradigma de recuperação de imagens com base no conteúdo, usando como caraterística principal da imagem a distribuição de contatos (modelados como cores de acordo com a natureza da interação química) no espaço. Comparamos esta abordagem com outras métricas baseadas no registro de imagens. A primeira delas foi denominada raio médio de dispersão, por computar a média dos custos de se deslocar os contatos de um mapa para ser transformado em outro. A outra foi baseada no *earth mover's distance* e foi resolvida com base no famoso problema do transporte. Todas estas métricas propostas mostraram excelentes resultados na recuperação de proteínas de 5 famílias testadas (Globinas, Apolipoproteínas, Plastocianinas, R.B.P.s e Tioredoxinas) misturadas a proteínas de topologias diversas.

Com isto, mostramos ser os mapas de contatos bastante conservados em cada família de proteínas o que serve de indício de que o padrão de contatos em uma cadeia proteica deve ser um importante componente da assinatura estrutural de cada família.

Propusemos então uma metodologia baseada em algoritmos de agrupamento com base na densidade dos pontos (DBSCAN) para obter automaticamente os grupos de contatos de cada mapa e caracterizar cada grupo como um vetor. Posteriormente, utilizamos um modelo de otimização para casar os vetores de dois mapas de contatos e contabilizar a dissimilaridade entre eles. Mostramos que, utilizando os contatos hidrofóbicos e pontes de hidrogênio (tipos de contatos mais freqüêntes e os únicos que formam *clusters* nos mapas), fomos capazes de definir um padrão de vetores representativos da família Globina. Mostramos, inclusive que este padrão pode ser usado para recuperação de Globinas misturadas a proteínas de enovelamentos diversos com alta

precisão.

Finalmente, construímos e disponibilizamos uma ferramenta na internet que possibilita a consulta a várias bases de cadeias de proteínas e a visualização de comparação de estruturas de proteínas e seus mapas de contatos.

Como um trabalho a parte, mostramos o potencial dos algoritmos desenvolvidos na identificação de padrões de contatos entre interfaces de cadeias de complexos de proteínas. Mostramos que o algoritmo foi capaz de identificar diferentes padrões de interações entre diversas sub-famílias de Serino-Proteases (Tripsinas, Quimotripsinas, Trombinas, Matriptases e Kalikreínas) e seu inibidor BPTI.

5.1 Perspectivas

Nesta seção levantamos algumas questões sobre o futuro dos trabalhos desenvolvidos nesta tese. Primeiramente, discutimos itens que gostaríamos de ter implementado e não foi possível principalmente por questões de tempo. A seguir, serão apresentados possíveis rumos para o trabalho.

A primeira questão relaciona-se com a calibração de dois dos nossos classificadores. Utilizamos a base de Mioglobinas misturadas a outras proteínas de enovelamentos variados no processo de calibração, ou seja, definição dos valores de parâmetros que maximiza a precisão dos classificadores. Conforme explicado na Seção 4.1, utilizamos o SCOP como banco de dados padrão ouro, ou seja, ele nos fornece a classificação correta para cada cadeia proteica. Com base nesta classificação correta, calculamos a precisão dos classificadores propostos com diversas configurações de parâmetros de entrada e escolhemos o valor de parâmetro que maximiza a precião do sistema de classificação. Um possível viés na escolha deste parâmetro é que ele foi selecionado com base em apenas uma família de proteínas. Gostaríamos de repetir estes experimentos com famílias variadas e estudar a influência da família no valor ótimo deste parâmetro. O intuito de tais estudos seria o de entender melhor os parâmetros definindo se existe ou não um parâmetro único que possa ser utilizado para todas as famílias ou se existe um valor específico para cada família.

Outro item que gostaríamos de ter implementado neste trabalho é uma análise comparativa e criteriosa entre a nossa metodologia e outras propostas na literatura. O principal problema que enfrentamos foi conseguir programas de uso aberto para que pudéssemos fazer os testes com as mesmas bases de dados que apresentamos. A maioria dos autores não disponibiliza o software e apresenta os resultados em bases específicas e pré-computadas em interfaces *web*. Nesses casos, é bastante complicado conseguir dados em larga escala e de forma automática para nossa análise comparativa. Acred-

itamos que para esta análise seria necessário eleger algumas das metodologias mais interessantes e tentar conseguir os softwares dos autores ou, no pior caso, reimplementá-los.

Uma meta também muito importante e que ainda não conseguimos finalizar foi o cômputo das nossas métricas para todo o PDB. Isto não foi possível devido a restrições de recursos computacionals principalmente, apesar de nossos algoritmos não terem alta complexidade computacional e volume de dados a processar é bastante grande. O algoritmo de maior complexidade é $O(n^3)$ onde n é o números de contatos. Para uma globina de cerca de 150 resíduos, usando o valor de corte de 7Å obtemos cerca de 300 contatos hidrofóbicos. Assim a comparação entre duas globinas teria que fazer cálculos proporcionais a 300^3. Imagine como seria a comparação a nível de todo o PDB. Seriam necessárias $[k * (k-1)]/2 \approx 3.200.000.000$ comparações onde k é o número de cadeias do PDB. Mesmo uma comparação a nível de ASTRAL 40 (um subconjunto do PDB no qual não existem cadeias com mais de 40% de similaridade) seria bastante demorada. Estamos fazendo estes cálculos do intuito de disponibilizar estes resultados em nosso servidor *web*. Uma das maiores dificuldades que estamos encontrando é que existe um pequeno número de cadeias muito grandes e estas cadeias são extremamente demoradas tanto de se calcular os contatos quanto de serem comparadas com cada uma das outras milhares de cadeias do PDB.

Dando continuidade ao tema de estudo desta pesquisa, gostaríamos de nos aprofundar na elucidação de assinaturas estruturais com base em contatos preservados. Neste trabalho, provamos ser possível classificar famílias de proteínas com base apenas na localização espacial dos contatos. Mostramos ainda que existem agrupamentos de contatos conservados na família de globinas e que devem ser uma componente importante de sua assinatura estrutural, ou seja, são um conjunto de características responsáveis pela estrutura e função da família. Gostaríamos de definir os contatos preservados de forma mais precisa identificando os contatos que se preservam ou os contatos que, mesmo não preservados, sejam equivalentes em proteínas de mesma estrutura e seqüências diversas. Estamos iniciando nossos trabalhos nesta área através da modelagem de proteínas como grafos e de algoritmos de isomorfismo de subgrafos.

Apêndice A

Seqüências das Proteínas Usadas nos Experimentos

A.1 Globinas

```
1FAW_B ____________________VHWSAEEKQLITGLWGKVN_VADCGA_ 25
1HBR_B ____________________VHWTAEEKQLITGLWGKVN_VAECGA_ 25
1WMU_B ____________________VHWTSEEKQYITSLWAKVN_VGEVGG_ 25
1A9W_E ____________________VHFTAEEKAAVTSLWSKMN_VEEAGG_ 25
1IRD_B ____________________VHLTPEEKSAVTALWGKVN_VDEVGG_ 25
2PGH_B ____________________VHLSAEEKEAVLGLWGKVN_VDEVGG_ 25
1G08_B ____________________MLTAEEKAAVTAFWGKVK_VDEVGG_ 24
1JEB_B ____________________VHLTDAEKAAVSGLWGKVN_ADEVGG_ 25
1S5X_B ____________________VEWTDKERSIISDIFSHMD_YDDIGP_ 25
1XQ5_B ____________________VVWTDFERATIADIFSKLD_YEAVGG_ 25
1SPG_B ____________________VDWTDAERAAIKALWGKID_VGEIGP_ 25
1GCV_B ____________________VHWTQEERDEISKTFQGTD_MKTVVT_ 25
1CG5_B ____________________VKLSEDQEHYIKGVWKDVD_HKQITA_ 25
1CG5_A ____________________VLSSQNKKAIEELGNLIKANAEAWGA_ 26
1GCV_A ____________________AFTACEKQTIGKIAQVLAKSPEAYGA_ 26
1G08_A ____________________VLSAADKGNVKAAWGKVGGHAAEYGA_ 26
1IRD_A ____________________VLSPADKTNVKAAWGKVGAHAGEYGA_ 26
1FAW_A ____________________VLSAADKTNVKGVFSKIGGHAEEYGA_ 26
1JEB_A ____________________SLTKTERTIIVSMWAKISTQADTIGT_ 26
1HBR_A ____________________MLTAEDKKLIQQAWEKAASHQEEFGA_ 26
1WMU_A ____________________MLTEDDKQLIQHVWEKVLEHQEDFGA_ 26
```

```
1S5X_A ________________________SLSDKDKAAVRALWSKIGKSADAIGN_ 26
1XQ5_A ________________________SLSSKDKDTVKALWGKIADKAEEIGS_ 26
1MWC_A ________________________GLSDGEWQLVLNVWGKVEADVAGHGQ_ 26
2MM1_A ________________________GLSDGEWQLVLNVWGKVEADIPGHGQ_ 26
1GJN_A ________________________GLSDGEWQQVLNVWGKVEADIAGHGQ_ 26
1EMY_A ________________________GLSDGEWELVLKTWGKVEADIPGHGE_ 26
1BZ6_A ________________________VLSEGEWQLVLHVWAKVEADVAGHGQ_ 26
1LHT_A ________________________GLSDDEWNHVLGIWAKVEPDLSAHGQ_ 26
1MYT_A ____________________________ADFDAVLKCWGPVEADYTTMGG_ 22
1OJ6_A ________________________MERPEPELIRQSWRAVSRSPLEHGT_ 25
1Q1F_A ________________________MERPESELIRQSWRVVSRSPLEHGT_ 25
1HBG_A ________________________GLSAAQRQVIAATWKDIAGADNGAGVGK 28
1JL7_A ________________________GLSAAQRQVVASTWKDIAGADNGAGVGK 28
3SDH_A ___________________PSVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGV_ 34
5HBI_A ___________________PSVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGV_ 34
1DLW_A ____________________________SLFEQLGG_QAAVQAVT 16
1UVY_A ____________________________SLFEQLGG_QAAVQAVT 16
1DLY_A MMRTVQLRTLRPCIRAQQQPVRPSTSATAAAATAPAPARKCPSSLFAKLGG_REAVEAAV 59
1IDR_A MGLLSRLR________________________KREPISIYDKIGG_HEAIEVVV 29
1RTE_A MGLLSRLR________________________KREPISIYDKIGG_HEAIEVVV 29
1MOH_A ________________________SLEAAQKSNVTSSWAKASAAWGTAGP_ 26
1MBA_A ________________________SLSAAEADLAGKSWAPVFANKNANGL_ 26
1IT2_A ___________________PIIDQGPLPTLTDGDKKAINKIWPKIYKEYEQYSL_ 35
1ITH_A ________________________GLTAAQIKAIQDHWFLNIKGCLQAAAD_ 27
2GDM_A ________________________GALTESQAALVKSSWEEFNANIPKHTH_ 27
1KR7_A ____________________________MVNWAAVVD_______ 9
1UX8_A ____________________________MGQSFNAPYEAIG_EELLSQLV 21
1H97_A ________________________TLTKHEQDILLKELGPHVDTPAHIVETGL 29
1ASH_A ________________________ANKTRELCMKSLEHAKVDTSNEARQDGI 28

1FAW_B EALARLLIVYPWTQRFFSSFG_NLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDN___ 80
1HBR_B EALARLLIVYPWTQRFFASFG_NLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDN___ 80
1WMU_B EALARLLIVYPWTQRFFASFG_NLSSANAILHNAKVLAHGQKVLTSFGEAVKNLDN___ 80
1A9W_E EALGRLLVVYPWTQRFFDSFG_NLSSPSAILGNPKVKAHGKKVLTSFGDAIKNMDN___ 80
1IRD_B EALGRLLVVYPWTQRFFESFG_DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN___ 80
2PGH_B EALGRLLVVYPWTQRFFESFG_DLSNADAVMGNPKVKAHGKKVLQSFSDGLKHLDN___ 80
1G08_B EALGRLLVVYPWTQRFFESFG_DLSTADAVMNNPKVKAHGKKVLDSFSNGMKHLDD___ 79
1JEB_B EALGRLLVVYPWTQRYFDSFG_DLSSASAIMGNAKVKAHGKKVITAFNDGLNHLDS___ 80
```

```
1S5X_B KALSRCLIVYPWTQRHFSGFG_NLYNAEAIIGNANVAAHGIKVLHGLDRGVKNMDN___ 80
1XQ5_B ATLARCLIVYPWTQRYFGNFG_NLYNAAAIMGNPMIAKHGTTILHGLDRAVKNMDN___ 80
1SPG_B QALSRLLIVYPWTQRHFKGFG_NISTNAAILGNAKVAEHGKTVMGGLDRAVQNMDN___ 80
1GCV_B QALDRMFKVYPWTNRYFQKRT_DFRSS_________IHAGIVVGALQDAVKHMDD___ 70
1CG5_B KALERVFVVYPWTTRLFSKLQ_GLFSANDIG___VQQHADKVQRALGEAIDDLKK___ 76
1CG5_A DALARLFELHPQTKTYFSKFS_GFEACNE____QVKKHGKRVMNALADATHHLDN___ 76
1GCV_A ECLARLFVTHPGSKSYF_EYK_DYSAAGA____KVQVHGGKVIRAVVKAAEHVDD___ 75
1G08_A EALERMFLSFPTTKTYFPHF_DLSHGSA____QVKGHGAKVAAALTKAVEHLDD___ 75
1IRD_A EALERMFLSFPTTKTYFPHF_DLSHGSA____QVKGHGKKVADALTNAVAHVDD___ 75
1FAW_A ETLERMFTAYPQTKTYFPHF_DLQHGSA____QIKAHGKKVAAALVEAVNHIDD___ 75
1JEB_A ETLERLFLSHPQTKTYFPHF_DLHPGSA____QLRAHGSKVVAAVGDAVKSIDD___ 75
1HBR_A EALTRMFTTYPQTKTYFPHF_DLSPGSD____QVRGHGKKVLGALGNAVKNVDN___ 75
1WMU_A EALERMFIVYPSTKTYFPHF_DLHHDSE____QIRHHGKKVVGALGDAVKHIDN___ 75
1S5X_A DALSRMIVVYPQTKTYFSHWP_DVTPGSP____HIKAHGKKVMGGIALAVSKIDD___ 76
1XQ5_A DALSRMLAVYPQTKTYFSHWK_DLSPGSA____PVNKHGKTIMGGIVDAVASIDD___ 76
1MWC_A EVLIRLFKGHPETLEKFDKFK_HLKSEDEMKASEDLKKHGNTVLTALGGILKKKGH___ 81
2MM1_A EVLIRLFKGHPETLEKFDRFK_HLKSEDEMKASEDLKKHGATVLTALGGILKKKGH___ 81
1GJN_A EVLIRLFTGHPETLEKFDKFK_HLKTEAEMKASEDLKKHGTVVLTALGGILKKKGH___ 81
1EMY_A TVFVRLFTGHPETLEKFDKFK_HLKTEGEMKASEDLKKQGVTVLTALGGILKKKGH___ 81
1BZ6_A DILIRLFKSHPETLEKFDRFK_HLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH___ 81
1LHT_A EVIIRLFQLHPETQERFAKFK_NLTTIDALKSSEEVKKHGTTVLTALGRILKQKNN___ 81
1MYT_A LVLTRLFKEHPETQKLFPKFA_GIA_QADIAGNAAISAHGATVLKKLGELLKAKGS___ 76
1OJ6_A VLFARLFALEPDLLPLFQYNGRQFSSPEDSLSSPEFLDHIRKVMLVIDAAVTNVEDL_S 83
1Q1F_A VLFARLFALEPSLLPLFQYNGRQFSSPEDSLSSPEFLDHIRKVMLVIDAAVTNVEDL_S 83
1HBG_A KCLIKFLSAHPQMAAVFGFSG___ASDPGVAALGAK____VLAQIGVAVSHLGDE_G 77
1JL7_A ECLSKFISAHPEMAAVFGFSG___ASDPGVAELGAK____VLAQIGVAVSHLGDE_G 77
3SDH_A ALMTTLFADNQETIGYFKRLG___NVSQGMANDKLRGHSITLMYALQNFIDQLDNP_D 88
5HBI_A ALMTTLFADNQETIGYFKRLG___DVSQGMANDKLRGHSIILMYALQNFIDQLDNP_D 88
1DLW_A AQFYANIQADATVATFFNGID__________MPNQTNKTAAFLCAALGGPNA___ 58
1UVY_A AQFYANIQADATVATFFNGID__________MPNQTNKTAAFLCAALGGPNA___ 58
1DLY_A DKFYNKIVADPTVSTYFSNTD__________MKVQRSKQFAFLAYALGGASE___ 101
1IDR_A EDFYVRVLADDQLSAFFSGTN__________MSRLKGKQVEFFAAALGGPEP___ 71
1RTE_A EDFYVRVLADDQLSAFFSGTN__________MSRLKGKQVEFFAAALGGPEP___ 71
1MOH_A EFFMALFDAHDDVFAKFSGLF_SGAAKGTVKNTPEMAAQAQSFKGLVSNWVDNLDNA_G 83
1MBA_A DFLVALFEKFPDSANFFADFK_GKSVADIKASPKLRDVSSRIFTRLNEFVNNAANA_G 82
1IT2_A NILLRFLKCFPQAQASFPKFS__TKKSNLEQDPEVKHQAVVIFNKVNEIINSMDNQ_E 90
1ITH_A SIFFKYLTAYPGDLAFFHKFS_SVPLYGLRSNPAYKAQTLTVINYLDKVVDALGG___ 81
2GDM_A RFFILVLEIAPAAKDLFSFLK__GTSEVPQNNPELQAHAGKVFKLVYEAAIQLEVTGVV 84
```

```
1KR7_A DFYQELFKAHPEYQNKFGFKG__VALGSLKGNAAYKTQAGKTVDYINAAIGGSAD___ 62
1UX8_A DTFYERVASHPLLKPIFPSDL__________TETARKQKQFLTQYLGGPPLYT_ 64
1H97_A GAYHALFTAHPQYISHFSRLE_GHTIENVMQSEGIKHYARTLTEAIVHMLKEISN_DA 85
1ASH_A DLYKHMFENYPPLRKYFKSRE_EYTAEDVQNDPFFAKQGQKILLACHVLCATYDDR_E 84

1FAW_B _IKNTFAQLSELHC_DKLHVDPENFRLLGDILIIVLAAHFA_KEFTPECQAAWQKLVRV 136
1HBR_B _IKNTFSQLSELHC_DKLHVDPENFRLLGDILIIVLAAHFS_KDFTPECQAAWQKLVRV 136
1WMU_B _IKKTFAQLSELHC_EKLHVDPENFKLLGNILIIVLATHFP_KEFTPASQAAWTKLVNA 136
1A9W_E _LKPAFAKLSELHC_DKLHVDPENFKLLGNVMVIILATHFG_KEFTPEVQAAWQKLVSA 136
1IRD_B _LKGTFATLSELHC_DKLHVDPENFRLLGNVLVCVLAHHFG_KEFTPPVQAAYQKVVAG 136
2PGH_B _LKGTFAKLSELHC_DQLHVDPENFRLLGNVIVVVLARRLG_HDFNPDVQAAFQKVVAG 136
1G08_B _LKGTFAALSELHC_DKLHVDPENFKLLGNVLVVVLARNFG_KEFTPVLQADFQKVVAG 135
1JEB_B _LKGTFASLSELHC_DKLHVDPENFRLLGNMIVIVLGHHLG_KDFTPAAQAAFQKVVAG 136
1S5X_B _IAATYADLSTLHS_EKLHVDPDNFKLLSDCITIVLAAKMG_HAFTAETQGAFQKFLAV 136
1XQ5_B _IKATYAELSVLHS_EKLHVDPDNFKLLSDCLTIVVAAQLG_KAFSGEVQAAFQKFLSV 136
1SPG_B _IKNVYKQLSIKHS_EKIHVDPDNFRLLGEIITMCVGAKFGPSAFTPEIHEAWQKFLAV 137
1GCV_B _VKTLFKDLSKKHA_DDLHVDPGSFHLLTDCIIVELAYLRK_DCFTPHIQGIWDKFFEV 126
1CG5_B _VEINFQNLSGKH__QEIGVDTQNFKLLGQTFMVELALHYK_KTFRPKEHAAAYKFFRL 131
1CG5_A _LHLHLEDLARKHG_ENLLVDPHNFHLFADCIVVTLAVNL_QAFTPVTHCAVDKFLEL 131
1GCV_A _LHSHLETLALTHG_KKLLVDPQNFPMLSECIIVTLATHL_TEFSPDTHCAVDKLLSA 130
1G08_A _LPGALSELSDLHA_HKLRVDPVNFKLLSHSLLVTLASHLP_SDFTPAVHASLDKFLAN 131
1IRD_A _MPNALSALSDLHA_HKLRVDPVNFKLLSHCLLVTLAAHLP_AEFTPAVHASLDKFLAS 131
1FAW_A _IAGALSKLSDLHA_QKLRVDPVNFKFLGHCFLVVVAIHHP_SALTPEVHASLDKFLCA 131
1JEB_A _IGGALSKLSELHA_YILRVDPVNFKLLSHCLLVTLAARFP_ADFTAEAHAAWDKFLSV 131
1HBR_A _LSQAMAELSNLHA_YNLRVDPVNFKLLSQCIQVVLAVHMG_KDYTPEVHAAFDKFLSA 131
1WMU_A _LSATLSELSNLHA_YNLRVDPVNFKLLSHCFQVVLGAHLG_REYTPQVQVAYDKFLAA 131
1S5X_A _LKTGLMELSEQHA_YKLRVDPANFKILNHCILVVISTMFP_KEFTPEAHVSLDKFLSG 132
1XQ5_A _LNAGLLALSELHA_FTLRVDPANFKILSHCILVLLAVKFP_KDFTPEVHISYDKFFSA 132
1MWC_A _HEAELTPLAQSHA_TKHKIPVKYLEFISEAIIQVLQSKHP_GDFGADAQGAMSKALEL 137
2MM1_A _HEAEIKPLAQSHA_TKHKIPVKYLEFISEAIIQVLQSKHP_GDFGADAQGAMNKALEL 137
1GJN_A _HEAELKPLAQSHA_TKHKIPIKYLEFISDAIIHVLHSKHP_GDFGADAQGAMTKALEL 137
1EMY_A _HEAEIQPLAQSHA_TKHKIPIKYLEFISDAIIHVLQSKHP_AEFGADAQGAMKKALEL 137
1BZ6_A _HEAELKPLAQSHA_TKHKIPIKYLEFISEAIIHVLHSRHP_GDFGADAQGAMNKALEL 137
1LHT_A _HEQELKPLAESHA_TKHKIPVKYLEFICEIIVKVIAEKHP_SDFGADSQAAMKKALEL 137
1MYT_A _HAAILKPLANSHA_TKHKIPINNFKLISEVLVKVMHEKAG__LDAGGQTALRNVMGI 130
1OJ6_A SLEEYLASLGRKHR__AVGVKLSSFSTVGESLLYMLEKSLG_PAFTPATRAAWSQLYGA 139
1Q1F_A SLEEYLTSLGRKHR__AVGVRLSSFSTVGESLLYMLEKSLG_PDFTPATRTAWSRLYGA 139
1HBG_A KMVAQMKAVGVRHKGYGNKHIKAQYFEPLGASLLSAMEHRIG_GKMNAAAKDAWAAAYAD 136
```

```
1JL7_A KMVAEMKAVGVRHKGYGNKHIKAEYFEPLGASLLSAMEHRIG_GKMNAAAKDAWAAAYGD 136
3SDH_A DLVCVVEKFAVNHI__TRKISAAEFGKINGPIKKVLASKN___FGDKYANAWAKLVAV 141
5HBI_A DLVCVVEKFAVNHI__TRKISAAEFGKINGPIKKVLASKN___FGDKYANAWAKLVAV 141
1DLW_A WTGRNLKEVHANMG_____VSNAQFTTVIGHLRSALTGAGVAAALVEQTVAVAETVRGD 112
1UVY_A WTGRNLKEVHANMG_____VSNAQFTTVIGHLRSALTGAGVAAALVEQTVAVAETVRGD 112
1DLY_A WKGKDMRTAHKDLVP___HLSDVHFQAVARHLSDTLTELGVPPEDITDAMAVVASTRTE 157
1IDR_A YTGAPMKQVHQGRG_____ITMHHFSLVAGHLADALTAAGVPSETITEILGVIAPLAVD 125
1RTE_A YTGAPMKQVHQGRG_____ITMHHFSLVAGHLADALTAAGVPSETITEILGVIAPLAVD 125
1MOH_A ALEGQCKTFAANHK__ARGISAGQLEAAFKVLSGFMKSYGG_____DEGAWTAVAGA 133
1MBA_A KMSAMLSQFAKEHVG__FGVGSAQFENVRSMFPGFVASVAA__PPAGADAAWTKLFGL 136
1IT2_A EIIKSLKDLSQKHK_TVFKVDSIWFKELSSIFVSTIDGGAE________FEKLFSI 137
1ITH_A NAGALMKAKVPSHD__AMGITPKHFGQLLKLVGGVFQEEFS__ADPTTVAAWGDAAGV 135
2GDM_A VTDATLKNLGSVHVS__KGVADAHFPVVKEAILKTIKEVVG_AKWSEELNSAWTIAYDE 140
1KR7_A ____AAGLASRHK__GRNVGSAEFHNAKACLAKACSAHGA____PDLGHAIDDILSH 109
1UX8_A _EEHGHPMLRARHLP__FPITNERADAWLSCMKDAMDHVGLEGEIREFLFGRLELTARH 120
1H97_A EVKKIAAQYGKDHT__SRKVTKDEFMSGEPIFTKYFQNLVK___DAEGKAAVEKFLKH 138
1ASH_A TFNAYTRELLDRHAR_DHVHMPPEVWTDFWKLFEEYLGKKTT__LDEPTKQAWHEIGRE 140

1FAW_B VAHALARKYH____ 146
1HBR_B VAHALARKYH____ 146
1WMU_B VAHALALGYH____ 146
1A9W_E VAIALAHKYH____ 146
1IRD_B VANALAHKYH____ 146
2PGH_B VANALAHKYH____ 146
1G08_B VANALAHRYH____ 145
1JEB_B VAAALAHKYH____ 146
1S5X_B VVSALGKQYH____ 146
1XQ5_B VVSALGKQYH____ 146
1SPG_B VVSALGRQYH____ 147
1GCV_B VIDAISKQYH____ 136
1CG5_B VAEALSSNYH____ 141
1CG5_A VAYELSSCYR____ 141
1GCV_A ICQELSSRYR____ 140
1G08_A VSTVLTSKYR____ 141
1IRD_A VSTVLTSKYR____ 141
1FAW_A VGTVLTAKYR____ 141
1JEB_A VSSVLTEKYR____ 141
1HBR_A VSAVLAEKYR____ 141
```

```
1WMU_A VSAVLAEKYR____ 141
1S5X_A VALALAERYR____ 142
1XQ5_A LARALAEKYR____ 142
1MWC_A FRNDMAAKYKELGFQG 153
2MM1_A FRKDMASNYKELGFQG 153
1GJN_A FRNDIAAKYKELGFQG 153
1EMY_A FRNDIAAKYKELGFQG 153
1BZ6_A FRKDIAAKYKELGYQG 153
1LHT_A FRNDMASKYKEFGFQG 153
1MYT_A IIADLEANYKELGFSG 146
1OJ6_A VVQAMSRGWDGE___ 151
1Q1F_A VVQAMSRGWDGE___ 151
1HBG_A ISGALISGLQS____ 147
1JL7_A ISGALISGLQS____ 147
3SDH_A VQAAL_________ 146
5HBI_A VQAAL_________ 146
1DLW_A VVTV__________ 116
1UVY_A VVTV__________ 116
1DLY_A VLNMPQQ_______ 164
1IDR_A VTSGESTTAPV____ 136
1RTE_A VTSGESTTAPV____ 136
1MOH_A LMGEIEPDM_____ 142
1MBA_A IIDALKAAGA____ 146
1IT2_A ICILLRSAY_____ 146
1ITH_A LVAAMK________ 141
2GDM_A LAIVIKKEMDDAA__ 153
1KR7_A L____________ 110
1UX8_A MVNQTEAEDRSS___ 132
1H97_A VFPMMAAEI_____ 147
1ASH_A FAKEINKHGR____ 150
```

A.2 Mioglobinas

```
103M_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
2MGF_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1CH2_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
```

```
1J52_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1CPW_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1MLL_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1MLN_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1A6M_A _VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 59
1SPE_A _VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 59
1L2K_A _VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 59
1YOI_A _VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 59
1UFP_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1UFJ_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1IRC_A MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
1DWT_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1XCH_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1DWS_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1GJN_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1WLA_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1YMC_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1YMB_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1AZI_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1NZ3_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDEFKHLKTEAEMKASE 59
1NZ4_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDEFKHLKTEAEMKASE 59
1NZ5_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDEFKHLKTEAEMKASE 59
1BJE_A _GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 59
1EMY_A _GLSDGEWELVLKTWGKVEADIPGHGETVFVRLFTGHPETLEKFDKFKHLKTEGEMKASE 59
1MDN_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1MNO_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1M6C_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1MNJ_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1MNK_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1YCA_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1YCB_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1MWC_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1MWD_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1MYG_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 59
1MYI_A _GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDSFKHLKSEDEMKASE 59
2MM1_A _GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKASE 59
1MBS_A _GLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSE 59
1LHS_A _GLSDDEWNHVLGIWAKVEPDLSAHGQEVIIRLFQLHPETQERFAKFKNLTTIDALKSSE 59
```

```
1LHT_A _GLSDDEWNHVLGIWAKVEPDLSAHGQEVIIRLFQLHPETQERFAKFKNLTTIDALKSSE 59
1MYT_A ____ADFDAVLKCWGPVEADYTTMGGLVLTRLFKEHPETQKLFPKFAGIA_QADIAGNA 54
1MBA_A _SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS_VADIKASP 58
2FAL_A XSLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS_VADIKASP 59
3MBA_A _SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS_VADIKASP 58
4MBA_A _SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS_VADIKASP 58
5MBA_A _SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS_VADIKASP 58
2FAM_A XSLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS_VADIKASP 59
1DM1_A _SLSAAEADLAGKSWAPVFANKNANGDAFLVALFEKFPDSANFFADFKGKS_VADIKASP 58

103M_A DLKKAGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
2MGF_A DLKKQGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
1CH2_A DLKKHGVTVLTALGAILKKKGHHEAELKPFAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
1J52_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
1CPW_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYWEFISEAIIHVLHSRH 120
1MLL_A DLKKHGVTFLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
1MLN_A DLKKHGVTILTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
1A6M_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 119
1SPE_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 119
1L2K_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 119
1YOI_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 119
1UFP_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
1UFJ_A DLKKHGVTVLTGLGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
1IRC_A DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSGATKHKIPIKYLEFISEAIIHVLHSRH 120
1DWT_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1XCH_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYNEFISDAIIHVLHSKH 119
1DWS_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1GJN_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1WLA_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1YMC_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1YMB_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1AZI_A DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1NZ3_A DLKEHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1NZ4_A DLKEHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1NZ5_A DLKEHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1BJE_A DLKKTGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 119
1EMY_A DLKKQGVTVLTALGGILKKKGHHEAEIQPLAQSHATKHKIPIKYLEFISDAIIHVLQSKH 119
1MDN_A DLKKHGNTNLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
```

```
1MNO_A DLKKHGNTNLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1M6C_A DLKKHGNTNLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1MNJ_A DLKKVGNTILTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1MNK_A DLKKVGNTTLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1YCA_A DLKKHGNTTLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1YCB_A DLKKHGNTTLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1MWC_A DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1MWD_A DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1MYG_A DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1MYI_A DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
2MM1_A DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 119
1MBS_A DLRKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSKH 119
1LHS_A EVKKHGTTVLTALGRILKQKNNHEQELKPLAESHATKHKIPVKYLEFICEIIVKVIAEKH 119
1LHT_A EVKKHGTTVLTALGRILKQKNNHEQELKPLAESHATKHKIPVKYLEFICEIIVKVIAEKH 119
1MYT_A AISAHGATVLKKLGELLKAKGSHAAILKPLANSHATKHKIPINNFKLISEVLVKVMHEKA 114
1MBA_A KLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVAS 118
2FAL_A KLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVAS 119
3MBA_A KLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVAS 118
4MBA_A KLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVAS 118
5MBA_A KLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVAS 118
2FAM_A KLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVAS 119
1DM1_A KLRDHSSTIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVAS 118

103M_A PGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
2MGF_A PGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1CH2_A PGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1J52_A PGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1CPW_A PGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1MLL_A PGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1MLN_A PGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1A6M_A PGDFGADAQGAMNKALELFRKDIAAKYKELGY__ 151
1SPE_A PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 153
1L2K_A PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 153
1YOI_A PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 153
1UFP_A PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1UFJ_A PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1IRC_A PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
1DWT_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
```

```
1XCH_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1DWS_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1GJN_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1WLA_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1YMC_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1YMB_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1AZI_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1NZ3_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1NZ4_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1NZ5_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1BJE_A PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
1EMY_A PAEFGADAQGAMKKALELFRNDIAAKYKELGFQG 153
1MDN_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1MNO_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1M6C_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1MNJ_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1MNK_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1YCA_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1YCB_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1MWC_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1MWD_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1MYG_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
1MYI_A PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 153
2MM1_A PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 153
1MBS_A PAEFGADAQAAMKKALELFRNDIAAKYKELGFHG 153
1LHS_A PSDFGADSQAAMKKALELFRNDMASKYKEFGFQG 153
1LHT_A PSDFGADSQAAMKKALELFRNDMASKYKEFGFQG 153
1MYT_A G__LDAGGQTALRNVMGIIIADLEANYKELGFSG 146
1MBA_A VAAPPAGADAAWTKLFGLIIDALKAAGA____ 146
2FAL_A VAAPPAGADAAWTKLFGLIIDALKAAGA____ 147
3MBA_A VAAPPAGADAAWTKLFGLIIDALKAAGA____ 146
4MBA_A VAAPPAGADAAWTKLFGLIIDALKAAGA____ 146
5MBA_A VAAPPAGADAAWTKLFGLIIDALKAAGA____ 146
2FAM_A VAAPPAGADAAWTKLFGLIIDALKAAGK____ 147
1DM1_A VAAPPAGADAAWTKLFGLIIDALKAAGK____ 146
```

Apêndice B

Publicações

Referências Bibliográficas

[Anfinsen, 1973] Anfinsen, C. (1973). Studies on the principles that govern the folding of protein chains. *Les Prix Nobel en 1972*, pp. 103–119.

[Anfinsen et al., 1955] Anfinsen, C.; Harrington, W.; Hvidt, A.; Linderstrom-Lang, K.; Ottensen, M. e Schellman, J. (1955). Studies on the structural basis of ribonuclease activity. *Biochimica et Biophysica Acta*, 17:141–142.

[Anfinsen et al., 1954] Anfinsen, C.; Redfield, R.; Choate, W.; Page, J. e Carroll, W. (1954). Studies on the gross structure, cross-linkages and terminal sequences in ribonuclease. *Journal of Biological Chemistry*, 2007(1):201–210.

[Bairoch et al., 2004] Bairoch, A.; Apweiler, R.; Wu, C.; Barker, W.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M.; Natale, D.; O'Donovan, C.; Redaschi, N. e Yeh, L. (2004). The universal protein resource (uniprot). *Nucleic Acids Res.*, 233:154–159.

[Barthel et al., 2007] Barthel, D.; Hirst, J.; Blazewicz, J.; Burke, E. e Krasnogor, N. (2007). Procksi: a decision support system for protein (structure) comparison, knowledge, similarity and information. *BMC Bioinformatics*, 8(416).

[Berman et al., 2000] Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. e Bourne, P. (2000). The protein data bank. *Nucleic Acids Res.*, 28:235–242.

[Brenner et al., 1995] Brenner, S.; Chothia, C.; Hubbard, T. e Murzin, A. (1995). Understanding protein structure: using scop for fold interpretation. *Methods in Enzymology*, 266:635–643.

[Brenner et al., 2000] Brenner, S.; Koehl, P. e Levitt, M. (2000). The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res.*, 28:254–256.

[Brown, 1992] Brown, L. (1992). A survey of image registration techniques. In *ACM Computing Surveys*, volume 4, pp. 325–376.

[Caprara et al., 2004] Caprara, A.; Carr, R.; Istrail, S.; Lancia, G. e Walenz, B. (2004). 1001 optimal pdb structure alignment: integer programming methods for finding the maximum contact map overlap. *J. Comput. Biol.*, 11:27–52.

[Chandonia et al., 2004] Chandonia, J.; Hon, G.; Walker, N.; Conte, L. L.; Koehl, P.; Levitt, M. e Brenner, S. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Res.*, 32:D189–D192.

[Chandonia et al., 2002] Chandonia, J.; Walker, N.; Conte, L. L.; Koehl, P. e Brenner, M. L. S. (2002). ASTRAL compendium enhancements. *Nucleic Acids Res.*, 30:260–263.

[Chung et al., 2007] Chung, J.; Beaver, J.; Scheeff, E. e Bourne, P. (2007). Con-struct map: a comparative contact map analysis tool. *Bioinformatics*, 23(18):2491–2492.

[Cormen et al., 2001] Cormen, T.; Leiserson, C.; Rivest, R. e Stein, C. (2001). *Introduction to algorithms*. MIT Press and McGraw-Hill.

[Dantzig, 1951] Dantzig, G. (1951). *Application of the simplex method to a transportatio problem*. John Wiley and sons.

[Del-Bimbo, 1999] Del-Bimbo, A. (1999). *Visual information retrieval*. MorganKaufmann.

[Ester et al., 1996] Ester, M.; Kriegel, H.; Sander, J. e Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*.

[Fawcett, 2006] Fawcett, . (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

[Fernandes-Jr. et al., 2004] Fernandes-Jr., F.; Carceroni, R.; Lopes, C.; Meira-Jr., W.; Melo, R.; Araujo, A.; Santoro, M. e Silveira, C. (2004). An image-matching approach to protein similarity analysis. In *SIBGRAPI '04: Proceedings of the Computer Graphics and Image Processing, XVII Brazilian Symposium on (SIBGRAPI'04)*, pp. 17–24, Washington, DC, USA. IEEE Computer Society.

[Guting, 1994] Guting, R. (1994). An introduction to spatial database systems. *The International Journal of Very Large Data Bases*, 3(4):357–399.

[Holm e Sander, 1991] Holm, L. e Sander, C. (1991). Detection of common tridimensional substructures in proteins. *Proteins*, 11:51–58.

[Hough, 1962] Hough, P. (1962). Method and means for recognizing complex patterns. Technical report.

[Hu et al., 2002] Hu, J.; Shen, X.; Shao, Y.; Bystroff, C. e Zaki, M. (2002). Mining protein contact maps. In *2nd BIOKDD: Workshop on Data Mining in Bioinformatics*.

[Huang et al., 1997] Huang, J.; Kumar, S.; Mitra, M.; Zhu, W. e Zabih, R. (1997). Image indexing using color correlograms. In *Computer Vision and Pattern Recognition (CVPR'97)*, pp. 762–768.

[Kaufman e Rousseeuw, 1990] Kaufman, L. e Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons.

[Kohavi, 2004] Kohavi, F. P. R. (2004). Machine learning. *Machine Learning*, 30(2-3):127–132.

[Krasnogor e Pelta, 2004] Krasnogor, N. e Pelta, D. (2004). Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, 20:1015–1021.

[Kutulakos, 2000] Kutulakos, K. (2000). Approximate n-view stereo. In *European Conf. on Computer Vision*, pp. 67–83.

[Lancia et al., 2001] Lancia, G.; Carr, R.; Walenz, B. e Istrail, S. (2001). 101 optimal pdb substructure alignments: a branch and cut algorithm for the maximum contact map overlap problem. In *5th Annual International Conference on Computational Molecular Biology (RECOMB)*, pp. 192–202.

[Leach, 2001] Leach, A. (2001). *Molecular Modelling: Principles and Applications (2nd Edition)*. Prentice Hall.

[Levinthal, 1968] Levinthal, C. (1968). Are there pathaways for protein folding? *Journal of Chimie Physique et de Physico-Chimie Biologique*, 65:44–45.

[Lopes, 2006] Lopes, J. (2006). Ligações químicas e interações intermoleculares (apostila). In *Curso de Educação continuada, SEE-MG/CECIMIG-UFMG*.

[Maintz e VIergever, 1998] Maintz, J. e VIergever, M. (1998). A survey of medical image registration. In *Medical Image Analysis*, volume 2, pp. 1–36.

[Mancini et al., 2004] Mancini, A.; Higa, R.; Oliveira, A.; Dominiquini, F.; Kuser, P.; Yamagishi, M.; Togawa, R. e Neshich, G. (2004). STING contacts: a web-based

application for identification and analysis of amino acids contacts within protein structure and across protein interfaces. *Bioinformatics*, 20(13):2145–2147.

[Melo et al., 2008] Melo, R.; Fernandes-Jr., F.; Carceroni, R.; Lopes, C.; Murray, C.; Meira-Jr, W.; Araújo, A.; Silveira, C. e Santoro, M. (2008). Similarity-based versus feature-based analysis of structural protein similarity. *Manuscrito submetido à revista Pattern Analysis and Applications*.

[Melo et al., 2007a] Melo, R.; Gomide, J.; Dias, P.; Meira-Jr., W. e Santoro, M. (2007a). Mining structural signatures of proteins. In *III Workshop em Algoritmos e Aplicações de Minaração de Dados*.

[Melo et al., 2006] Melo, R.; Lopes, C.; Fernandes-Jr., F.; Silveira, C.; Santoro, M.; Carceroni, R.; Meira-Jr., W. e Araujo, A. (2006). A contact map matching approach to protein structure similarity analysis. *Genet. Mol. Res.*, 5(2):284–308.

[Melo et al., 2007b] Melo, R.; Ribeiro, C.; Murray, C.; Veloso, C.; Silveira, C.; Neshich, G.; Meira-Jr., W.; Carceroni, R. e Santoro, M. (2007b). Finding protein-protein interaction patterns by contact map matching. *Genet. Mol. Res.*, 6(4):946–963.

[Mojsilovic et al., 2004] Mojsilovic, A.; Gomes, J. e Rogowitz, B. (2004). Semantic-friendly indexing and quering of images based on the extraction ofthe objective semantic cues. *Int. J. Computer Vision*, 56(1-2):79–107.

[Murzin et al., 1995] Murzin, A.; Brenner, S.; Hubbard, T. e Chothia, C. (1995). Scop: A structural classification of proteins database for investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540.

[Neshich et al., 2005] Neshich, G.; Borro, L.; Higa, R.; Kuser, P.; Yamagishi, M.; Franco, E.; Krauchenco, J.; Ribeiro, R. F. A.; Bezerra, G.; Velludo, T.; Jimenez, T.; Furukawa, N.; Teshima, H.; Kitajima, K.; Bava, A.; Sarai, A.; Togawa, R. e Mancini, A. (2005). Diamond sting: an expanded functionality for the sting suite of programs allowing the comprehensive sequence/structure/function/stability analysis with added capability for handling local files. *Nucleic Acids Res. : Web Server Issue*, 33.

[Neshich et al., 2006a] Neshich, G.; Mazoni, I.; Oliveira, S.; Yamagishi, M.; Kuser-FalcÃ£o, P.; Borro, L.; Morita, D.; Souza, K.; Almeida, G.; Rodrigues, D.; Jardine, J.; Togawa, R.; Mancini, A.; Higa, R.; Cruz, S.; Vieira, F.; dos Santos, E.; Melo, R. e Santoro, M. (2006a). The star STING server: a multiplatform environment for protein structure analysis. *Genet. Mol. Res.*, 5(2).

[Neshich et al., 2006b] Neshich, G.; Mazoni, I.; Oliveira, S.; Yamagishi, M.; Kuser-FalcÃ£o, P.; Borro, L.; Morita, D.; Souza, K.; Almeida, G.; Rodrigues, D.; Jardine, J.; Togawa, R.; Mancini, A.; Higa, R.; Cruz, S.; Vieira, F.; Santos, E.; Melo, R. e Santoro, M. (2006b). The star sting server: A multiplatform environment for protein structure analysis. *Genet. Mol. Res.*, 5(4):717–722.

[Neshich et al., 2003] Neshich, G.; Togawa, R.; Mancini, A.; Kuser, P.; Yamagishi, M.; Pappas-Jr, G.; Torres, W.; e Campos, T. F.; Ferreira, L.; Luna, F.; Oliveira, A.; Miura, R.; Inoue, M.; Horita, L.; de Souza, D.; Dominiquini, F.; Alvaro, A.; Lima, C.; Ogawa, F.; Gomes, G.; Palandrani, J.; dos Santos, G.; de Freitas, E.; Mattiuz, A.; Costa, I.; de Almeida, C.; Souza, S.; Baudet, C. e Higa, R. (2003). STING millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, 31(13):3386–3392.

[Pauling e Corey, 1951] Pauling, L. e Corey, R. (1951). The plated sheet, a new layer configuration of polypeptide chains. *PNAS*, 37:251–256.

[Pauling et al., 1951] Pauling, L.; Corey, R. e H.R.Branson (1951). The structure od proteins: two hydrogen-bonded helical configurations of the polypeptide chain. In *Proc. Nat. Acad. Sci. Wash.*, volume 37, pp. 205–211.

[Pearl et al., 2003] Pearl, F.; Bennett, C.; Brazy, J.; Harrison, A.; Martin, N.; Shepherd, A.; Sillitoe, I.; Thornton, J. e Orengo, C. (2003). The cath database: as extended protein family resource for structural and functional genomics. *Nucleic Acid Res.*, 31(1):452–455.

[Pentland et al., 1994] Pentland, A.; Picard, R. e Sclaroff, S. (1994). Photobook: content-based manipulation ofimage databases. In *SPIE Storage and Retrieval for Image and Video Databases*.

[Rubner et al., 1998] Rubner, Y.; Tomasi, C. e Guibas, L. (1998). A metric for distributions with applications to image databases. In *IEEE International Conf. on Computer Vision*.

[Silveira et al., 2008] Silveira, C.; Pires, D.; Melo, R.; Ribeiro, C.; Veloso, C.; J.C.D.Lopes; Meira-Jr, W.; Neshich, G.; Ramos, C.; Habesch, R. e Santoro, M. (2008). Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Submitted to Proteins: Structure, Functiona and Bioinformatics*.

[Sobolev et al., 1999] Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. e Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15:327–332.

[Wagner, 1986] Wagner, H. (1986). *Principles of operations research with applications to managerial decisions*. Prentice-Hall.

[Wetlaufer e Ristow, 1973] Wetlaufer, D. e Ristow, S. (1973). Acquisition of three-dimensional structure of proteins. *Annual Review of Biochemistry*, 42:135–158.

[Yang e Honig, 1999] Yang, A. e Honig, B. (1999). Sequence to structure alignment in comparative modelling. *Proteins: Struc., Func. and Genet.*, 3:66–72.