VALDETE MARIA GONÇALVES DE ALMEIDA

HYDROPACE: UMA METODOLOGIA PARA ANÁLISE DE INIBIÇÃO CRUZADA EM SERINO PROTEASES ATRAVÉS DE CENTROIDES DE REGIÕES HIDROFÓBICAS

Belo Horizonte
19 de novembro de 2011

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

HYDROPACE: UMA METODOLOGIA PARA ANÁLISE DE INIBIÇÃO CRUZADA EM SERINO PROTEASES ATRAVÉS DE CENTROIDES DE REGIÕES HIDROFÓBICAS

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.

VALDETE MARIA GONÇALVES DE ALMEIDA

Belo Horizonte
19 de novembro de 2011

UNIVERSIDADE FEDERAL DE MINAS GERAIS

HydroPaCe: uma metodologia para análise de inibição cruzada em serino proteases através de centroides de regiões hidrofóbicas

VALDETE MARIA GONÇALVES DE ALMEIDA

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

Ph. D. Raquel C. de Melo Minardi – Co-orientador
Universidade Federal de Minas Gerais

Ph. D. Carlos Henrique da Silveira – Co-orientador
Universidade Federal de Itajubá

Belo Horizonte, 19 de novembro de 2011

Resumo

Interfaces de interações proteína-proteína contêm informações importantes sobre o reconhecimento molecular. Este trabalho versa sobre a descoberta de padrões conservados nesta região. Nós entendemos que é essencial compreender como substratos e inibidores são ligados ou para predizer a ligação molecular. Quando um inibidor liga-se a enzimas com estruturas, sequências ou mecanismos diferentes, nós chamamos esta falta de especificidade de *Inibição Cruzada*. A identificação de variantes é um tarefa difícil no qual os métodos tradicionais podem falhar. Para entender como a inibição cruzada ocorre, nós modelamos o problema e propomos uma metodologia, na qual chamamos de HydroPaCe (***Hydro****fobic* ***Pa****tch* ***Ce****ntroid*) e avaliamos, por meio de métrica de cobertura penalizada PRM (*Penalized Recall Metric*), para detectar padrões conservados. Os átomos hidrofóbicos que compõem a interface foram modelados como grafos de interações apolares e os *patches* ou regiões conectadas foram computadas e sumarizadas em centroides geométricos (HP-centroides), e sua conservação foi detectada. Nós analisamos dois casos de inibição cruzada, sendo um com o inibidor Ovomovoide (5 complexos não redundantes) e outro com o inibidor Eglina C (4 complexos não redundantes). Além disso, duas famílias de serino proteases, as Tipo Tripsina (35 estruturas não redundantes) e Tipo Subtilisina (9 estruturas não redundantes), com cadeias simples foram analisadas, usando técnica de projeção da interface. Os padrões encontrados nas interfaces das inibições cruzadas foram encontrados na maioria das enzimas destas famílias. Acreditamos que nossa metodologia atingiu um nível de abstração para obter propriedades invariantes em inibição cruzada. Nós mostramos exemplos onde os HP-centroides foram preditos com sucesso em enzimas por inibidores já estudados, de acordo com a base de dados BRENDA. Por fim, nós disponibilizamos todos os códigos, usados nos passos metodológicos, bem como, tutoriais e exemplos em www.dcc.ufmg.br/∼raquelcm/hydropace.

Abstract

Interfaces of protein-protein interactions contain important information about molecular recognition. This work aims at the discovery of conserved patterns in this region. We believe it is essential to understand how substrates and inhibitors are bound and for predicting molecular binding. When an inhibitor binds to enzymes with different structures, sequences or mechanisms, we call this lack of specificity of Cross-Inhibition. The identification of variants is a difficult task for which traditional methods may fail. To understand how the cross-inhibition occurs, we model the problem and propose a methodology, which we call HydroPaCe(HydrofobicPatch Centroid) and we evaluate, through penalized recall metric - PRM, to detect preserved patterns. The hydrophobic atoms that belong to the interface were modeled as graphs of apolar interactions and hydrofobic connected regions or patches were computed and summarized by geometric centroids (HP-centroids), and their conservation is detected. We analyze two cases of cross-inhibition, one with the inhibitor Ovomovoid (5 non-redundant complexes) and another with the inhibitor Eglin C (4 non-redundant complexes). In addition, two families of serine proteases with single chain (apo-enzymes) were analyzed using projection of the interface technique: The Trypsin-like (35 non-redundant structures) and the Subtilisin-like (9 non-redundant structures). The patterns found in the interfaces of cross-inhibitions were found in most enzymes these families. We believe that our methodology achieves an appropriate level of abstraction to obtain invariant properties in cross-inhibition. We show examples in which HP-centroids successfully predicted enzymes that could be inhibited by the studied inhibitors according to BRENDA database. Finally, we provide all codes used in the methodological steps, as well as tutorials and examples in www.dcc.ufmg.br/∼raquelcm/hydropace.

".....Qualquer coisa que você possa fazer ou sonhar, você pode começar. A coragem contém em si mesma, poder, o gênio e magia."

Goethe

Agradecimentos

Agradeço a Deus pela vida.

Ao meu pai Osvaldo que me encorajou no início desta jornada, com suas simples palavras de incentivo. Pai, sei que hoje você não está mais entre nós, mas acredito que sempre estará olhando por mim. A minha mãe Maria por sempre me incluir em suas orações e por tudo que não há como citar ou descrever ao longo de minha existência.

Agradeço ao Fernando, meu marido e companheiro, pelo carinho e apoio incondicional para que eu fizesse este doutorado. Fernando, você me ensinou que a tolerância e a paciência são grandes virtudes. A você eu dedico todo meu amor.

A minhas irmãs Ediane e Rosilene, pelo apoio, carinho e dedicação nos momentos difíceis. Agradeço também aos meus sobrinhos João Vitor e Luís Felipe pela alegria e energia da infância que sempre me faz feliz.

A minha sogra Ana que do seu jeitinho sempre torceu por mim. Aos cunhados Rogério, Rodrigo e Andréia e as sobrinhas Mariana, Ana Júlia, Wanessa, Jéssica e Estefani pelo apoio em todos os momentos.

Aos professores e mestres Marcelo Santoro, Carlos Henrique e Wagner Meira, que, mais que orientadores, foram amigos que ofereceram sua grande inteligência e capacidade para que eu pudesse concluir este doutorado.

A professora Raquel Minardi que foi mais que uma orientadora, foi uma amiga e conselheira, que ofereceu seu talento para me conduzir. Raquel, não tenho palavras para agradecer o quanto você contribuiu com este trabalho.

Ao grande amigo Douglas Pires que sempre me apoiou nos momentos mais difíceis, usando de sua capacidade intelectual para me ajudar nos problemas metodológicos. Douglas, para você meu irmãozinho do coração, meu muito obrigada. Conte sempre comigo.

Ao amigo Angelo Bruno e a todos os amigos que conheci ao longo desses quatro anos; os amigos da Bioinformática, os amigos biólogos, a todos vocês meu abraço apertado.

A todos os professores que estiveram presentes durante esse doutorado e que contribuíram muito com seus conhecimentos.

Aos amigos do LBS, pela troca de conhecimentos e inesquecíveis momentos de descontração que tornaram esta jornada mais suave.

A instuição de fomento CAPES pelo apóio financeiro para que eu pudesse me dedicar exclusivamente à pesquisa.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AA: Agrupamento Aglomerativo

ASA: Accessibility Solvent Area

BRENDA: BRaunschweig ENzyme Database

CCC: Centróide de Componente Conexo

DC: Dependente de Corte

DV: Diagrama de Voronoi

EBCC: Centróide de Comunidade baseado no método Edge Betweeness

EC: Enzyme Classification

HP-centroide: Hydrofobic Patch Centroide

HydroPaCe: Hydrofobic Patch Centroid

IC: Independente de Corte

IFR: Interface Forming Residue

IUBMB: International Union of Biochemistry and Molecular Biology

MO: Modelo de Otimização

PDB: Protein Data Bank

PDBest: PDB Enhanced Structure Toolkit

PRM: Penalized Recall Metric

RCL: Reative Centre Loop

SCOP: Structural Classification of Proteins

SGCC: Centróide de Comunidade baseado no método Spin Glass

SMS: STING Millennium Suite

TD: Tesselação de Delaunay

Capítulo 1

Introdução

Inibição enzimática ocorre quando uma molécula liga-se a uma enzima diminuindo a sua atividade, podendo ser reversível ou irreversível. Ela é um fenômeno complexo que envolve reconhecimento molecular e interações [Bahadur e Zacharias (2008); Shulman-Peleg et al. (2007); Caffrey et al. (2004); Tuncbag et al. (2011); Csermely (2008)]. É um mecanismo importante envolvendo regulações metabólicas inter e intra celular, processos inflamatórios e imunológicos, replicação de vírus e muitas outras funções biológicas [Rawlings et al. (2004)]. Uma vez que esse fenômeno natural é corretamente entendido, ele pode ser usado com propósitos biotecnológicos no desenvolvimento de novos fármacos, inseticidas, pesticidas, entre outros.

Os inibidores enzimáticos podem ser proteicos ou não proteicos e eles podem diminuir a habilidade de uma enzima para encontrar substratos, diminuir a atividade catalítica ou a combinação de ambos. A inibição pode ocorrer de três maneiras diferentes. A primeira é a inibição competitiva e ocorre quando o inibidor tem uma afinidade pelo sítio ativo, disputando o acesso com o substrato. A segunda é a inibição não competitiva e ocorre quando o processo de inibição não afeta a ligação no substrato, mas reduz a atividade enzimática. A terceira é a combinação dos dois primeiros tipos e ocorre quando o inibidor e o substrato ligam-se ao mesmo tempo na enzima e o primeiro afeta a ligação do último.

Um caso particular é a inibição de proteases e, neste contexto, a base de dados MEROPS é atualmente um dos mais importantes repositórios [Rawlings et al. (2008)]. Ele agrupa hierarquicamente proteases e inibidores em famílias (entidades relacionadas pela sequência) e clãs (entidades relacionadas pela estrutura). Uma criteriosa busca na base de dados MEROPS ressalta um bem conhecido, mas intrigante fenômeno: a falta de especificidade de alguns inibidores, envolvendo estruturas tridimensionais e mecanismos catalíticos diferentes. Por exemplo: os inibidores Ovomucoide e Eglina C agem em diferentes clãs de proteases, como PA(S) (Tipo Tripsina: enovelamento do tipo *all* β)[1] e SB (Tipo Subtilisina: enovelamento do tipo α/β)[2]; o inibidor *Kunitz* de Tripsina da

[1] *all* β: estrutura formada quase exclusivamente por folhas β, apresentadas em um arranjo quase ou completamente anti-paralelo

[2] α/β: estrutura formada por camadas externas de α hélices e um núcleo de folhas β paralelas.

soja decai a atividade proteolítica tanto em serino quanto em metalo proteases, que têm mecanismos enzimáticos muito diferentes. Nós chamamos esta falta de especificidade do inibidor de *Inibição Cruzada*. Nosso principal desafio é desenvolver uma metodologia que pode ajudar a entender e predizer esse fenômeno.

O reconhecimento e a ligação entre enzima e inibidor é determinado pela organização complexa das interações químicas e fatores entrópicos envolvendo o sistema protease-inibidor-solvente. Felizmente, a energia experimental da ligação de muitos complexos inibidores e proteases já foi termodinamicamente determinada. É conhecido, por exemplo, que a ligação de Ovomucoide e Elastase, em 25 °C, é caracterizada por uma energia livre de Gibbs negativa, na qual, a variação da entalpia é quase insignificante, mas a mudança de entropia é amplamente positiva [Baker e Murphy (1997)]. Isso nos deu fortes indícios de que devemos concentrar a nossa atenção, especialmente na busca por padrões de interações hidrofóbicas conservadas. Nós definimos esses padrões como regiões hidrofóbicas invariáveis ou *patches* que estão em contato com as mesmas regiões do inibidor.

Embora existam muitos estudos bioquímicos analisando a diversidade de processos de inibição [Laskowski e Qasim (2000a); Bode et al. (1986); Qasim et al. (1997); Chakrabarti e Janin (2002a)], a caracterização experimental de inibição é um processo de trabalho intensivo. A grande quantidade de possíveis inibidores para uma determinada enzima podem fazer com que estes testes sejam custosos e, por isso, métodos *in silico* podem contribuir para predizer reconhecimento enzima-inibidor.

Apesar de sua importância evidente, nós percebemos que há uma falta de modelos e algoritmos capazes de identificar padrões de reconhecimento e interação que poderiam ajudar a esclarecer como a inibição cruzada acontece. Neste contexto, um padrão é um conjunto de atributos conservados, na interface, usados para explicar ou predizer ligações.

Tradicionalmente, métodos de comparação de sequências e/ou alinhamento estrutural têm sido utilizados na detecção de conservação [Zhang et al. (2011); Ribeiro et al. (2010); Melo et al. (2007)]. De acordo com Tuncbag et al. (2011), estruturas são mais conservadas do que as sequências e resíduos da interface ou IRF, do inglês *Interface Forming Residues*, são ainda mais conservado do que toda a estrutura. No entanto, na inibição cruzada podemos lidar com sequências muito diferentes e estruturas tridimensionais completamente dissimilares, o que faz com que os métodos clássicos sejam inadequados.

De fato, com os métodos tradicionais, a detecção de padrões na inibição cruzada limita-se essencialmente nos resíduos já conhecidos e conservados que têm uma participação direta no processo de catálise, como a tríade catalítica, sítios de especificidades e cavidade do oxianionte. Observamos que para avaliar corretamente a eventual contribuição hidrofóbica de toda a interface protease/inibidor, devemos abstrair a semântica de resíduo e avaliar *patches* em nível atômico. Abordagem semelhante tem sido utilizada na caracterização de núcleos de domínios proteicos com estruturas semelhantes, mas muito divergentes nas sequências [Soundararajan et al. (2010a)]. Uma análise em nível atômico é mais adequada, tendo em vista que todos os resíduos têm porções apolares. Lisina,

por exemplo, é considerado um resíduo carregado positivamente (em pH neutro), mas há também vários grupos metis hidrofóbicos.

O reconhecimento enzima-inibidor é determinado por uma rede de interações entre os átomos, portanto, a modelagem em grafo é uma abordagem direta. Nós modelamos os átomos hidrofóbicos como vértices de um grafo e os contatos entre eles como as arestas. Usamos o grafo para obter *patches* hidrofóbicos conservados ou, em outras palavras, componentes conectados.

Supondo que a propriedade mais importante de um *patch* hidrofóbico é onde ele está posicionado para interagir com o ligante, nos abstraímos a sua composição, forma, volume e densidade e representamos um *patch* como um centroide geométrico, que nós chamamos de HP-centroide (*Hydrophobic Patch Centroid*). Neste trabalho, nós propomos uma nova metodologia que envolve modelos e algoritmos para detectar HP-centroides hidrofóbicos conservados em inibição cruzada.

Finalmente, nós apresentamos um estudo de caso qualitativo que consiste em dois exemplos de inibição cruzada. Ambos pertencem à família de serino protease: nós focamos em Tipo Tripsina e Tipo Subtilisina. Elas apresentam estruturas tridimensionais completamente diferentes e a identidade de sequência é menor que 20% [Wallace et al. (1996)]. Entretanto, elas possuem os mesmos resíduos que formam a tríade catalítica (SER-HIS-ASP) em seus sítios ativos. No primeiro caso, nós analisamos 5 complexos com Tipo Tripsina e Tipo Subtilisina inibidas pelo inibidor Eglina C [Betzel et al. (1993)] e no segundo caso, nós analisamos 4 complexos com as mesmas famílias com o inibidor Ovomucoide [Papamokos et al. (1982)]. Além disso, nós verificamos que os HP-centroides obtidos nos complexos também estão presentes em um conjunto de apo estruturas com sequências diversas, sendo fortemente conservados na família.

Capítulo 2

Fundamentação Teórica

2.1 Proteases

As proteases, também conhecidas como proteinases, peptidases ou enzimas proteolíticas, são classes de enzimas que catalisam a clivagem de ligações peptídicas de proteínas. O processo de clivagem, ou seja, a quebra de ligações peptídicas necessita do envolvimento de uma molécula de água, sendo classificado como hidrolase. Muitas dessas enzimas adquirem amplas atividades enzimáticas quando se enovelam espontaneamente em suas formas tridimensionais características, porém, outras são sintetizadas como precursores inativos, que subsequentemente são ativados por clivagem de uma ou mais ligações peptídicas específicas. Estes precursores inativos são chamados de zimogênios ou proenzimas [Berg et al. (2005)].

Proteases são vistas como moléculas sinalizadoras extremamente importantes que estão envolvidas em numerosos processos vitais. Elas são estritamente reguladoras, e a desregulação de suas atividades pode conduzir a patologias, tais como, doenças cardiovasculares e inflamatórias, câncer, osteoporose e desordens neurológicas. Elas ocorrem naturalmente em todos os organismos e correspondem a aproximadamente 2% no genoma humano. Elas estão envolvidas em uma grande variedade de reações metabólicas, como a digestão de proteínas do alimento e cascatas altamente reguladas, como, por exemplo, a da coagulação sanguínea, o sistema do complemento, as vias de apoptose, a cascata ativadora da profenoloxidase nos invertebrados, e replicação e transcrição do DNA [Turk (2006)].

Proteases participam no mecanismo invasivo de tumores; infeções de micro-organismos requerem proteases para replicação ou usam proteases como fator de virulência. Isto tem facilitado o desenvolvimento de terapias de proteases alvos para doenças de grande relevância para a vida humana, como por exemplo a AIDS, tornando muitas proteases focos de atenção para as indústrias farmacêuticas como potenciais alvos para desenvolvimento de novos fármacos ou como biomarcadores de diagnósticos e prognósticos. Em plantas as proteases também têm um papel importante, pois contribuem para o processamento,

maturação ou destruição de conjuntos específicos de proteínas em resposta as variações das condições ambientais [López-Otín e Bond (2008)].

Representantes de enzimas com importantes papeis em processos fisiológicos, as proteases também têm aplicação comercial, estando entre os três maiores grupos de enzimas industriais, sendo responsáveis por 60% da venda internacional de enzimas. São também importantes ferramentas nas indústrias de biotecnologia, porque têm utilidade como reagentes ou na fabricação de inúmeros produtos. É notável a aplicação na indústria de detergentes e de alimentos; tendo em vista os recentes acordos mundiais para o uso de tecnologias não poluentes, as proteases estão substituindo os compostos tóxicos poluentes [Beynon e Bond (2001)].

As proteases são encontradas em micro-organismos, como vírus, bactérias, protozoários, leveduras e fungos. Os micro-organismos representam uma excelente fonte de proteases, devido a sua grande diversidade bioquímica e facilidade de manipulação genética [Rao et al. (1998)].

Os sítios catalíticos de uma protease estão comumente localizados em sulcos na superfície da molécula. O substrato específico, determinado pelas propriedades dos sítios de ligação, são acomodados ao longo dos sulcos onde acontece a hidrólise da ligação peptídica. As proteases podem quebrar ligações peptídicas em sequências específicas de aminoácidos (proteólise limitada), ao passo que outras degradam o peptídeo integralmente (proteólise ilimitada). A clivagem específica de uma proteína pode tanto neutralizá-la, quanto permitir que ela assuma uma conformação ativa, o que pode servir de sinalização para ciclos celulares.

A especificidade de uma protease é descrita pelo uso de um modelo conceitual no qual cada subsítio específico é capaz de acomodar a cadeia lateral de um resíduo de aminoácido simples. Os sítios catalíticos da enzima são numerados, S1, S2...Sn para o N-terminal e S'1, S'2...S'n para o C-terminal e os aminoácidos do substrato são acomodados e numerados P1, P2...Pn, e P'1, P'2...P'n, respectivamente. Os resíduos P1-P1' formam uma ligação chamada "*scissile bond*" ou ligação peptídica do substrato que é clivada pela protease na hidrólise, como mostra esquema abaixo [Beynon e Bond (2001)].

O sítio catalítico da enzima, marcado com asterisco (*) indica o local da catálise e a ligação peptídica que é clivada no substrato é indicada com sinal de mais (+).

$$\text{Substrato: -}P_3\text{ - }P_2\text{ - }P_1\text{ + }P_1'\text{ - }P_2'\text{ - }P_3'\text{ -}$$
$$\text{Enzima: -}S_3\text{ - }S_2\text{ - }S_1\text{ * }S_1'\text{ - }S_2'\text{ - }S_3'\text{ -}$$

As proteases são divididas em endopeptidases e exopeptidases, de acordo com a posição da ligação peptídica a ser clivada na cadeia polipeptídica. A figura 2.1 mostra as divisões deste mecanismo.

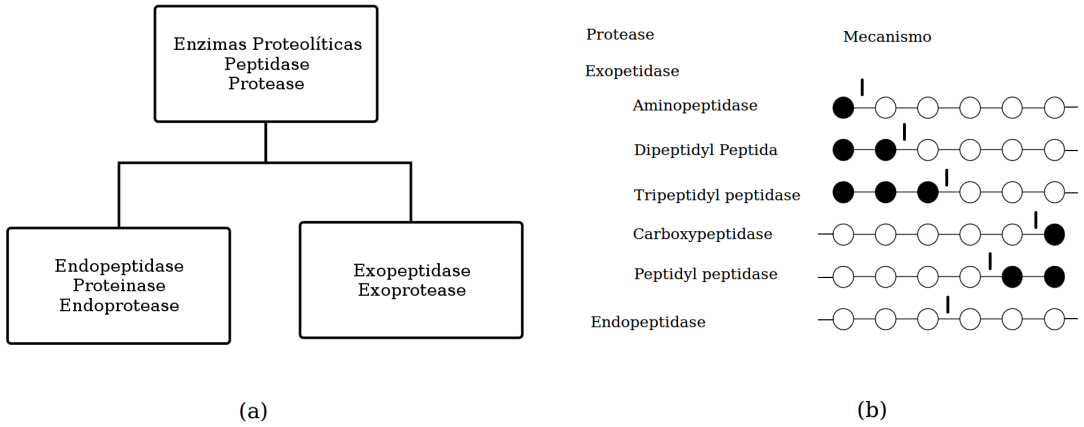


Figura 2.1: Subdivisão das enzimas proteolíticas.Em (a), temos a representação dos tipos de proteases existentes com seus sinônimos e em (b), temos a representação do modo de ação, sendo que os traços indicam o local da clivagem. Figura adaptada de [Beynon e Bond (2001)]

As endopeptidases, também conhecidas como proteinases ou endoproteases, atuam preferencialmente nas regiões internas da cadeia polipeptídica, entre as regiões N- e C-terminais. A presença de grupos α-amino ou α-carboxila terminais tem um efeito negativo na atividade destas enzimas.

As exopeptidases, também conhecidas como exoproteases, atuam somente nos finais das cadeias polipeptídicas na região N- ou C- terminal. Aquelas que atuam na região amino terminal livre liberando um único resíduo de aminoácido são chamadas de amino-peptidases, liberando dois resíduos(dipeptídeo) são chamadas de dipeptidil-peptidases ou liberando três resíduos (tripeptídeo) são chamadas de tripeptidil-peptidases. As exopep-tidases que atuam na região carboxi terminal livre liberando um único aminoácido são chamadas de carboxipeptidases ou liberando dois resíduos (dipeptídeo), são chamadas de peptidil-dipeptidases.

As endo e exo-peptidases também podem ser classificadas em cinco tipos, de acordo com mecanismos catalíticos: serina, cisteína e treonina proteases que formam complexos enzimáticos covalentes, e aspártico e metalo proteases que formam complexos enzimáticos não covalentes. O processo de inibição é diferente para as duas principais classes. A primeira possui aminoácidos fortemente nucleofílicos em seu sítio catalítico. A segunda categoria inclui duas classes de proteases que catalisam a hidrólise de ligações peptídicas sem ataque nucleofílico por um grupo funcional da enzima [Beynon e Bond (2001)].

2.1.1 Classificação de proteases

A classificação e nomenclatura das enzimas são dadas pela IUBMB (*International Union of Biochemistry and Molecular Biology*) que estabeleceu um sistema de classificação de enzimas utilizado mundialmente, chamado EC. (*number Enzyme Classification*) que classifica hierarquicamente todas as enzimas em seis categorias: (1) Oxidoredutases, (2)

Transferases, (3) Hidrolases, (4) Liases, (5) Isomerases e (6)Ligases. Além disso, cada uma destas categorias pode ser subdividida. Por exemplo, as Hidrolases (3) que quebram ligações peptídicas são chamadas de proteases ou peptidases.

2.1.1.1 Classificação EC

As proteases são classificadas, seguindo critério EC (*Enzyme Classification*), na subclasse (3.4) e podem ser divididas entre quatorze sub-subclasses, separadas em exopeptidases (3.4.11/19) e endopeptidases (3.4.21-25/99), conforme mostra a tabela 2.1[IUBMB (1999)].

Tabela 2.1: O sistema EC de classificação de peptidases.

Tipo Hidrolise	Sub-subclasse	Tipo de peptidase
Exopeptidase	3.4.11	Aminopeptidases
	3.4.13	Dipeptidases
	3.4.14	Dipetidil-peptidases
	3.4.15	Peptidil-dipeptidases
	3.4.16	Serino-tipo carboxipeptidases
	3.4.17	Metallocarboxipeptidases
	3.4.18	Cisteine-tipo carboxipeptidases
	3.4.19	Omega peptidases
Endopeptidase	3.4.21	serino endopeptidases
	3.4.22	Cisteina endopeptidases
	3.4.23	Aspartico endopeptidases
	3.4.24	Metallo endopeptidases
	3.4.25	Treonine endopeptidase
	3.4.99	Endopeptiddases de tipos desconhecidas

2.1.1.2 Classificação MEROPS

O sistema de classificação EC divide todas as peptidases em somente 14 subclasses e deste total somente 6 são endopeptidases. Entretanto, elas são muito diferentes, sendo necessário um agrupamento estrutural, o que refletiria relacionamentos evolucionários. Desta maneira, foi proposto em 1992 por Rawlings e colaboradores do *Wellcome Trust Sanger Institute* a criação de um sistema de classificação de peptidases que agrupasse de acordo com características estruturais e relacionamentos evolucionários. Em 1996 o sistema de classificação foi publicado na internet como a base de dados MEROPS[1]. [Beynon e Bond (2001)]

Atualmente, a base de dados MEROPS é uma fonte de pesquisa para peptidases e as proteínas que as inibem. Possui três níveis de classificação de forma hierárquica: Clãs, Famílias e Peptidases [Rawlings et al. (2008)].

[1]A base MEROPS pode ser acessada em http://www.merops.co.uk

Os clãs são grupos de famílias onde todas as peptidases evoluíram de um único ancestral, ou seja, são homólogas. O melhor tipo de evidência para suportar a formação de um clã é a similaridade nas estruturas tridimensionais de suas peptidases. Para identificar um clã é atribuída a ele uma letra maiúscula, indicando o tipo de catálise das peptidases do grupo, ou seja, A-aspártico, C-cisteína, M-metalo, S-serina, T-treonina, U-não classificado e ainda P para identificar clãs que contém peptidases de mais de um tipo de catálise (C, S e T). Além disso, os clãs são divididos em subclãs para identificar as homologias entre as famílias contidas no clã; assim, é atribuída ao identificador do clã mais uma letra sequencialmente, como por exemplo, (AA, AB e AC).

As famílias são grupos de peptidases com base na estatística significativa de similaridade na comparação da sequência de aminoácidos. Cada família é representada por uma peptidase e muitos membros da família mostram significantes similaridades na sequência de aminoácido, com relação ao outro membro da família, na parte da molécula que é responsável pela atividade da peptidase. Porém, uma característica essencial de uma família é que os membros contidos nela são evolucionariamente relacionados. Para identificar uma família é atribuída a ela uma letra com um número, por exemplo, (C14). A letra denota o tipo de catálise, assim como nos clãs, ou seja, (S,C,T,A , M ou U, para serina, cisteína, treonina, aspártico, metalo ou não classificadas, respectivamente) e os números tipos de variações relevantes que subdivide a família de peptidase. A figura 2.2, mostra uma estatística atual do número de famílias contidas em cada clã.

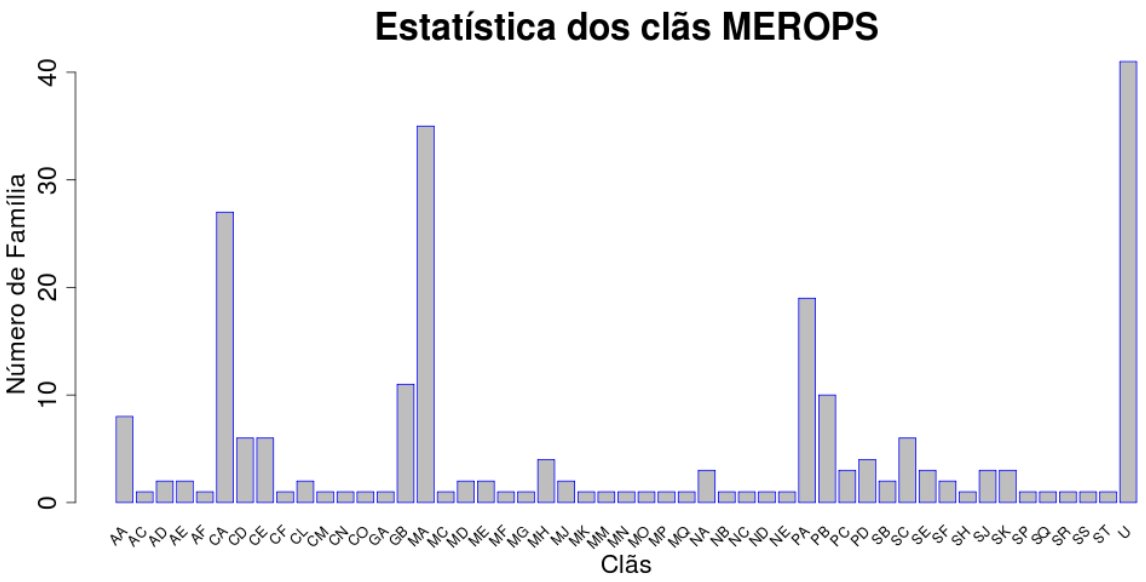


Figura 2.2: Gráfico representativo do número de famílias inseridas dentro de cada clã MEROPS.

As peptidases são distinguidas umas das outras pela diferença na atividade que elas desempenham. Quando identificadas, são agrupadas dentro de uma família que por sua vez pertence a um clã.

2.1.1.3 Classificação SCOP

O SCOP[2] (*Structural Classification of Proteins*) é uma base de classificão de estruturas em proteínas em geral. É gerada por inspeção manual e auxiliada por um conjunto de métodos automatizados e que tem como objetivo fornecer uma detalhada descrição das relações evolutivas entre todas as estruturas conhecidas e depositadas no PDB. Possui uma organização estrutural de forma hierárquica, ou seja, Classe/Enovelamento/Superfamília/Família e Domínio [Murzin et al. (1995)].

As Classes são definidas com base na composição das cadeias em termos de estruturas secundárias: α (formadas, na maioria, por α hélices) ou β (formadas, na maioria, por folhas-β) ou uma junção delas. O Enovelamento é uma classificação de proteínas com base em um conjunto de características de estruturas secundárias. Superfamília contém proteínas que são possivelmente relacionadas evolutivamente compartilhando o mesmo enovelamento e desempenhando funções bastante similares. As Famílias são compostas de proteínas de alta similaridade tanto na sequência primária quanto na estrutura tridimensional. O Domínio é uma parte da cadeia polipeptídica que pode se enovelar independentemente para formar uma estrutura compacta e estável.

2.2 Base de dados

Existem muitas bases de dados que contém informações sobre enzimas e proteínas em geral e a maioria são baseadas em relatos da literatura. Além disso, elas podem ser de estruturas como o PDB (*Protein Data Bank*) ou apenas de dados de sequências.

2.2.1 PFAM

O PFAM [3] é uma extensa e acurada base de dados que contêm informações sobre domínios e famílias de proteínas. É formada a partir das bases de dados UniProt, GenPept[4] e nas sequências de projetos meta genômicos e é baseada em alinhamentos múltiplos de sequências e em modelos estatísticos (modelo oculto de Markov). As famílias de proteínas estão divididas em duas categorias, Pfam-A e Pfam-B.

O Pfam-A consiste em uma parte curada manualmente que contém um pequeno conjunto de entidade representativo da família. Para cada entidade um alinhamento da sequência da proteína e um modelo oculto de Markov é armazenado. Os modelos ocultos de Markov podem ser usados para pesquisar bancos de dados de sequência. As entradas no Pfam-A não cobrem todas as proteínas conhecidas, sendo necessário um suplemento que é gerado automaticamente chamado Pfam-B.

[2] O SCOP é uma base de classificação estrutural hierárquica de proteínas
[3] A base Pfam pode ser acessada em http://pfam.sanger.ac.uk/
[4] UniProt e GenPept são bases de dados de sequências e funções de proteínas

O Pfam-B contém um grande número de pequenas famílias derivadas de agrupamentos produzidos por um algoritmo chamado ADDA. Embora de qualidade inferior, as famílias descritas no Pfam-B podem ser úteis quando não encontradas no Pfam-A.

O Pfam também é subdividido em clãs. Um clã define uma simples classificação hierárquica contendo uma coleção de entidades, alocadas em famílias, do Pfam-A que são homólogas, permitindo uma melhor transferência de informação estrutural e/ou funcional entre famílias e uma melhor predição de função e estrutura para famílias de funções desconhecidas. A formação de um clã é feita manualmente, baseada em uma variedade de informações como a literatura, estruturas conhecidas, comparações perfil a perfil e outras bases de dados de origem como o SCOP.

A versão atual do Pfam (25.0) contém mais de 12 mil diferentes famílias de proteínas. As famílias são identificadas na base de dados do Pfam através de um identificador que se inicia com as letras PF seguidas de 5 dígitos. Por exemplo, as famílias Tipo Subtilisina PF00082 e Tipo Tripsina PF00089 [Finn et al. (2010)].

2.2.2 BRENDA

A base BRENDA[5] (*BRaunschweig ENzyme Database*) é uma abrangente base de dados que provem informações sobre enzimas. A maioria dos dados são manualmente extraídos da literatura, portanto, não garante que todas as enzimas relatadas em sua base de dados têm estruturas tridimensionais resolvidas.

As informações contidas são sobre função, estrutura (quando houver), ocorrência, preparação e aplicação de enzimas bem como propriedades de mutantes e variantes de engenharia [Scheer et al. (2011)].

2.3 serino proteases

As serino proteases são enzimas proteolíticas que estão presentes em vários organismos. Atuam como enzimas digestivas de procariotos e eucariotos e são caracterizadas pela presença do resíduo de aminoácido serina, particularmente reativo, em seu sítio ativo, essencial para a atividade enzimática [Voet e Voet (2002)].

Desde a determinação da estrutura tridimensional da Quimotripsina, enzima digestiva produzida pelo pâncreas, por Blow e seus colaboradores em 1967, o interesse pelas serino proteases é contínuo. Elas estão amplamente distribuídas na natureza e são encontradas em todos os reinos da vida celular, bem como muitos genomas virais. Em mamíferos, algumas enzimas, tais como, a Quimotripsina, a Tripsina, a Elastase, a Calicreína e Trombina, podem participar de numerosos processos fisiológicos, como a digestão, a coagulação sanguínea, fertilização, ativação da resposta imune e em vários estágios de doenças como enfisemas, metástase de tumor e artrites [Lesk e Fordham (1996)]. Elas podem ser encon-

[5]BRENDA está disponível em http://www.brenda-enzymes.org

tradas também em micro-organismos, como a Subtilisina produzida por diversos gêneros de bacilos como *Bacillus subtilis*.[Voet e Voet (2002)].

De acordo com a base de dados MEROPS, as serino proteases estão agrupadas nos clãs P (misto de C, S, T) e S (Serina) de acordo com a similaridade estrutural e a homologia e dentro de diversas famílias de acordo com a similaridade de sequência. Dentre os clãs de serino proteases citados anteriormente podemos destacar o PA contendo a família S1 (conhecida como Quimotripsina) e o clã SB contendo as famílias S8 e S53 (conhecidas como Tipo Subtilisinas). Elas possuem muitas semelhanças com relação aos resíduos que formam a tríade catalítica, apesar de estarem em clãs diferentes, ou seja, não são homólogas. Embora as famílias Tipo Tripsinas e Tipo Subtilisina não sejam classificadas pelo MEROPS como sendo da mesma família, elas pertencem ao mesmo grupo no sistema de classificação enzimática (E.C)[Rawlings et al. (2008)].

As serino proteases são classificadas enzimaticamente no sistema E.C. como (3.4.21) que usa o clássico mecanismo de ação da tríade catalítica **SER/HIS/ASP**, onde a serina é nucleofílica, a histidina é a base e ácido gerais, e o ácido aspártico ajuda a orientar o resíduo de histidina e a neutralizar a carga que desenvolve na histidina durante o estado de transição [Page e Di Cera (2008)]

O mecanismo de catálise inicia-se quando a enzima liga-se ao substrato formando um complexo de Michaelis-Menten e em seguida ocorre o ataque da Ser reativa à carbonila da ligação peptídica do substrato para formar um intermediário acyl-enzima. Para que isso ocorra, é necessário a presença dos resíduos de aspartato (ASP) e histidina (HIS) que interagem através de ligações de hidrogênio em um mecanismo de dupla transferência de próton, comumente chamado de "charge-relay"[Ekici et al. (2008)].

Existem muitas famílias de serino proteases, entretanto, duas se destacam por algumas características peculiares. As famílias Tipo Tripsina e Tipo Subtilisina não são homólogas, mas compartilham os mesmos resíduos da tríade catalítica para quebrarem seus substratos e possuem estruturas tridimensionais diferentes.

2.3.1 Tipo Subtilisina

As enzimas do Tipo Subtilisina são a segunda maior família de serino protease, em termos de número de sequências e peptidases caracterizadas e estão presentes em plantas, bactérias, fungos, protozoários, vírus e com poucos representantes no genoma animal [Page e Di Cera (2008)]. A maioria dos membros da família possui uma tríade catalítica com seus resíduos de aminoácidos na ordem ASP 32, HIS 64 e SER 221 e são ativados em pH alcalino com preferência para clivar seus substratos após resíduos hidrofóbicos, porém alguns membros preferem clivar após aminoácidos dibásicos [Siezen e Leunissen (1997)].

Quanto a sua estrutura tridimensional, as Tipo Subtilisinas são classificadas como α/β. É formada em sua grande maioria por três camadas com sete folhas β sobrepostas entre duas porções de hélices, conforme mostra a figura 2.3.

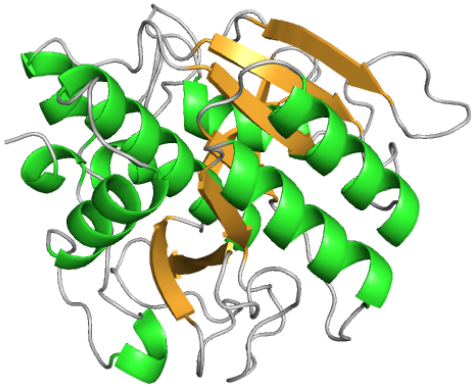


Figura 2.3: Estrutura tridimensional das Tipo Subtilisina (PDB ID:1GCI)

Possuem resíduos conservados (responsáveis pela catálise), ligações dissulfeto, sítios de ligação de cálcio (contribuem para a estabilidade térmica), sítio de ligação de substrato e interações iônicas e aromáticas [Roland et al. (1997)]. De acordo com MEROPS, as Tipo Subtilisinas são classificadas no clã SB, divididas em duas famílias (S8 e S53).

2.3.2 Tipo Tripsina

As enzimas do tipo tripsina são enzimas proteolíticas do tipo serino proteases e incluem várias enzimas de mamíferos como quimotripsina, tripsina, elastase, calicreína, trombina etc. Sua estrutura tridimensional é composta por grande maioria de unidades denominadas folhas β, conforme mostra figura 2.4. Assim como nas enzimas do Tipo Subtilisinas, as Tipo Tripsinas possuem três resíduos no bolsão catalítico, os quais são denominados de tríade catalítica.

Figura 2.4: Estrutura tridimensional de Tipo Tripsina (PDB ID: 1HJ8)

As enzimas do Tipo Tripsina estão onipresente em muitos seres vivos e desempenham diversos papéis, especialmente no sistema digestivo, em diferentes filos. Em animais eucariotos são responsáveis por muitas funções essenciais, incluindo processos de digestão dos

alimentos, hemostasia, resposta de defesa imunológica, e do sistema nervoso, reprodução e transdução de sinal. No sistema nervoso central, algumas enzimas como a trombina e plasmina, tem recebido grande atenção. Essas proteases são expressas no cérebro, e têm um papel funcional na regulação das consequências de acidente vascular cerebral isquêmico, plasticidade sináptica, neuro degeneração e neuro regeneração [Wang et al. (2008)].

A classificação proteolítica das enzimas do Tipo Tripsina, segundo MEROPS, é representada, em geral, pelo clã PA e são altamente expressas em eucariotos com componentes raros em genomas procarióticos.

O Clã PA contém o maior número de família de peptidase do tipo serina e talvez o grupo mais estudado de enzimas. Existem muitos estudos envolvendo enzimas digestivas como a tripsina e quimotripsina, que clivam cadeias polipeptídicas do lado C-terminal próximo a uma cadeia lateral de (ARG e LYS) carregada positivamente ou resíduo hidrofóbico grande (PHE, TRP, TYR), respectivamente.

A maioria das peptidases do clã PA tem seletividade por cadeias laterais da ARG no substrato. No entanto, muitos membros virais ou bacterianos da família são específico para Gln. A diferença é marcante no mecanismo pelo qual a seletividade do substrato é alcançada nas duas subfamílias de clã PA (S1A e S1B).

S1A e S1B são filogeneticamente distintos grupos de enzimas, no entanto, compartilham uma arquitetura comum, conhecida com dois β-barris. As peptidases S1B são encontradas em toda vida celular e são responsáveis pela deposição de proteínas intracelulares. Em contrapartida, as peptidases de S1A, como por exemplo a tripsina, medeiam uma variedade de processos extracelulares. Peptidases S1A têm uma distribuição limitada nas plantas, procariontes e arquéia. Quase todo o clã PA utiliza a tríade canônica catalítica, mas alguns membros da família de origem viral usam um sítio ativo tiol de um resíduo de Cys. A Catálise é favorecida por uma ligação entre o ASP-102 e HIS-57 (seguindo a numeração da quimotripsinogênio), o que facilita a captação do próton da SER-195 gerando um potente nucleófilo [Page e Di Cera (2008)] .

2.4 Convergência evolutiva em famílias de serino proteases

A convergência evolutiva ou evolução convergente é um fenômeno evolutivo observado em seres vivos quando estes desenvolvem características semelhantes a partir de origens diferentes (ou análogas). Esse fenômeno está associado à seleção natural, onde mutações que geram alterações morfológicas propícias a determinado ambiente são selecionadas em detrimento de outras menos adequadas. Desta forma, seres vivos que compartilhem o mesmo habitat, ou mesmos hábitos de vida, podem desenvolver estruturas similares que os tornam capazes de sobreviver àquelas condições[Berg et al. (2005)].

Em proteínas esse fenômeno pode ser observado nos membros das famílias de serino

proteases, tipo Subtilisina (α/β) e tipo tripsina (toda β). Elas estão distribuídas em diferentes clãs e consequentemente não são homólogas; isto pode ser elucidado pelas diferenças marcantes em suas estruturas tridimensionais, entretanto, a topologia de seus centros ativos é muito semelhante, ou seja, os resíduos que formam a catálise enzimática, conhecidos como tríade catalítica, são os mesmos apesar de aparecerem em ordens diferentes na sequência da cadeia polipeptídica. Em membros da família Tipo Tripsina a tríade é formada por HIS/ASP/SER, enquanto na família Subtilisina a tríade é formada por ASP/HIS/SER. Além dessa variação na ordem, a posição dos resíduos também não são iguais para as duas famílias. Apesar de todas essas diferenças elas hidrolisam seus substratos pelo mesmo mecanismo (veja figura 2.5).

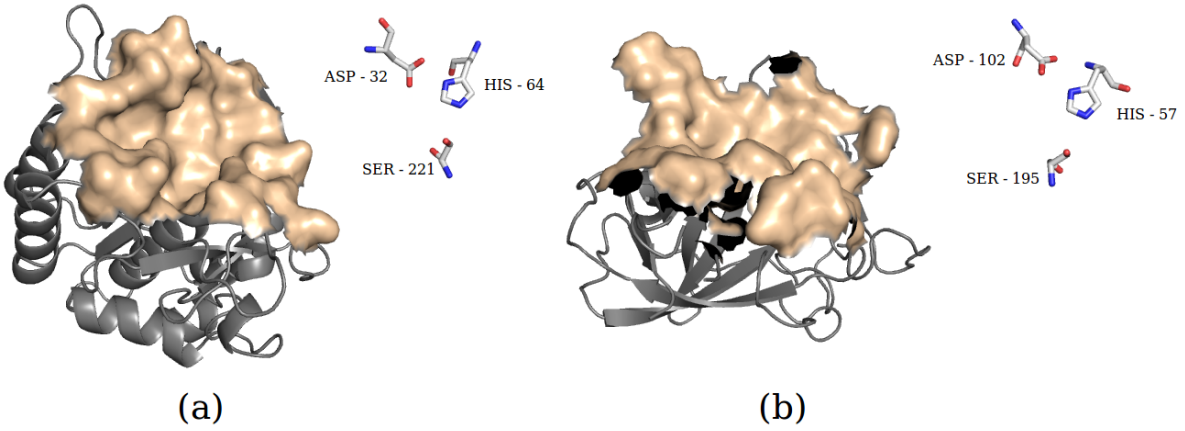


Figura 2.5: Semelhança na topologia dos resíduos da tríade catalítica de Tipo Subtilisina e Tipo Tripsina. Em (a), temos Tipo Subtilisina 1R0R e sua tríade catalítica e em (b), temos Tipo Tripsina 1PPF e sua tríade catalítica.

A figura 2.6 mostra o alinhamento das sequências da Tipo Subtilisina (Subtilisin Carlsberg de *Bacillus licheniformis*-1R0R) e Tipo Tripsina (Human Leukocyte Elastase de *Homo sapiens*-1PPF) com um *score* de identidade de 6%.

```
1ROR_E|PDBID|CHAIN|SEQUENCE      AQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQASHPDLNVVGGASF 50
1PPF_E|PDBID|CHAIN|SEQUENCE      ---------IVGGRRARPHAWP------FMVSLQLRGGH-----FCGATL 30
                                          :: . :.:.:.:        ::.  ::..*     . **::

1ROR_E|PDBID|CHAIN|SEQUENCE      VAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVAPSVSLYAVKVLNSSGS 100
1PPF_E|PDBID|CHAIN|SEQUENCE      IAPNFVMSAAHCVANVNVRAVRVVLGAHNLSRREPTRQVFAVQRIFENGY 80
                                 :* :   : .:  ..  . :* .: .: .:    *: .::**: : ..*

1ROR_E|PDBID|CHAIN|SEQUENCE      GSYSGIVSGIEWATTNGMDVINMSLGGASGSTAMKQAVDNAYARGVVVVA 150
1PPF_E|PDBID|CHAIN|SEQUENCE      DPVN-LLNDIVILQLNGSATINANVQ-----VAQLPAQGRRLGNGVQCLA 124
                                 .. . ::..*     **  .** .:      .*   * ..  ..**  :*

1ROR_E|PDBID|CHAIN|SEQUENCE      AAGNSGNSGSTNTIGYPAKYDSVIAVGAVDSNSNRASFSSVGAELEVMAP 200
1PPF_E|PDBID|CHAIN|SEQUENCE      MG--WGLLGRNRGIASVLQELNVTVVTSLCRRSNVCTL--------VRGR 164
                                   .   * * .. *.   :  .* .* ::  .** .::        * .

1ROR_E|PDBID|CHAIN|SEQUENCE      GAGVYSTYPTNTYATLNGTSMASPHVAGAAALILSKHPNLSASQVRNRLS 250
1PPF_E|PDBID|CHAIN|SEQUENCE      QAGVCFGDSGSPLVCNGLIHGIASFVRGGCASGLYPDAFAPVAQFVNWID 214
                                  ***    . .. .  .     :..* *..*  *  ..  ..:*. * :.

1ROR_E|PDBID|CHAIN|SEQUENCE      STATYLGSSFYYGKGLINVEAAAQ 274
1PPF_E|PDBID|CHAIN|SEQUENCE      SIIQ-------------------- 218
```

Figura 2.6: Alinhamento das sequências da Tipo Subtilisina (1R0R) e tripsina (1PPF)

Apesar de serem proteases distantes em suas sequências primárias, elas são classificadas pelo mesmo tipo enzimático (serino) e conservam propriedades em seus bolsões catalíticos.

A família tipo tripsina é considerada a maior de todas as famílias de serino proteases, pelo número de proteínas que foram sequenciadas ou pelo número de atividades distintas. Os membros dessa família, como por exemplo, quimotripsina, tripsina, elastase, termolisina, pepsina e endopeptidase V8, possuem alta similaridade, o que indica que elas se originaram a partir da duplicação de um gene de uma serino protease ancestral, seguida pela evolução divergente das enzimas resultantes [Voet e Voet (2002)]. Todas as enzimas dessa família contêm uma tríade catalítica, no centro ativo, formada por HIS, ASP e SER, nesta ordem, podendo variar a posição desses aminoácidos na cadeia polipeptídica.

Estão distribuídas em famílias de bactérias, protozoários, fungos, plantas e vírus, mas prevalece em animais, desempenhando muitas funções biológicas, como a digestão intestinal e respostas imunes. Possui relevância farmacêutica e biotecnológica, pois membros dessa família são potenciais alvos de drogas.

A maioria dos membros dessa família tem na posição N-terminal um peptídio sinal. Elas são sintetizadas na forma de precursores com uma extensão N-terminal que é clivada para formar a enzima ativa. A clivagem do precursor é essencial para habilitar o rearranjo estrutural que acontece próximo ao N-terminal exposto. Geralmente a clivagem ocorre antes dos resíduos ILE ou VAL. Sua estrutura tridimensional é formada de dois domínios, cada um contendo um β barril, abertos em ângulos para o outro [Rawlings et al. (2008)].

2.5 Inibidores de serino proteases

Proteases, em geral, desempenham muitas funções essenciais para a vida, entretanto, sem controle ou variações em seu funcionamento podem causar danos ou estar associadas a muitas doenças. Os inibidores são usados como forma principal de controle, uma vez que a enzima foi ativada ou contribui para a manifestação de doenças. Eles formam complexos completamente inativos (inibição total) ou parcialmente ativos (inibição parcial) com suas enzimas cognatas.

Os inibidores de serino proteases são muito comuns na natureza e têm a função regulatória e de proteção. Em certas plantas, por exemplo, são liberados inibidores de proteases em resposta a picadas de insetos provocando a morte do inseto por inanição devido a inativação das enzimas digestivas. Inibidores de serino proteases constituem cerca de 10% das quase 200 proteínas presentes no soro sanguíneo. Por exemplo, o inibidor de proteinases a1, que é secretado pelo fígado, inibe a Elastase dos leucócitos, pois, a ação da Elastase dos leucócitos faz parte do processo inflamatório. Variantes patológicos do inibidor de proteinase a1 com atividade reduzida estão associadas com enfisema pulmonar [Voet e Voet (2002)].

Existem diversos grupos de inibidores de serino proteases, incluindo inibidores químicos sintéticos para pesquisas ou com propósitos terapêuticos, e também inibidores naturais

proteicos. Este trabalho aborda somente os inibidores proteicos de serino proteases.

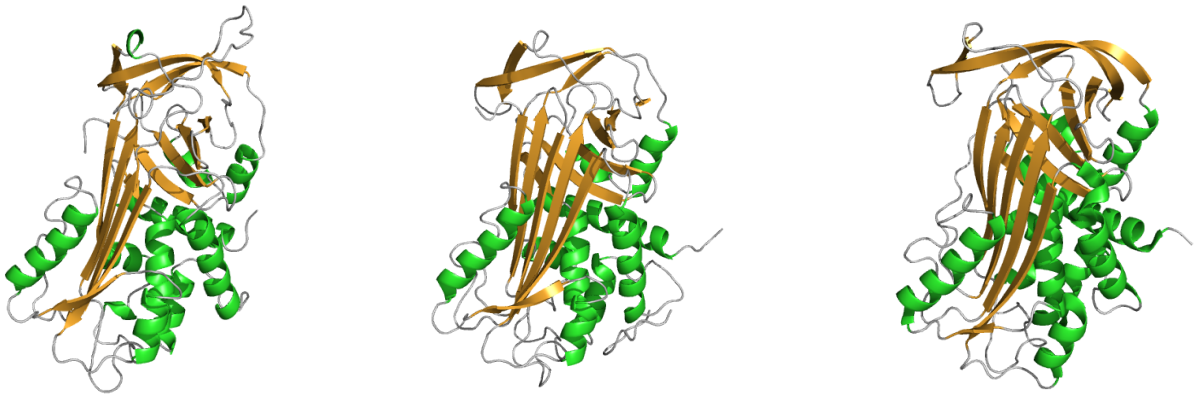
Segundo Krowarsch et al. (2003) existem três classes de inibidores proteicos de serino proteases. O primeiro deles é o grupo das Serpinas (abreviação do inglês *Serine protease inhibitors*). O segundo grupo são dos inibidores canônicos que é constituído por 18 famílias diferentes, porém, nem todas com estrutura tridimensional conhecida e o terceiro grupo são dos inibidores não canônicos.

2.5.1 Serpinas

As serpinas são classificadas como super família de inibidores de serino proteases e estão envolvidas em inúmeros processos biológicos fundamentais, tais como a coagulação sanguínea (chamadas de macroglobulina), ativação do complemento, fibrinólise, angiogêneses, inflamação e supressão de tumor, e são expressas de alguma maneira em células específicas [Gent et al. (2003)].

Todas as serpinas foram classificadas em 16 grupos. As serpinas de vertebrados estão classificadas em 6 subgrupos. No plasma humano elas representam aproximadamente 2% do total de proteínases, do qual 70% são α-1-antitripsina.

Quanto a sua estrutura e mecanismo de ação elas são caracterizadas por um número variado de folhas β e 8 ou 9 α hélices (veja figura 2.7). O modo de ação inclui o "*reative centre loop(RCL)*", uma região na proteína com o centro de clivagem entre P1 e P1'. Em geral os RCL são compostos de aproximadamente 20 resíduos de aminoácidos próximos a região C- terminal [Mangan et al. (2008)].



(a) (b) (c)

Figura 2.7: Estruturas tridimensionais de inibidores da classe Serpina. (a) representa a estrutura de α1-Antitripsina (1PSI), (b) representa a estrutura de uma Antitrombina (1ANT) e (c) representa a estrutura de Plasminogênio Ativador Inibitor-1A (1DVN).

2.5.2 Inibidores canônicos

O mecanismo de interação dos inibidores de serino proteases com suas enzimas cognatas é consideravelmente bem detalhado. Estudos de sequenciamento e cristalografia de

raio-X tem mostrado que esses inibidores não são todos homólogos, consistindo de muitas (pelo menos 18) famílias [Laskowski e Qasim (2000b)].

Surpreendentemente, entretanto, a grande maioria dos inibidores de serino proteases interage seguindo um "mecanismo padrão". Esta situação é análoga a existência de pelo menos duas famílias de serino proteases não homólogas, a família Tipo Subtilisina e Tipo Tripsinas, as quais não compartilham uma estrutura tridimensional comum, mas hidrolisam seus substratos e são inibidas pelos seus inibidores por meio do mesmo mecanismo (veja tabela 2.2).

Tabela 2.2: Famílias de inibidores de serino proteases com estrutura cristalina resolvida. Inibidores de origem [a]animal, de [b]plantas e de [c]micro-organismos.

Nome da família	PDB ID Representativo	
	Livre	Complexo
BPTI Kunitz[a]	1BPI	2PTC
STI Kunitz[b]	1AVU	1AVW
BBI[b]	1PI2	1SMF
Batata I[b]	1MIT	1ACB
Batata II[b]	1TIH	1ACB
Abobora[b]	2CTI	1PPE
Kazal[b]	2OVO	1CHO
SSI[c]	2SSI	2SIC
Antistasina[a]	1SKZ	1HIA
Chelonianina[a]	2REL	1FLE
Ascaris[a]	1ATA	1EAI
Ecotina[c]	1ECY	1AZZ

Ao contrário dos inibidores da família das serpinas, os inibidores canônicos com estruturas tridimensionais conhecidas, são bem menores, como pode ser vista na figura 2.8

(a) (b) (c) (d)

Figura 2.8: Estruturas tridimensionais de inibidores canônicos de serino proteases. (a) representa a estrutura do BPTI (1BPI), extraído de *Bos taurus*, (b) representa a estrutura de inibidor tipo batata I (1MIT), extraído de *Cucúrbita máxima*, (c) representa a estrutura do inibidor tipo batata II (1TIH), extraído de *Nicotiana alata* e (d) representa a estrutura de STI (1AVU), extraído da *Glycine max*.

2.5.3 Inibidores não canônicos

Os inibidores não canônicos interagem através de seus segmentos N-terminal no qual liga ao sítio ativo formando uma curto paralelo de folha β. Esta classe de inibidores

também formam extensivas interações secundárias, fora do sítio ativo, com a proteases alvos, que provem um área de contato adicional e contribui significativamente para força, velocidade e especificidade do reconhecimento [Huntington e Carrell (2001)].

Um exemplo clássico de um inibidor não canônico é a Hirudina, também conhecida como antitrombina, que é reconhecida pela enzima Trombina. A Hirudina, que é extraída da saliva da sanguessuga *Hirudo medicinallis*, tem ação anticoagulante e pode ser usada na fabricação de fármacos capazes de dissolver coágulos sanguíneos [Stubbs et al. (1995)]. Além disso, inibidores sintéticos baseados em Hirudina são atualmente produzidos por indústrias farmacêutica e são usados na profilaxia e no tratamento de trombose e de outros distúrbios.

Hirudina é também o nome de uma família de inibidores, incluindo a própria Hirudina, que possuem origens evolucionárias comuns. Eles possuem aproximadamente 7KDa com uma cadeia polipeptídica estabilizada por três pontes dissulfeto altamente conservadas.

Os inibidores do tipo não canônicos são menos abundantes que os inibidores canônicos ou as serpinas. Eles ocorrem somente em organismos sugadores de sangue e inibem proteases envolvidas em formação de coágulos (Trombina ou Fator Xa). Somente poucas delas foram caracterizadas em termos de estruturas e cinética de interações com proteases alvos.

A figura 2.9 mostra um exemplo de estrutura tridimensional desta classe de inibidores específicos. O arquivo PDB ID 1TOC foi adaptado para visualização somente da cadeia do inibidor que está em complexo.

Figura 2.9: Exemplo de estrutura tridimensional de um inibidor não canônico de serino proteases (PDB ID 1TOC)

2.6 Inibição Cruzada

A inibição enzimática é um fenômeno que ocorre quando uma molécula se liga a uma enzima e altera a sua atividade. Conforme detalhado mas seções anteriores, existem vários tipos de inibição e dentre este tipos, inúmeras famílias de inibidores. O reconhecimento molecular é um processo complexo que envolve interações atômicas presentes nos resíduos contidos em uma região conhecida como interface molecular ou IFR.

Ao que se sabe, os resíduos e a geometria de uma IFR são bem preservados em uma família. O que pode explicar o fato de existir inibidores atuando em diferentes membros de

uma mesma família. No entanto, o que dizer quando enzimas de famílias não homólogas, ou seja, que não compartilham a mesma estrutura tridimensional, são inibidas pelo mesmo inibidor?

A ocorrência deste fenômeno é observado em pelo menos dois inibidores (Eglina C e Ovomucoide), envolvendo dados estruturais de duas diferentes famílias de serino protease; as Tipo Tripsina e Tipo Subtilisina. Entretanto, quando faz-se uma busca em bases de dados de sequências ou bases de dados de ocorrências experimentais (literatura) a quantidade de instâncias é bem maior.

O fato é que não existe uma terminologia apropriada para descrever esta falta de especificidade de alguns inibidores. Neste trabalho nós propomos o uso do termo *Inibição Cruzada*, uma vez que, o inibidor cruza as fronteiras de uma família para inibir outra. Neste sentido, nosso desafio é tentar entender e predizer por que isso ocorre. Para tanto, nós desenvolvemos uma metodologia para investigar, de forma minuciosa, as interfaces moleculares.

Nossa metodologia ultrapassa as análises tradicionais (alinhamentos sequências/estruturas) e investiga a interface de contato, a nível atômico, na busca por padrões conservados. Para isso, nós usamos os conceitos de grafos para modelar este problema e posteriormente usar ferramentas computacionais para analisar estes grafos e tentar identificar os padrões conservados que justifiquem a inibição cruzada.

2.7 Modelagem via grafos

Nesta seção, nós descrevemos uma breve explicação do que são grafos e como eles podem nós ajudar a entender a inibição cruzada. Explicamos também, como nós mapeamos as interações na interface dos complexos enzima/inibidor.

A teoria de grafos é uma área da matemática e computação que estuda a relação entre objetos de um determinado conjunto, sendo que cada objeto é denominado de vértice e a relação entre eles de aresta. Formalmente, um grafo é $G = (V; E)$ é uma coleção V de nós (ou vértices) conectados por um conjunto E de arestas (do inglês *Edges*). Grafos são usados para descrever, modelar e analisar diversos fenômenos físicos tais como redes de energia elétrica e comunicação, sistemas sociais como redes sociais ou corporativas, hierarquias políticas e redes biológicas Colizza et al. (2006)Sales-Pardo e Amaral (2007).

Com o avanço das técnicas de alto desempenho para sequenciamento genômico em larga escala e o crescimento das bases dados de interações moleculares, o desenvolvimento de modelos e algoritmos para análise de redes biológicas tem papel fundamental. Estas redes podem descrever interações atômicas em proteínas, redes de transcrição, redes metabólicas, dados de expressão gênica e a combinação de redes formadas a partir de superposições de conjuntos de dados de rede múltiplas Bachman e Liu (2009).

Em enzimas, as interações atômicas que ocorrem na interface de contato podem ser vista como as arestas e átomos os vértices. Em nosso trabalho, nós selecionamos os

resíduos que compõem a IFR e mapeamos os contatos entre os átomos hidrofóbicos. A técnica usada para identificar os contatos é baseado na geometria computacional e será explicada com detalhes na seção de materiais e métodos.

O mapeamento das redes de interações atômicas na interface dos complexos envolvendo inibição cruzada tem papel importante para delimitar regiões ou componentes conexos hidrofóbicos, ou seja, os grafos. Entretanto, os grafos têm tamanho e elementos bem diversificados, sendo assim necessária uma abstração que foi além da caracterização de padrões no nível dos resíduos. Para representar cada grafo na interface foi necessário calcular um centroide geométrico hidrofóbico capaz de comparar regiões em diferentes interfaces.

Com esta metodologia de análise, nós propomos identificar padrões hidrofóbicos conservados na interface molecular de complexos de serino protease que justifiquem a ocorrência de inibição cruzada

Capítulo 3

Objetivos

3.1 Objetivo Geral

Identificar padrões hidrofóbicos conservados na interface molecular em complexos, com inibidores proteicos canônicos, de serino proteases que justifiquem a ocorrência de inibição cruzada.

3.1.1 Objetivos específicos

- Identificar estruturas em complexos com inibição cruzada na base de dados do PDB;
- Construir uma base de dados não redundante (similaridade de sequência menor que 50%) de apo enzimas de famílias de serino proteases;
- Estudar métodos para identidicar resíduos na interface molecular dos complexos;
- Estudar metodologias para identificação de contatos em interfaces enzima-inibidor;
- Identificar padrões conservados em grafos representativos de interações em interfaces enzima-inibidor;
- Propor algoritmos e estratégias de avaliação de padrões de interações.

Capítulo 4

Materiais e Métodos

4.1 Seleção dos dados

Nosso trabalho começou pela pesquisa por estruturas tridimensionais de enzimas do Tipo Tripsina e Tipo Subtilisina na base de dados PDB [Berman et al. (2000)]. As estruturas encontradas foram separadas em complexos e apo enzimas (cadeias simples). As estruturas que tinham identidade de sequência maior que 50% foram descartadas para garantir a diversidade e reduzir a redundância. As estruturas muito similares podem gerar uma análise tendênciosa ou enviesada. Além disso, somente complexos com casos de inibição cruzada foram considerados nas análises. Após busca das estruturas, nós tínhamos 4 bases para analisar:

1. 5 complexos com inibição cruzada com o inibidor Eglina C (tabela 4.1);

Tabela 4.1: Tabela de inibição cruzada com Eglina C

PDB ID	Molécula	Família	EC. Number
1ACB	Alpha-Chymotrypsin	**Tipo Tripsina**	3.4.21.1
1TEC	Thermitase	Tipo Subtilisina	3.4.21.66
1CSE	Subtilisin Carlsberg	Tipo Subtilisina	3.4.21.62
1MEE	Mesentericopeptidase	Tipo Subtilisina	3.4.21.14
1SBN	Subtilisin NOVO BPN'	Tipo Subtilisina	3.4.21.62

2. 4 complexos com inibição cruzada com o inibidor Ovomucoide (tabela 4.2);

Tabela 4.2: Tabela de inibição cruzada com Ovomucoide

PDB ID	Molécula	Família	EC. Number
1R0R	Subtilisin Carlsberg	**Tipo Subtilisina**	3.4.21.62
1PPF	Human Leukocyte Elastase	Tipo Tripsina	3.4.21.37
1CHO	Alpha-Chymotrypsin A	Tipo Tripsina	3.4.21.1
3SGB	Proteinase B (SGPB)	Tipo Tripsina	-

3. 35 apo enzimas do Tipo Tripsina (incluindo o complexo modelo PDB ID 1PPF:E) com baixa identidade de sequência, conforme mostra a tabela 4.3. Além disso, nós identificamos algumas destas apo enzimas na base de dados BRENDA (*).

Tabela 4.3: Tabela de apo enzimas do Tipo Tripsina

PDB ID	Cadeia	Molécula	Organismo	EC. Number	Ident.
1PPF	E	Human L. Elastase	Homo sapiens	3.4.21.37	
1AUT	C	Activated Protein C	Homo sapiens	3.4.21.46	
1BIO	A	Comp. Factor D	Homo sapiens	3.4.21.46	
1DLE	A	Comp. Factor B	Homo sapiens	3.4.21.-	
1EAX	A	Tumorigenicity 14	Homo sapiens	3.4.21.-	
1EKB	B	Enteropeptidase	Bos taurus	3.4.21.9	
1ELV	A	Comp C1S	Homo sapiens	3.4.21.42	
1EQ9	A	Chymotrypsin	Solenopsis invicta	3.4.21.1	*
1FIW	A	β-Acrosin	Ovis aries	3.4.21.-	
1GG6	B	γ chymotrypsin	Bos taurus	3.4.21.1	*
1GJ7	B	Urokinase	Homo sapiens	3.4.21.73	
1GVK	B	Elastase 1	Sus scrofa	3.4.21.-	
1GVZ	A	Kallikrein	Equus caballus	3.4.21.-	
1HAO	H	α-Thrombin	Homo sapiens	3.4.21.5	
1HAP	H	α-Thrombin	Homo sapiens	3.4.21.5	
1HJ9	A	β-Trypsin	Bos taurus	3.4.21.4	*
1HUT	H	α-Thrombin	Homo sapiens	3.4.21.5	
1LO6	A	Kallikrein 6	Homo sapiens	3.4.21.-	
1M9U	A	Earthworm Fibrinolytic	Eisenia foetida	3.4.21.-	
1MD8	A	C1R Comp.	Homo sapiens	3.4.21.41	
1NPM	A	Neuropsin	Mus musculus	3.4.21.-	
1OP0	A	Venom serine proteinase	Deinagkistrodon acutus	3.4.21.-	
1ORF	A	Granzyme A precursor	Homo sapiens	3.4.21.78	
1P57	B	Serine protease hepsin	Homo sapiens	3.4.21.-	*
1PQ7	A	Trypsin	Fusarium oxysporum	3.4.21.4	
1RFN	A	Coagulation Factor IX	Homo sapiens	3.4.21.-	
1RTF	B	Plasminogen Activator	Homo sapiens	3.4.21.68	
1T32	A	Cathepsin G	Homo sapiens	3.4.21.20	
1Z8G	A	Serine protease hepsin	Homo sapiens	3.4.21.-	
2BZ6	H	Blood Coag. Factor VIIA	Homo sapiens	3.4.21.21	
2F91	A	hepatopancreas trypsin	Pontastacus leptodactylus	3.4.21.4	*
2FPZ	A	Tryptase beta-2	Homo sapiens	3.4.21.59	*
2HLC	A	Collagenase	Hypoderma lineatum	3.4.21.-	
2PKA	B	Kallikrein A	Sus scrofa	3.4.21.35	
2QY0	B	Comp. C1r subcomponent	Homo sapiens	3.4.21.41	

4. 9 apo enzimas do Tipo Subtilisina (incluindo o complexo modelo PDB ID 1TEC:E) com baixa identidade de sequência, conforme mostra a tabela 4.4. Além disso, nós identificamos algumas destas apo enzimas na base de dados BRENDA (*).

Tabela 4.4: Tabela de apo enzimas do Tipo Subtilisina

PDB ID	Cadeia	Molécula	Organismo	EC. Number	Ident.
1TEC	E	Thermitase	Thermoactinomyces vulgaris	3.4.21.66	*
1GCI	A	Sutilisin	Bacillus lentus	3.4.21.62	*
1P8J	A	Furin precursor	Mus musculus	3.4.21.75	*
1THM	A	Thermitase	Thermoactinomyces vulgaris	3.4.21.66	*
1V6C	A	alkaline serine protease	Pseudoalteromonas sp.	3.4.21.-	
1WMD	A	protease	Bacillus sp.	3.4.21.-	
1ID4	A	Assemblin protease	Human herpesvirus	3.4.21.97	
2IXT	A	36KDA protease	Bacillus sphaericus	3.4.21.-	
2PWA	A	Proteinase K	Engyodontium album	3.4.21.64	

Após descrição das nossas bases de dados para análise, nós apresentamos detalhadamente, nas próximas seções, todos os passos da metodologia proposta. A figura 4.1 expõe minuciosamente, por meio do fluxograma, todos os passos metodológicos no processo de descoberta de padrões conservados nas interfaces moleculares envolvendo complexos de inibição cruzada e interfaces projetadas em apo enzimas. Em (a), (b), (c) , (d) e (e), as estruturas da esquerda referem-se a Tipo Subtilisina e as estruturas da direita referem-se a Tipo Tripsina.

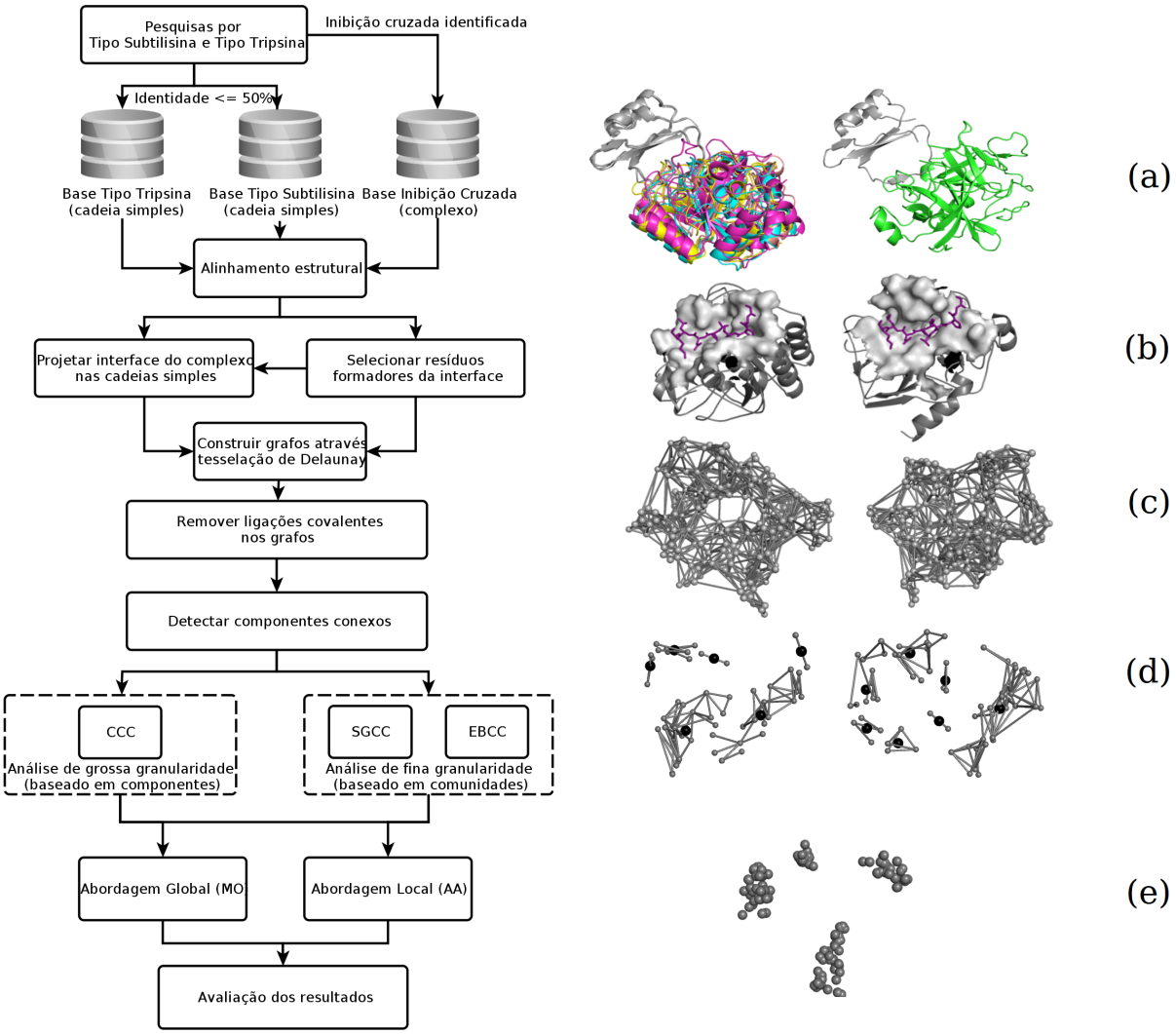


Figura 4.1: Diagrama de fluxo da metodologia

4.2 Preparação dos dados

Após definir o nosso escopo de trabalho, ou seja, os dados a serem analisados, nós iremos detalhar nesta seção os passos da preparação das moléculas a serem analisadas.

4.2.1 Normalização e alinhamento das estruturas

Inicialmente, as estruturas foram submetidas a um processo de normalização ou parametrização usando o PDBest (*PDB Enhanced Structure Toolkit*) [Pires et al. (2007)]. O PDBest é uma ferramenta desenvolvida pelo nosso grupo que tem o objetivo de acessar os arquivos PDBs e corrigir possíveis inconsistências e anomalias contidas nas estruturas. Dentre as anomalias observadas, estão a numeração descontínua dos resíduos e átomos e a duplicação de átomos.

Formado por um conjunto de *scripts* Perl que aplica um conjunto de regras, definidas pelo usuário, o PDBest modifica os arquivos PDBs e fornece uma versão mais limpa e filtrada. As estruturas das duas bases de apo enzimas foram submetidas, sendo que cada arquivo foi filtrado, separando as cadeias de interesse (quando dímeros) e re-enumerando os átomos e resíduos. As estruturas em complexos não foram alteradas porque nós precisávamos mantê-las íntegras, pois os resíduos que compõem a interface não foram calculados por nós. A alteração da numeração destes resíduos implicaria em erros nas análises. Na seção 4.2.2 explicaremos como e onde os resíduos que compõem as interfaces dos complexos foram obtidos.

As apo enzimas também sofreram um processo de solvatação usando um módulo do PDBest que carrega o programa Gromacs [Erik et al. (2001)]. A solvatação foi necessária para que, em passos futuros, a falta do inibidor não comprometesse o cálculo dos contatos.

Após a parametrização das estruturas, um passo muito importante e essencial de nossa metodologia é o alinhamento estrutural. O alinhamento das estruturas é fundamental para que as interfaces estejam em posições tridimensionais similares para que haja a comparação entre as regiões hidrofóbicas.

Os complexos com inibição cruzada foram alinhados pela cadeia do inibidor. A figura 4.2(a), mostra um exemplo de alinhamento de 5 complexos, usando a cadeia do inibidor como referência (cor cinza), sendo considerada uma boa sobreposição. Enquanto isso, é possível observar que as cadeias das enzimas não foram bem sobrepostas por se tratarem de estruturas dissimilares.

As apo enzimas foram alinhadas pelas próprias cadeias. A figura 4.2(b) mostra um exemplo de sobreposição de cadeias simples de uma família de protease, sendo considerada uma boa sobreposição. Neste caso, nós sobrepomos várias cadeias simples e um complexos com inibidor, que pode ser visto na cor cinza. Em ambos os casos de alinhamento, as várias cadeias das enzimas possuem cores diferentes, a fim de identificá-las.

Os alinhamentos foram realizados usando o programa Multiprot [Shatsky et al. (2004)][1]. O Multiprot realiza alinhamento estrutural múltiplo baseado em critérios geométricos. A visualização das estruturas foi feita através do programa PyMOL[2].

[1] O Multiprot está disponível na versão Web (http://bioinfo3d.cs.tau.ac.il/MultiProt/) ou pode ser usado localmente

[2] O PyMOL é uma ferramenta gratuita e pode ser livremente acessada em:http://www.pymol.org/

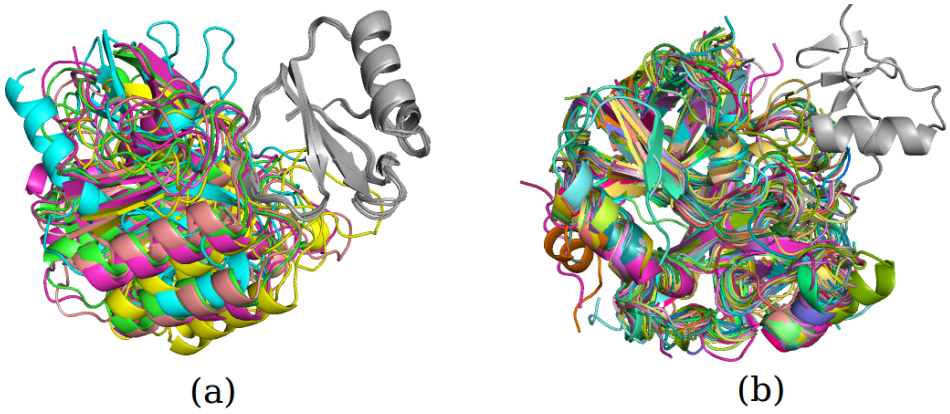
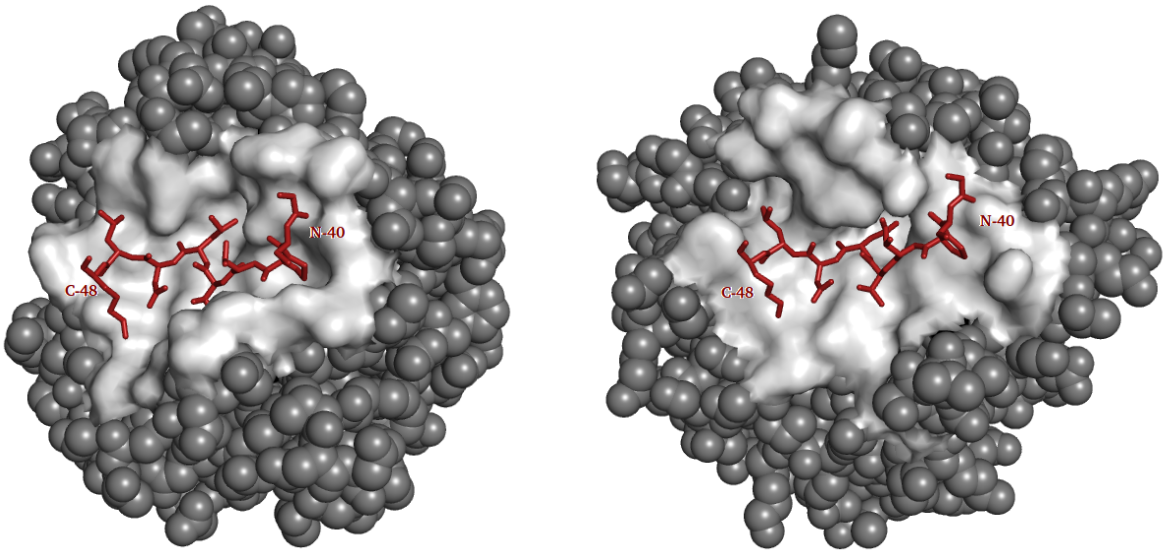


Figura 4.2: Exemplos de alinhamento das estruturas. Em (a), temos exemplo de alinhamento pelo cadeia do inibidor e em (b), temos exemplos de alinhamento pela cadeia das apo enzimas (sem complexos)

4.2.2 Interface Molecular

A interface molecular é uma importante região de contato e interação entre duas moléculas. Ela está localizada no bolsão catalítico e os resíduos que fazem algum tipo de contato atômico com o ligante são ditos como resíduos da interface ou simplesmente IFR (do inglês *Interface Forming Residues*). A figura 4.3 mostra a interface de duas diferentes proteases em contato com o mesmo inibidor (Eglina C). Os resíduos que compõem a interface estão representados como superfície lisa (cor cinza claro). Quanto ao inibidor, nós restringimos apenas a visualização da alça que ancora no bolsão catalítico. A alça (cor vermelha) é composta de 9 resíduos (GSPVTLDLR), exibida de N- para C-terminal.



(a) (b)

Figura 4.3: Exemplos de interface molecular. Em (a), Tipo Subtilisina PDB ID 1TEC:E e em (b), Tipo Tripsina PDB ID 1ACB:E.

As IFRs podem ser determinadas basicamente por três métodos diferentes. O primeiro define a interface simplesmente pelo uso de uma distância de corte (*cut-off distance*) entre os resíduos das moléculas interagindo [Chothia e Janin (1975); Conte et al. (1999)]. O segundo método calcula as interações baseado nas diferenças da área acessível ao solvente quando os monômeros são separados [Janin et al. (1990); Chakrabarti e Janin (2002b)] e o último método define interface através da geometria computacional, usando diagramas de Voronoi e teoria de *Alpha Shape* [Richards (1974); Pontius et al. (1996)]. Nos três métodos apresentados, usados para calcular a interface molecular, é necessário a presença do ligante.

Neste trabalho, as IFRs dos complexos foram calculadas com base no método de acessibilidade ao solvente, conhecido como ASA do inglês (*Accessibility Solvent Area*). O método ASA é mais usado e portanto mais consolidado, além disso, existem várias ferramentas que disponibilizam a IFR de todas as estruturas do PDB. Nós usamos a interface calculada pela plataforma STING Millennium Suite (SMS) [Neshich et al. (2003)].

Entretanto, as IFRs das apo enzimas não podem ser determinadas por nenhum dos três métodos, uma vez que não há a presença do ligante. Neste caso, nós propomos a predição da IFR, através do alinhamento das estruturas, usando um complexo enzima/inibidor com interface conhecida como modelo. Basicamente, os resíduos das apo enzimas que se alinharam aos resíduos da interface do complexo foram considerados como pertencentes à interface das apo enzimas. Aqui, nós desenvolvemos algoritmos capazes de capturar e armazenar estes resíduos. Este processo foi executado para ambas as famílias (Tipo Tripsina e Tipo Subtilisina) de enzimas de cadeias simples contidas em nossas bases de dados de análises. Cada uma das famílias usou como modelo um complexo com enzima da própria família, garantindo assim um bom alinhamento e minimizando possíveis erros na predição ou projeção dos resíduos.

4.3 Construção dos grafos

Apos normalizar todas as estruturas, recuperar os resíduos da IFR na base de dados da plataforma STING e projetar as IFRs para todas as apo enzimas, nós podemos prosseguir na extração dos grafos representando as interações nas IFRs.

Interações proteína-proteína, como enzima/inibidor, são grafos de interações covalentes e não covalentes. Estas interações podem ser analisadas a nível de resíduo ou átomo. Neste trabalho, os vértices de nossos grafos são os átomos e as interações presumidas entre eles são as arestas. A premissa mais importante neste trabalho é que a arquitetura de redes atômicas pode revelar princípios importantes de reconhecimento molecular [Soundararajan et al. (2010b)].

De acordo com da Silveira et al. (2009), existem duas principais abordagens para prospectar contatos (interações presumidas). As abordagens dependente de corte (DC) e independente de corte (IC). Em DC um contato é definido entre um dado par de átomos i, j

se a distância Euclidiana entre eles foi menor ou igual a uma distância de corte arbitrária. Em IC não há associação a nenhum valor de distância para determinar um contato. Nós escolhemos a abordagem IC porque ela é livre de detectar falsos contatos que podem estar oclusos. Embora nesse estudo, constatou-se que, a nível de resíduos, a abordagem DC revelou ser uma técnica mais simples, mais completa e confiável do que IC, nós escolhemos usar uma metodologia independente de corte porque, em nível atômico, não foi possível encontrar um valor de corte confiável.

Este paradigma usa algoritmos clássicos da geometria computacional como diagrama de Voronoi [Poupon (2004)] e seu problema dual; a Tesselação de Delaunay [Angelov et al. (2002) e Dupuis et al. (2005)].

4.3.1 Diagrama de Voronoi e Tesselação de Delaunay

O Diagrama de Voronoi (DV) (em homenagem ao matemático russo Georgy Voronoi) é uma técnica usada na área da matemática que estuda a decomposição de um espaço métrico em regiões de acordo com a distância a determinados pontos. Possui aplicação em inúmeras áreas do conhecimento com vários algoritmos disponíveis.

DV é uma estrutura geométrica que representa informações de proximidade sobre um conjunto de pontos ou objetos. Em um conjunto $S = \{p1, p2, ..., pn\}$ de pontos em um plano, chamados sítios, é uma divisão do plano em n regiões convexas, uma por sítio. Cada região Voronoi Vi contém todos os pontos mais próximos a pi do que qualquer outro sítio. As linhas retas dual do DV, obtidas pela adição de um segmento de reta entre cada par de sítio de S cujas regiões Voronoi partilham uma aresta, é chamada de Tesselação ou Triangulação de Delaunay (TD).

A figura 4.4 mostra as regiões, representadas por polígonos ou células de Voronoi e as ligações entre os sítios representando TD. Cada célula é definida por suas fronteiras indicadas pelas linhas pontilhadas e cada aresta é uma reta perpendicular que liga dois átomos

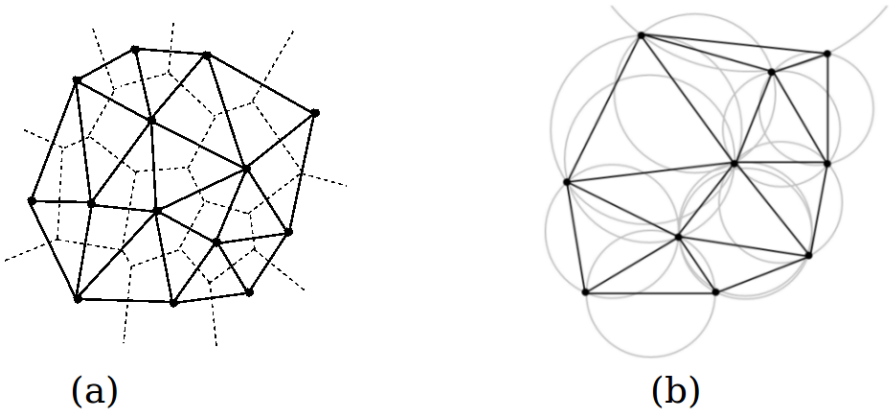


Figura 4.4: Em (a), temos DV (linhas pontilhadas) e TD (linhas sólidas) e em (b) temos os círculos que sempre delimita três sítios

Em nosso contexto, cada célula contém um átomo, e cada ponto no interior da célula é mais próximo do centro do seu átomo do que qualquer outro átomo. Três arestas encontram-se em um vértice e três sítios compor-se-iam em uma triangulação de Delaunay se e somente se o círculo circunscrito a eles não contivesse nenhum outro sítio. Aplicando esse método a todos os sítios irá decompor o espaço ocupado por eles em triângulos justapostos, fazendo emergir a notável propriedade de que somente os vizinhos mais próximos e não oclusos por outro estarão conectados.

A figura 4.5 mostra a técnica de TD aplicada, como exemplo, aos átomos que compõem a interface molecular (cor laranja) de duas diferentes proteases. O mapeamento da rede de átomos vizinhos, seguindo os critérios geométricos, foi naturalmente considerado como grafo.

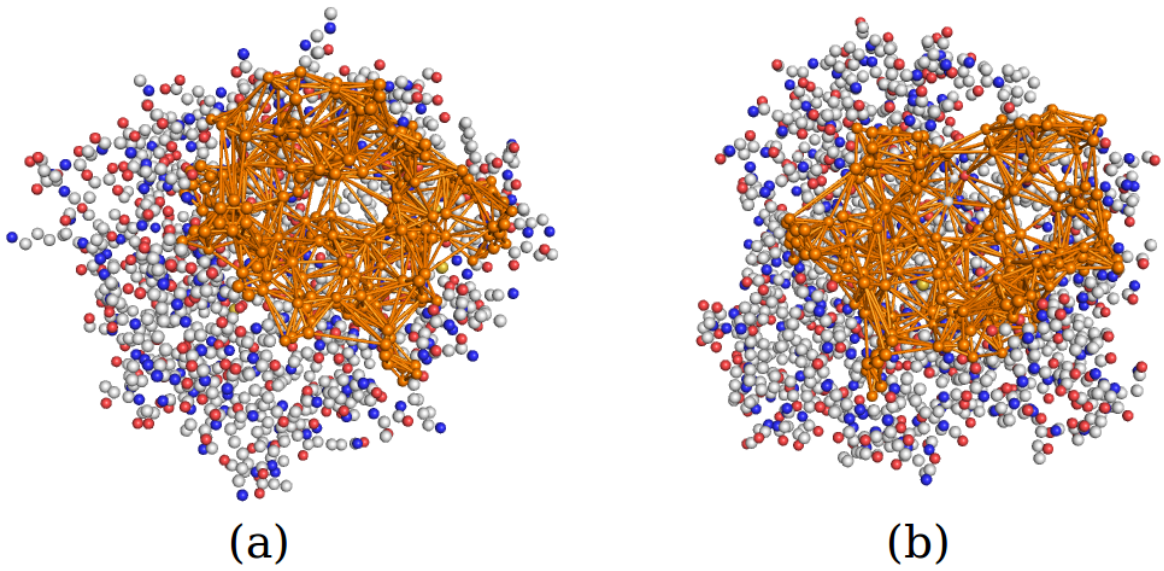


Figura 4.5: Exemplos de Tesselação de Delaunay na interface molecular de enzimas. Em (a), PDB ID 1TEC:E e em (b), PDB ID 1ACB:E

Neste trabalho, nosso interesse é somente por interações não covalentes. Assim, todas as interação atômicas (arestas) intra-resíduos e covalentes foram removidas. Uma vez que nós temos inferência geométrica de interações não oclusas, dadas seguindo critérios da técnica de TD, nós classificamos estas interações em hidrofóbicas e polares, baseadas em estudos de interações atômicas feitas por Sobolev et al. (1999).

Alguns estudos sobre as contribuições positivas de mudanças entrópicas no reconhecimento e ligação de proteases e inibidores foram discutidos no capítulo 1. Neste sentido, nós concentramos nossas análises somente nas interações entre átomos que estão em condições hidrofóbicas (apolares). Considerando que todos os resíduos têm porções apolares (átomos), nós construímos uma tabela que identifica estes átomos em cada resíduo. A tabela 4.5 mostra os resíduos (exceto a glicina por não ter cadeia lateral) e quais dos seus átomos são considerados estar em condições hidrofóbicas.

Tabela 4.5: Tabela de classificação de átomos em condições hidrofóbicas

Resíduo	Átomo	Resíduo	Átomo	Resíduo	Átomo	Resíduo	Átomo	Resíduo	Átomo
ALA	CB	GLU	CG	LYS	CB	PHE	CZ	TRP	CZ3
ARG	CB	GLU	CD	LYS	CG	PRO	CB	TRP	CH2
ARG	CG	HIS	CB	LYS	CD	PRO	CG	TYR	CB
ARG	CD	HIS	CG	LYS	CE	PRO	CD	TYR	CG
ARG	CZ	HIS	CD	MET	CB	SER	CB	TYR	CD1
ASN	CB	HIS	CE1	MET	CG	THR	CB	TYR	CD
ASN	CG	ILE	CB	MET	SD	THR	CG2	TYR	CE1
ASP	CB	ILE	CG1	MET	CE	TRP	CB	TYR	CE2
ASP	CG	ILE	CG2	PHE	CB	TRP	CG	TYR	CZ
CYS	CB	ILE	CD1	PHE	CG	TRP	CD1	VAL	CB
GLN	CB	LEU	CB	PHE	CD1	TRP	CD	VAL	CG1
GLN	CG	LEU	CG	PHE	CD	TRP	CE2	VAL	CG2
GLN	CD	LEU	CD1	PHE	CE1	TRP	CE3		
GLU	CB	LEU	CD	PHE	CE2	TRP	CZ2		

Como base na tabela de classificação os átomos, nós restringimos nossas análises para este tipo de interação. A análise pode ser estendida para lidar com áreas polares.

Deste modo, as arestas que não representavam a ligação entre dois átomos hidrofóbicos foram removidas. As redes de interações entre os átomos hidrofóbicos revelou grafos de componentes conexos que representam regiões hidrofóbicas que pode contribuir para o reconhecimento molecular.

A figura 4.6 mostra os mesmos exemplos da figura 4.5, porém, aqui são exibidos apenas as ligações de nosso interesse, ou seja, as regiões hidrofóbicas. Podemos notar que as ligação remanescentes formam grafos representando os componentes conexos hidrofóbicos na interface de contato (cor laranja).

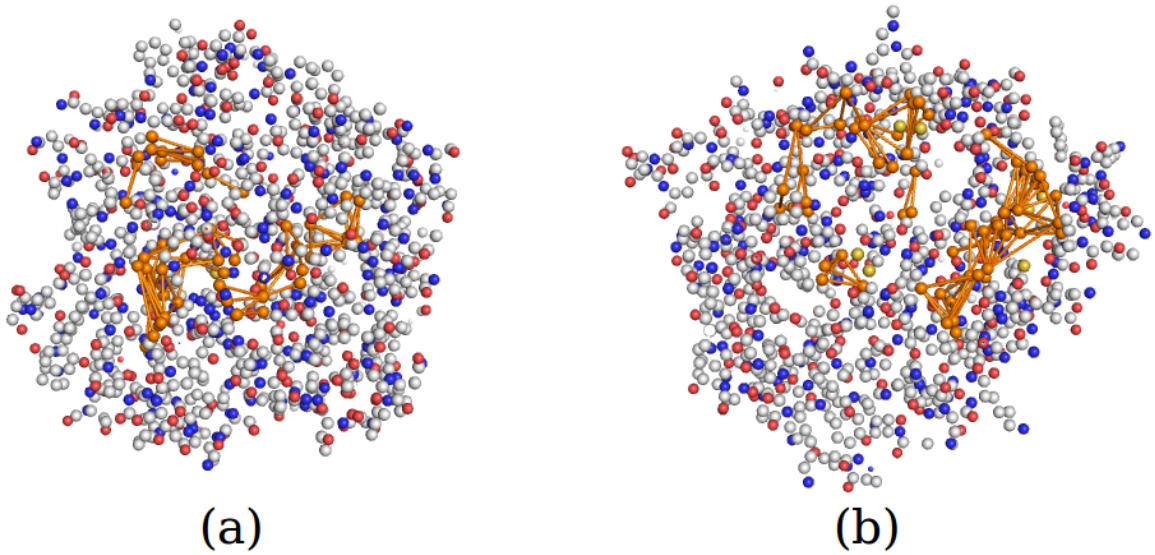


Figura 4.6: Exemplos de componentes conexos na interface molecular de enzimas. Em (a), PDB ID 1TEC:E e em (b), PDB ID 1ACB:E

4.4 Modelagem do problema

Os componentes conexos são subgrafos ou regiões onde os átomos estão todos conectados. Para identificar os componentes conexos nós usamos, com base na teoria de grafos, a busca em profundidade. Os componentes conexos são naturalmente representações de regiões hidrofóbicas ou *Patches*. Porém, as regiões hidrofóbicas podem ocorrer de formas e volumes muito diferentes. Neste sentido, nosso modelo considera dois níveis de abstração para representá-los, ambos baseados no conceito de centroides geométricos (HP-centroides). Os centroides geométricos são calculados sobre a distância média (em Å) das coordenadas x, y e z de cada átomo do componente conectado[3].

O primeiro, nós chamamos de análise de granularidade grossa e consiste em computar um centroide para cada componente conectado. O segundo, nós chamamos de análise de granularidade fina e fragmenta o componente conectado original em subgrafos densos, ou comunidades. Uma comunidade é um subgrafo onde os vértices (átomos) são muito mais conectados entre eles do que com o resto do componente. Nesta abordagem, os HP-centroides são computados usando algoritmos de detecção de comunidades em grafos. Nós comparamos dois tipos distintos de algoritmos para detectar comunidades que serão explicados das próximas seções

Na conclusão da modelagem de nosso problema, nosso método é baseado no cálculo das regiões hidrofóbicas e sua abstração por meio de centroides geométricos (HP-centroides) que podem representar as regiões hidrofóbicas como um todo (granularidade grossa) ou comunidades destes regiões hidrofóbicas (granularidade fina). Considerando o HP-centroide de um conjunto de complexos enzima-inibidor, nós propomos algoritmos para agrupar os centroides e identificar conservações sobre todos eles. Nós descrevemos os algoritmos na próxima seção e após explicá-los nós descrevemos uma métrica de avaliação para os grupos obtidos.

4.5 Algoritmos

Nesta seção, nós descrevemos com mais detalhes as duas abordagens utilizadas para abstrair as regiões hidrofóbicas: granularidades grossa e fina. Nós descrevemos os paradigmas para detecção de comunidades usado na decomposição de granularidade fina das regiões hidrofóbicas. Finalmente, nós descrevemos os algoritmos propostos para agrupar os HP-centroides: uma tentativa de equiparar os HP-centroides usando abordagem global e outra abordagem local que consiste em agrupar os centroides de maneira aglomerativa.

[3]Os termos componente conexo e componente conectado são sinônimos

4.5.1 Abordagem de granularidade grossa

4.5.1.1 Abordagem CCC

A análise baseada em CCC (centróide de componente conectado) é usada para representar a abordagem de granularidade grossa. Neste caso, os HP-centroides são calculados usando todos os átomos que pertencem ao componente conectado. A figura 4.7 mostra alguns exemplos de HP-centroides dentro de um componente conectado. Os átomos (esfera) e as suas arestas são mostrados na cor cinza. Os HP-centroides estão exibidos em esferas na cor preta e estão posicionados exatamente do centro dos componentes conectados.

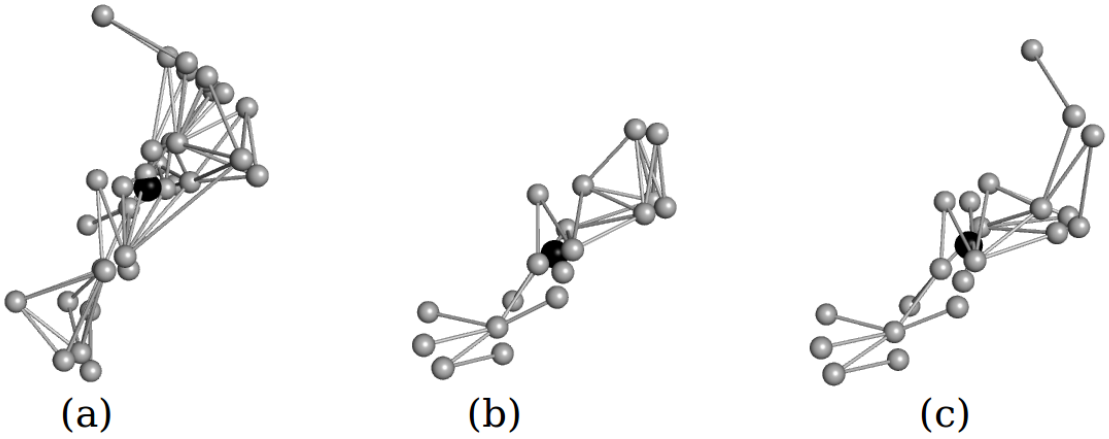


Figura 4.7: Exemplos de HP-centroides para CCC.

4.5.2 Abordagem de granularidade fina

Para a abordagem de granularidade fina nós testamos dois diferentes algoritmos. O primeiro é o *edge.betweenness community* que usa conceito de modularidade em grafos e o segundo é o *spin.glass community* que usa conceito de simulações para encontrar regiões mais densas nos grafos. Para viabilizar o uso dos algoritmos, nós desenvolvemos vários *scripts* na linguagem Perl. Os algoritmos usam bibliotecas do programa R [R Development Core Team (2008)]. Estas bibliotecas são disponíveis gratuitamente e estão contidas no pacote *igraph*.

4.5.2.1 Abordagem EBCC

A análise baseada em EBCC (centróide de comunidade baseado no método *Edge Betweeness*) [Newman e Girvan (2004)] é um dos algoritmos usados na abordagem de granularidade fina. Ela é uma abordagem divisiva onde as arestas mais centralizadoras são quebradas uma após a outra até que a modularidade do grafo seja maximizada. O conceito de modularidade está diretamente relacionado com a forma de encontrar módulos, ou seja, regiões mais densas ou aglomeradas. O algoritmo calcula a aresta central

através do *edge betweenness*. O *edge betweenness* é determinado em um grafo contabilizando o número de caminhos mínimos que atravessa a aresta. Quanto maior for o *edge betweenness*, mais usada é a aresta ou mais central. Em outras palavras, isso indica que não existem arestas redundantes para cruzar entre diferentes comunidades e que a aresta conecta duas diferentes comunidades.

A figura 4.8 mostra alguns exemplos de comunidades encontradas nos componentes conectados, usando o algoritmo de *edge.betweenness community*. Os átomos (esfera) e as suas arestas são mostrados na cor cinza. Os HP-centroides estão exibidos em esferas na cor preta. Neste caso, cada componente conectado foi subdividido em duas regiões consideradas com uma densidade maior de átomos.

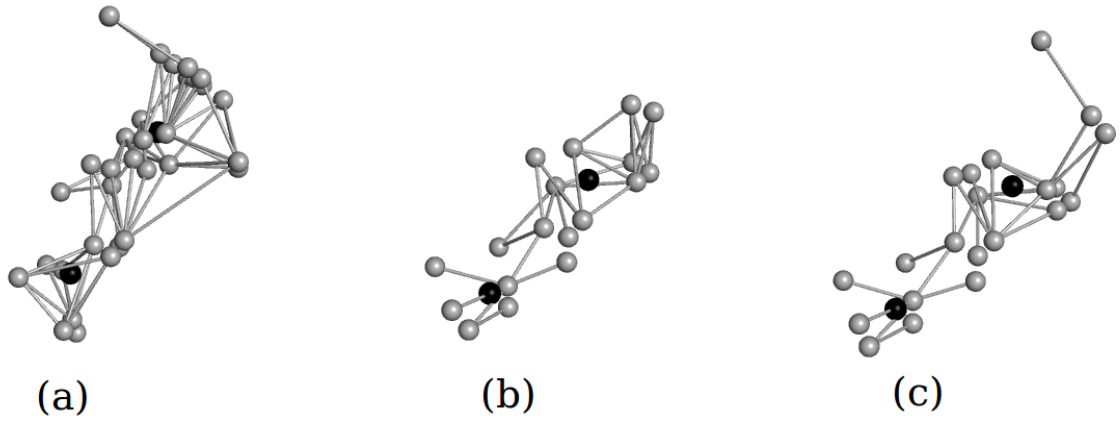


Figura 4.8: Exemplos de HP-centroides para EBCC

4.5.2.2 Abordagem SGCC

A análise baseada em SGCC (centróide de comunidade baseado no método *spin glass*) [Reichardt e Bornholdt (2006)] é usada para tentar encontrar comunidades nos grafos via um modelo de *spin-glass* e simulações de *annealing*. Isto é, ele usa simulações de *annealing* para maximizar a modularidade do grafo. A modularidade de uma possível divisão de um grafo em comunidades é definido como a fração das arestas que caem dentro da comunidade dada menos a fração esperada, caso as arestas fossem distribuídas aleatoriamente. Comumente, a distribuição aleatória das arestas é feita de forma a preservar o grau de cada vértice.

A figura 4.9 mostra alguns exemplos de comunidades encontradas nos componentes conectados, usando o algoritmo *spin.glass community*. Os átomos (esfera) e as suas arestas são mostrados na cor cinza. Os HP-centroides estão exibidos em esferas na cor preta. Igualmente mostrado na análise EBCC, cada componente conectado foi subdividido em duas regiões consideradas com uma densidade maior de átomos.

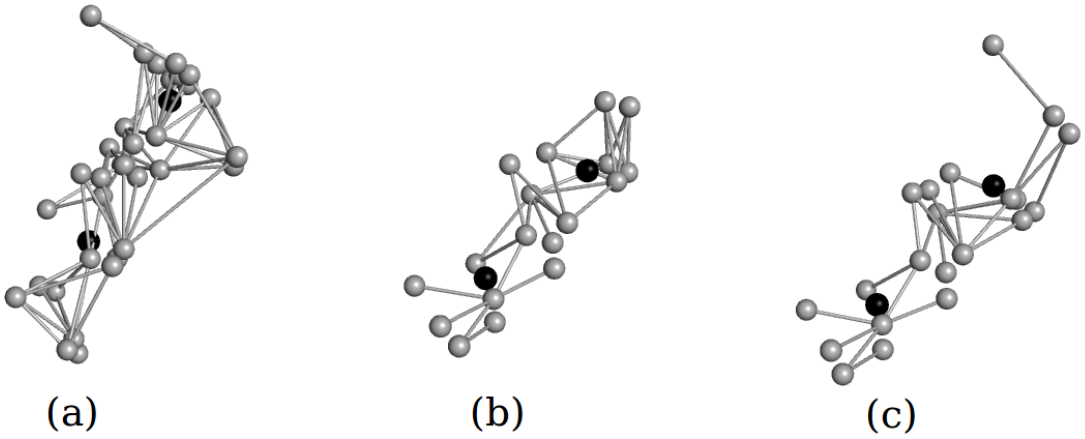


Figura 4.9: Exemplos de HP-centroides para SGCC

4.5.3 Comparação dos HP-centroides

Após calcular os HP-centroides para todas as entidades de nossas bases de dados, o passo final foi construir algoritmos que têm por objetivo comparar os HP-centroides. Para isso, nós usamos um método global e outro local. O método global busca casar os HP-centroides mais próximos entre si minimizando o somatório das distâncias entre os HP-centroides casados globalmente, ou seja, considerando a totalidade dos HP-centroides. O método local busca casar os HP-centroides que estejam próximos no espaço, sem considerar HP-centroides distantes e seus casamentos.

4.5.3.1 Comparação usando MO

A comparação dos HP-centroides baseado em MO (Modelo de Otimização) usa programação linear inteira, sendo o modelo proposto baseado diretamente do problema do transporte. Os pontos são equiparados globalmente minimizando a diferença entre o tamanho das arestas entre todos os possíveis pares de vértices. A programação linear é uma técnica para otimização de funções objetivas lineares dada uma lista de requisitos representados como equações ou inequações lineares. A interpretação geométrica destas restrições é um politopo no qual é chamado de região viável. Assim, nós apresentamos como este tipo de modelo matemático pode ser usado para representar o problema de casamento de grafos que representam os nossos *patches* hidrofóbicos. Neste grafo $G = (V, E)$, onde V são os vértices representando os centroides e E é um conjunto de arestas contendo todos os centroides. Então, nós queremos equiparar as arestas no grafo a fim de:

$$min \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} x_{ij}$$

onde n é o número de centroides da primeira enzima e m é o número de centroides da segunda enzima a ser combinada. x_{ij} são variáveis inteiras binárias que codificam possíveis

casamentos entre aresta i do primeiro grafo e j do segundo grafo e d_{ij} é a diferença em tamanho de arestas i e j. Nós otimizamos a equação exposta para seguir as restrições:

$$\forall i \sum_{j=1}^{m} x_{ij} = 1(1)$$

$$\forall j \sum_{i=1}^{n} x_{ij} \leq 1(2)$$

o que significa que (1) cada aresta do primeiro grafo deve ser combinada com uma aresta do segundo grafo e (2) cada aresta no segundo grafo pode ser combinada para até uma aresta do primeiro grafo. Este modelo foi construído usando um *script* Perl e resolvido pelo programa Glpsol.

4.5.3.2 Comparação usando AA

A comparação dos HP-centroides baseado em AA tem por objetivo equiparar os pontos que estão mais próximos. Ele equipara os HP-centroides mais próximos através de um processo iterativo aglomerativo de baixo para cima. Este caso, existe uma importante decisão sobre o ponto de parada do processo a fim de garantir que nós tenhamos grupos de alta qualidade. Abaixo nos descrevemos minuciosamente os passos do algoritmo, bem como, a avaliação de sua complexidade.

Entrada:

Um conjunto de HP-centroides (por exemplo, um conjunto de coordenadas tridimensionais);

O número de grupos.

Saída:

Um mapeamento de HP-centroides para grupos.

Cada passo do Algoritmo 1 é descrito abaixo:

- O primeiro passo é calcular a distância entre os pares de HP-centroides (Linha 1).
- Próximo, os pares de HP-centroides são ordenados (crescente) pela suas distâncias (Linha 2).
- O número atual de grupos é definido como o número de HP-centroides, e cada HP-centroide é atribuído a um grupo separado (Linhas 3-5).
- Após isto, enquanto o número desejado de grupos não é atingido, os grupos são unidos de acordo com os pares de HP-centroides mais próximos (Linhas 6-7).
- Se o par de HP-centroide selecionado pertence a um grupo diferente, os dois grupos são unidos e o mapeamento é propriamente modificado e o número atual dos grupos é reduzido (Linhas 8-10).

- Finalmente, o mapeamento dos HP-centroides para os grupos é recuperado (Linha 11).

Algorithm 1 Agrupamento Aglomerativo

Input: $CentroidSet, NumberOfClusters$
Output: $MapClusters$

```
1: Distances ← calculateDistances(CentroidSet)
2: SortedPairs ← sortAscendDist(CentroidSet, Distances)
3: CurrentNumClusters ← size(CentroidSet)
4: for all centroid i ∈ (CentroidSet) do
5:     MapClusters[i] ← i
6: while CurrentNumClusters < NumberOfClusters do
7:     (A, B) ← getNextPair(SortedPairs)
8:     if MapClusters[A] ≠ MapClusters[B] then
9:         mergeClusters(MapClusters, A, B)
10:        CurrentNumClusters ← CurrentNumClusters - 1
11: return MapClusters
```

A complexidade assintótica de tempo do algoritmo é $O(n^2)$, onde n é o número de HP-centroides, e é devido aos pares de cálculo da distância entre os HP-centroides.

4.6 Avaliação

A fim de executar uma avaliação qualitativa dos grupos formados pela equiparação dos HP-centroides, nós propomos uma métrica baseada no conceito de cobertura penalizada. A métrica penaliza os grupos que têm diferentes HP-centroides da mesma enzima sendo agrupados juntos (HP-centroides redundantes). Ela está descrita abaixo e recebeu o nome de PRM, do inglês (*Penalized Recall Metric*):

$$PRM = \frac{\mathbb{C}\binom{D}{2}}{\mathbb{C}\binom{P}{2}} - \frac{\mathbb{C}\binom{E}{2}}{\mathbb{C}\binom{P}{2}} \tag{4.1}$$

onde $\mathbb{C}\binom{D}{2}$ é o número de pares de HP-centroides de diferentes enzimas no mesmo grupo, $\mathbb{C}\binom{E}{2}$ é o número de pares de HP-centroides da mesma enzima no mesmo grupo e $\mathbb{C}\binom{P}{2}$ é o total de número de pares de HP-centroides.

A métrica produz valores em um intervalo [-1; +1], onde, -1 é o pior caso, com mínima cobertura e máxima redundância. Ela resultará em +1 quando nós temos máxima cobertura e mínima redundância. Quando nós temos valores próximos para cobertura e redundância, a métrica aproxima de 0.

A média da métrica PRM dos grupos foi usada para avaliar as três diferentes abordagens (CCC, EBCC e SGCC). Entretanto, ela não pode ser usada para comparar os

métodos MO e AA. Em MO, os grupos são formados com a variabilidade total por definição, em outras palavras, não haverá HP-centroides da mesma enzima no mesmo grupo.

Neste caso, nós usamos as distâncias médias intra- e inter-grupo. A princípio, um grupo de alta qualidade tem baixa distância intra-grupo e alta distância inter-grupos. Isto acontece porque em um agrupamento ideal, elementos similares devem ser agrupados juntos e elementos dissimilares devem ser separados em grupos diferentes.

Para concluir, nós comparamos as abordagens propostas à luz da cobertura penalizada (próximo de 1, melhor) e as distâncias médias intra- e inter-grupo (melhor ter baixa distância intra-grupo e alta distância inter-grupo).

Capítulo 5

Resultados e discussão

Neste capítulo, nós apresentamos e discutimos os resultados da aplicação da metodologia HydroPaCe na análise dos dois casos de inibição cruzada (Ovomucoide e Eglina C) e na predição das IFRs em apo enzimas, e busca de HP-centroides conservados nestas IFRs, de duas famílias de serino protease (Tipo Tripsina e Tipo Subtilisina). O capítulo está dividido em três principais seções: as análises das inibições cruzadas, a análise da predição das IFRs e a discussão dos resultados.

A primeira seção apresenta, individualmente, os resultados da análise para as inibições cruzadas com os inibidores Eglina C e Ovomucoide. A conservação dos HP-centroides nas três abordagens diferentes (CCC, SGCC e EBCC), bem como a avaliação individual de cada grupo de HP-centroides identificado é exposto. Além disso, a comparação das médias das distâncias intra-grupos e os valores da PRM são usados para avaliar as abordagens.

A segunda seção apresenta, individualmente, os resultados da análise para as IFRs projetadas. Neste caso, nós mostramos que usando complexos modelos, da inibição cruzada, é possível identificar os HP-centroides na grande maioria das apo enzimas analisadas. A conservação dos HP-centroides nas três abordagens diferentes são apresentados e a avaliação individual, através das distâncias intra grupos e os valores da PRM são expostos. Além disso, é possível comparar as três abordagens por meio das médias das distâncias intra-grupos e os valores da PRM.

Finalmente, na terceira e última seção, nós discutimos a contribuição da metodologia HydroPaCe na busca por padrões conservados que podem explicar porque a inibição cruzada ocorre e como ela pode ser predita em enzimas de IFRs desconhecidas.

5.1 Análise das inibições cruzadas

Em trabalhos anteriores [Melo et al. (2007); Ribeiro et al. (2010)], colegas do grupo de pesquisa estudaram o uso de sequências, estruturas e mapas de contatos para analisar contatos em inibição de enzimas. Embora haja similaridade entre alguns resíduos de sítios específicos em inibidores e em enzimas, pela análise de conservação somente a nível de

sequência, ainda que mutações conservativas são consideradas, é muito difícil encontrar padrões em inibição cruzada.

No contexto deste trabalho, a expressão inibição cruzada denota a falta de especificidade do inibidor. Sendo este, um problema muito complexo porque podem existir similaridades de sequências muito baixas e estruturas tridimensionais completamente diferentes. Isto torna impossível considerar os clássicos alinhamentos de sequência e estrutura. A presente metodologia HydroPaCe tem o objetivo de detectar um padrão extenso para explicar a inibição cruzada. Ela é flexível para tolerar modificações a nível de resíduo, abstraindo-se as mutações esperadas e usando uma representação geométrica mais robusta das propriedades dos resíduos que formam a IFR.

Nós próximos passos, nós discutimos os resultados dos algoritmos de equiparação global e local nas perspectivas das granularidades grossa e fina para as inibições cruzadas envolvendo os inibidores Eglina C e Ovomucoide.

5.1.1 Inibidor Eglina C

O inibidor Eglina C é uma pequena proteína monomérica (70 resíduos) que pertence a família de inibidores de serino protease do tipo *Potato Chymotrypsin Inhibitor I* e que ocorre naturalmente em sanguessugas *Hirudo medicinalis*. Funcionalmente, Eglina C pode inibir enzimas de mais de uma família de enzimas proteolíticas com estruturas não homólogas [Hyberts et al. (1992)]. A figura 5.1 mostra o inibidor Eglina C (cor azul) complexado com duas enzimas de famílias diferentes (cor rosa). É possível ainda, visualizar a superfície da interface de interação ou IFR em cor amarela.

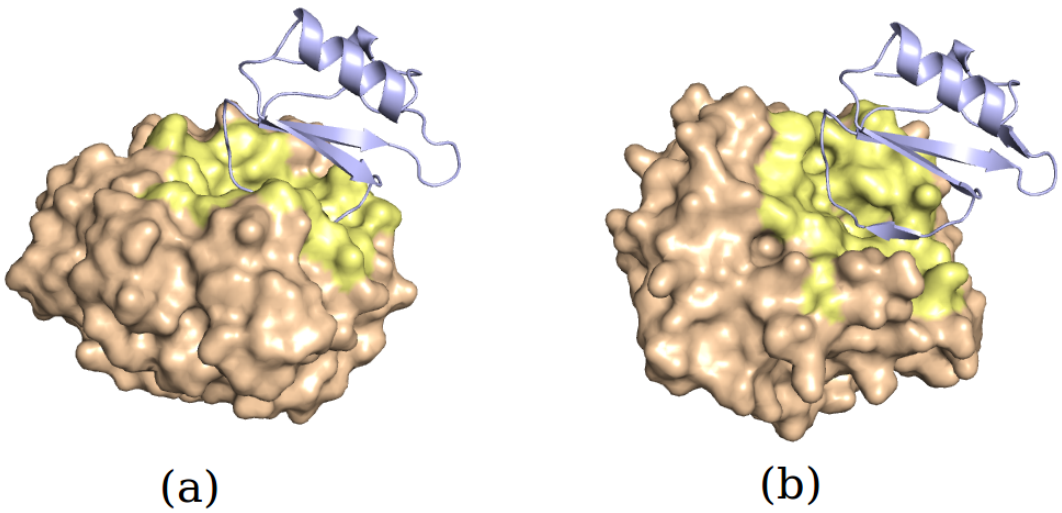


Figura 5.1: Exemplos de inibição cruzada com o inibidor Eglina C. Em (a), Tipo Subtilisina PDB ID 1TEC e em (b), Tipo Tripsina PDB ID 1ACB.

Na base de dados BRENDA, nós encontramos 12 diferentes *EC. numbers* identificando complexos onde ocorre inibição por Eglina C. Nesta seção, nós apresentamos a análise de 5

complexos diferentes com estruturas tridimensionais conhecidas, sendo 4 na família Tipo Subtilisina e 1 na família Tipo Tripsina.

Como explicado anteriormente, nossa metodologia usa diferentes abordagens para encontrar os HP-centroides. Os HP-centroides podem ser agrupados por dois diferentes métodos (MO e AA). O método MO não tem parâmetros e agrupa, obrigatoriamente, todos os HP-centroides. O método AA é guiado pelo número de grupos, fornecido como parâmetro de entrada no algoritmo de agrupamento. Neste caso, é usado o critério de proximidade para agrupar os HP-centroides. Deste modo, a análise usando o método AA pode ser estendida com a finalidade de escolha e avaliação dos grupos.

Granularidade grossa: CCC

Na análise de granularidade grossa (CCC) os resultados são exibidos na figura 5.2. O gráfico da esquerda mostra a distribuição do número de grupo em função da distância média intra-grupo e os valores da PRM média para cada grupo. O número de grupos possíveis é igual ao total de HP-centroides. Ou seja, o método é baseado em agrupamento aglomerativo que inicia considerando cada HP-centroide como um grupo. Nós observamos que com **12** grupos a distância média intra-grupo é estabilizada (gráfico da esquerda) e a PRM média é a maior. Com esta configuração, nós obtemos **5** com alta qualidade ($>$ 50%), dos **12** grupos considerados, de acordo com a métrica PRM (gráfico da direita). Este conjunto de HP-centroides apresenta uma alta cobertura. Por exemplo, eles estão representes em quase todos os complexos com inibição cruzada e além disso, existe somente um caso onde HP-centroides estão no mesmo grupo e que são do mesmo complexo. É importante observar que os grupos com valores de PRM menor que zero não são exibidos.

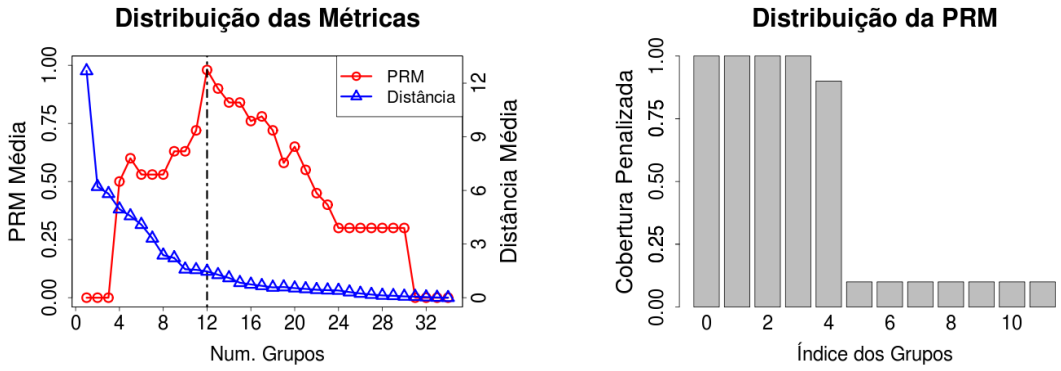


Figura 5.2: Eglina C (abordagem CCC): Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e os valores da PRM média (gráfico da esquerda) e os valores da métrica PRM para cada grupo (gráfico da direita)

A tabela 5.1 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos de agrupamentos dos HP-centroides AA e MO podem ser comparados, mesmo que em MO não existam valores para PRM. Neste caso, os valores das distâncias médias

inter- e intra-grupos são levadas em consideração para efeito de comparação.

Tabela 5.1: Análise CCC para Eglina C

COMPARAÇÃO DOS MÉTODOS (CCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	4.735	14.563	5	1	I	2.923	10.092	5	-
II	3.125	11.669	5	1	II	6.154	8.348	5	-
III	2.941	17.408	5	1	III	2.785	9.189	5	-
IV	2.844	12.220	5	1	IV	5.793	8.729	5	-
V	3.528	13.314	6	0.9	V	9.647	10.109	5	-
Médias	3.435	13.835	26	0.98	Médias	5.460	9.294	25	-

Comparando os dois métodos, notamos que a equiparação ou agrupamento dos HP-centroides, usando o método AA, é mais consistente. Os valores da distância média intra-grupo é menor e inter-grupo é maior em AA. Em outras palavras, os HP-centroides estão mais próximos dentro dos grupos e os grupos são mais distantes, o que são premissas para um bom agrupamento. Além disso, em AA o valor médio da cobertura (**0.98**) o que indica grupos de alta qualidade.

Granularidade fina: SGCC

Na abordagem SGCC (granularidade fina), observa-se que são necessários **24** grupos para que ocorra a estabilização das distâncias médias intra-grupo e o maior valor da PRM média. A métrica de avaliação aponta **4** de maior qualidade ($> 50\%$) dos **24** grupos com valor de cobertura considerável. Entretanto, é possível observar muitos outros grupos, ou seja, há muita variação dos HP-centroides dentro de um mesmo grupo, penalizando a cobertura. A figura 5.3 mostra que neste nível de abstração nós não podemos identificar um limiar que ressalte claramente grupos de alta qualidade de grupos de baixa qualidade.

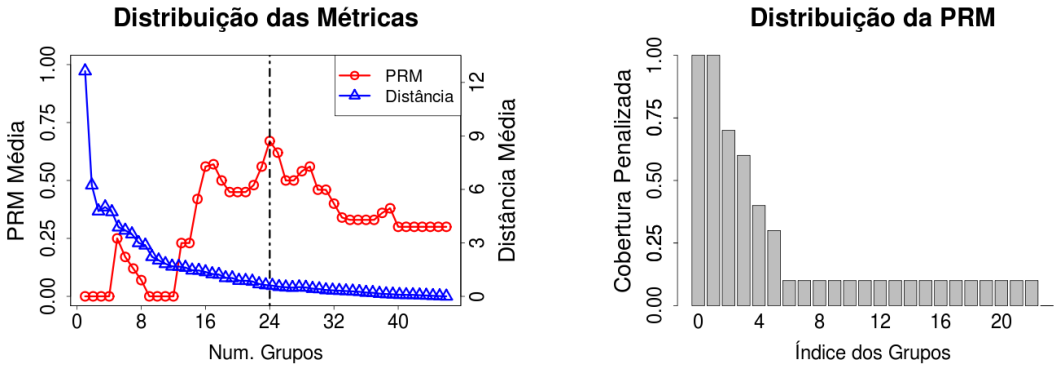


Figura 5.3: Eglina C (abordagem SGCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e os valores da PRM média (gráfico da esquerda) e os valores da métrica PRM para cada grupo (gráfico da direita)

A tabela 5.2 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não existam valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.2: Análise SGCC para Eglina C

COMPARAÇÃO DOS MÉTODOS (SGCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	2.840	11.485	5	1	I	7.363	16.348	5	-
II	2.361	17.882	5	1	II	5.227	10.236	5	-
III	2.371	13.504	8	0.70	III	2.823	9.964	5	-
IV	2.228	9.680	4	0.60	IV	4.960	8.375	5	-
Médias	2.450	13.138	22	0.82	Médias	5.093	11.231	20	-

Granularidade fina: EBCC

Na abordagem EBCC (granularidade fina), **19** grupos estabilizam as distâncias médias intra-grupo e apresenta maior valor da PRM média. A métrica de avaliação aponta **5** dos **19** grupos com valores de PRM mais conservados ($> 50\%$), representando regiões na interface dos cinco complexos. Embora haja certa semelhança com a abordagem CCC, os grupos são menos coesos. O que pode ser observado nos valores de cobertura penalizada. A figura 5.4 mostra que neste nível de abstração nós não podemos identificar um limiar que ressalte claramente grupos de alta qualidade de grupos de baixa qualidade.

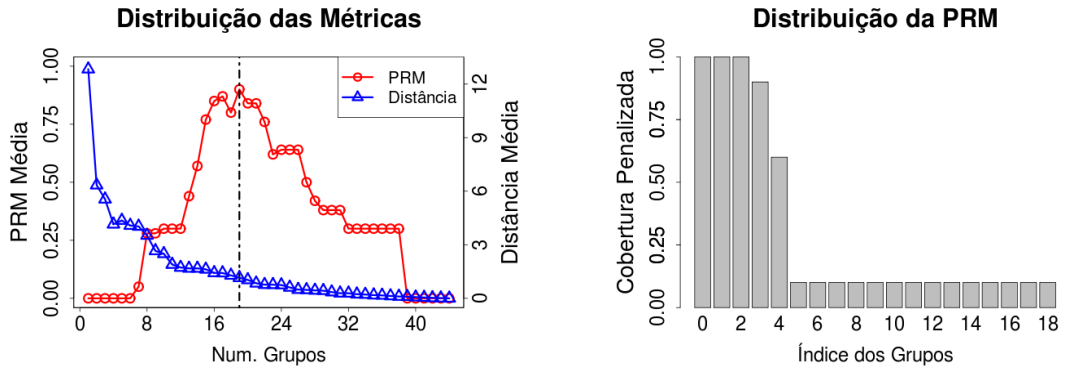


Figura 5.4: Eglina C (abordagem EBCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e os valores da PRM média (gráfico da esquerda) e os valores da métrica PRM para cada grupo (gráfico da direita)

A tabela 5.3 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não existam valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.3: Análise EBCC para Eglina C

COMPARAÇÃO DOS MÉTODOS (EBCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	2.448	16.456	5	1	I	9.444	10.736	5	-
II	3.076	11.255	5	1	II	2.176	11.300	5	-
III	2.421	11.629	5	1	III	5.793	9.007	5	-
IV	3.290	13.319	7	0.90	IV	3.076	9.852	5	-
V	2.160	10.692	4	0.60	V	6.205	9.035	5	-
Médias	2.679	12.670	26	0.90	Médias	5.339	9.986	25	-

Após a verificação individual de cada abordagem, nós apresentamos na tabela 5.4 uma síntese dos resultados para efeito de comparação dentre as três abordagens diferentes. Neste caso, nós calculamos e mostramos as médias gerais de cada uma das abordagens, juntamente com os modelos (AA e MO). Os melhores resultados são expostos em negrito, destacando-se o método AA que apresentou resultado da cobertura = **0.98**. A análise de granularidade grossa, em especial, apresentou os melhores resultados. As distâncias atingiram os valores baixos intra e altos inter grupos, que combinado com o valor na métrica possibilitou-nos concluir que a abordagem CCC, em conjunto com o método de agrupamento aglomerativo AA foi a mais apropriada.

Tabela 5.4: Comparação quantitativa entre as abordagens propostas para Eglina C

		Dist. Média Intra (Å)	Dist. Média Inter (Å)	Média PRM
CCC	AA	3.435	13.835	**0.98**
	MO	5.460	9.294	-
SGCC	AA	2.450	13.138	0.82
	MO	5.093	11.231	-
EBCC	AA	2.679	12.670	0.90
	MO	5.339	9.986	-

A semântica dos cinco[1] *patches* hidrofóbicos ou componentes conexos está representada pelos HP-centroides conservados que são apresentados na figura 5.5 {(a), (b), (c), (d) e (e)}). Nós podemos ver os grafos atômicos para a interface dos cinco complexos. Os átomos hidrofóbicos (representando os vértices) estão exibidos como pequenas esferas na cor cinza, juntamente com a identificação dos resíduos ao qual eles pertencem. As interações entre os átomos representam as arestas e os HP-centroides são exibidos como esferas coloridas localizadas no ponto central dos componentes conexos. Na última parte da figura (f), nós apresentamos a alça do inibidor (formada por 9 resíduos {GSPVTLDLR} - posições 40 a 48) em formato *sticks* na cor cinza, destacando os átomos hidrofóbicos em preto. Os cinco grupos de HP-centroides são sobrepostos em cores. Em escala de verde estão os HP-centroides das enzimas do Tipo Subtilisina e em vermelho os HP-centroides da enzima Tipo Tripsina.

[1]Neste caso, os 5 complexos analisados produziram, coincidentemente, 5 *patches* hidrofóbicos

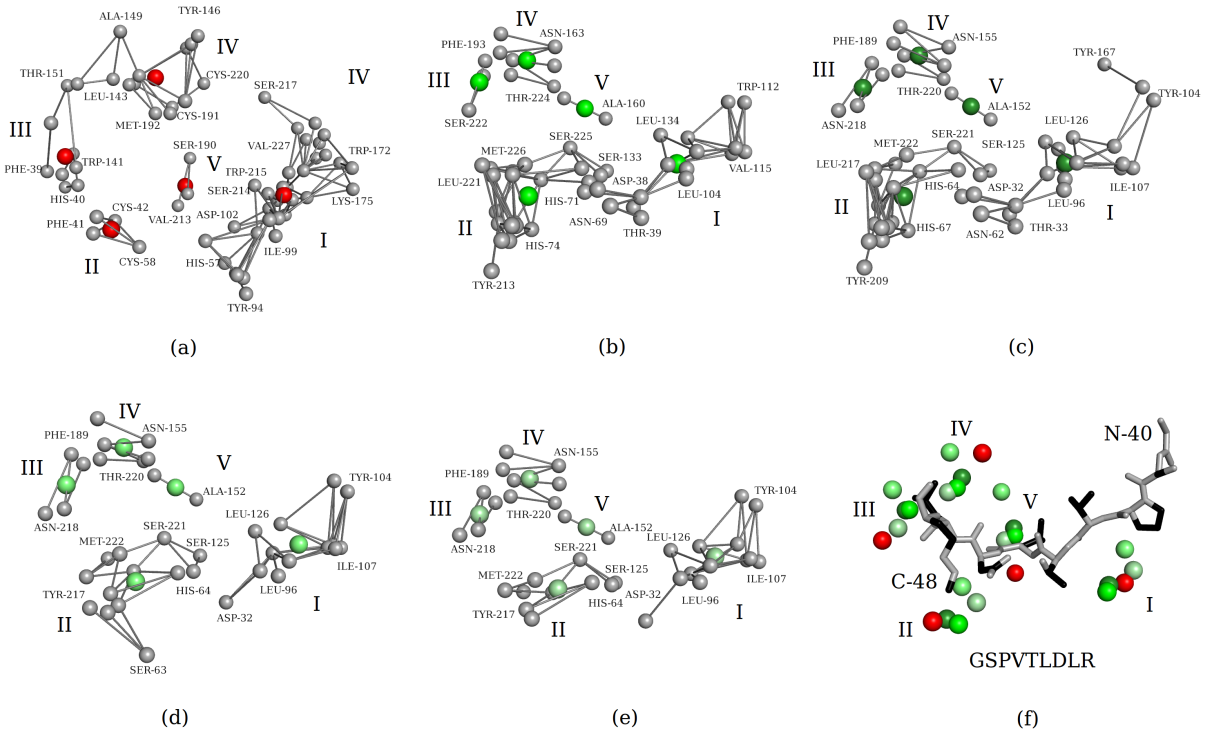


Figura 5.5: *Patches* hidrofóbicos para inibição cruzada com o inibidor Eglina C

O método proposto atinge um nível de abstração que ajuda a detectar padrões conservados na inibição cruzada. Por exemplo, quando nós comparamos os resíduos que compõem o grupo IV, nós podemos ver que na enzima Tipo Tripsina (a), há a presença dos resíduos de LEU-143, THR-151, ALA-149, TYR-146, CYS-220, CYS-191 e MET-192 e em contraparte em uma enzima Tipo Subtilisina (b), nós encontramos os resíduos PHE-193, ASN-163 e THR-224. Apesar de ter dissimilaridades na composição dos resíduos, no volume e na densidade do *patch*, a nossa metodologia (HydroPaCe) seleciona HP-centroides os quais são espacialmente conservados de acordo com o inibidor.

5.1.2 Inibidor Ovomucoide

Ovomucoides são os inibidores da protease glicoproteína aviária encontrada na clara do ovo. Particularmente, o inibidor Ovomucoide do peru pertence a família de inibidor do tipo *Kazal de serino protease* que ocorre naturalmente em *Meleagris gallopavo* e é um significante contaminante de preparação bruta de Ovomucoide. Atua sobre Tripsina e Quimotripsina, bem como Elastase suína e proteases de fungos [Robertson et al. (1988), Fujinaga et al. (1987)].

A figura 5.6 mostra o inibidor Ovomucoide (cor azul) complexado com duas enzimas (cor rosa) de famílias diferentes. É possível ainda, visualizar a interface de interação com o inibidor em cor amarela.

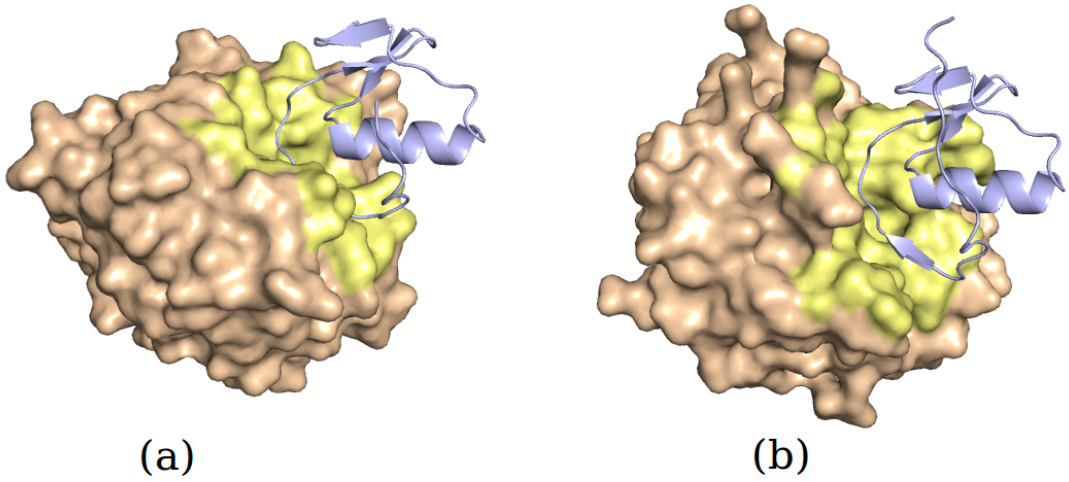


Figura 5.6: Exemplos de inibição cruzada com inibidor Ovomucoide. Em (a), Tipo Subtilisina PDB ID 1R0R e em (b), Tipo Tripsina PDB ID 1PPF.

Nós analisamos quatro complexos existentes, no qual 3 estavam em complexos com enzimas do Tipo Tripsina e 1 em complexo com uma enzima do Tipo Subtilisina. Em uma busca realizada na base de dados BRENDA, nós encontramos 5 *EC. numbers* diferentes que são conhecidos por serem inibidos por esta molécula.

Granularidade grossa: CCC

Na análise de granularidade grossa (CCC) os resultados são exibidos na figura 5.7. É possível observar que, no gráfico da esquerda, com **4** grupos há uma estabilização das distâncias médias intra-grupo e um maior valor para a PRM média. Com esta configuração nós obtemos **3** dos **4** grupos com alta qualidade ($> 50\%$), de acordo com a métrica PRM (gráfico da direita). Este conjunto de HP-centroides apresenta uma boa cobertura. Neste caso, os HP-centroides estão representados em todos os grupos.

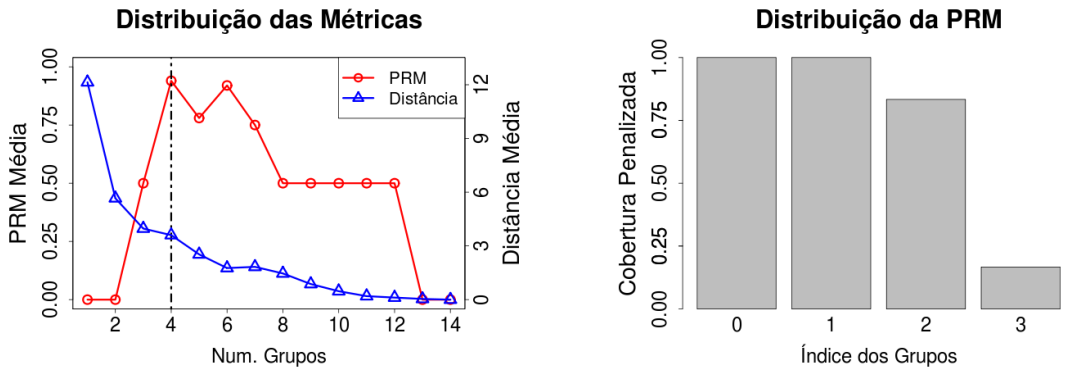


Figura 5.7: Ovomucoide (abordagem CCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e os valores da PRM média (gráfico da esquerda) e os valores da métrica PRM para cada grupo (gráfico da direita)

A tabela 5.5 mostra uma análise individual dos valores de cada grupo. Além disso, os

métodos de equiparação dos HP-centroides (AA e MO) podem ser comparados, mesmo que em MO não há valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.5: Análise CCC para Ovomucoide

COMPARAÇÃO DOS MÉTODOS (CCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	6.392	12.757	4	1	I	10.884	9.387	4	-
II	2.136	14.042	4	1	II	5.892	11.531	4	-
III	5.880	12.917	5	0.83	III	7.251	10.288	4	-
Médias	4.803	13.239	13	0.94	Médias	8.009	10.402	12	-

Granularidade fina: SGCC

Na abordagem SGCC (granularidade fina), observa-se que são necessários **14** grupos para que ocorra a estabilização das distâncias médias intra-grupo e as médias da PRM é a maior. A métrica de avaliação aponta apenas **2** dos **14** grupos com valores consideráveis de PRM ($> 50\%$) os quais representam *patches* hidrofóbicos na interface dos quatro complexos. Entretanto, é possível observar que existem vários outros grupos com baixa PRM, ou seja, há pouca conservação dos HP-centroides dentro de um mesmo grupo, penalizando muito a cobertura.

A figura 5.8 mostra que neste nível de abstração nós não podemos identificar um limiar que ressalte claramente grupos de alta qualidade de grupos de baixa qualidade.

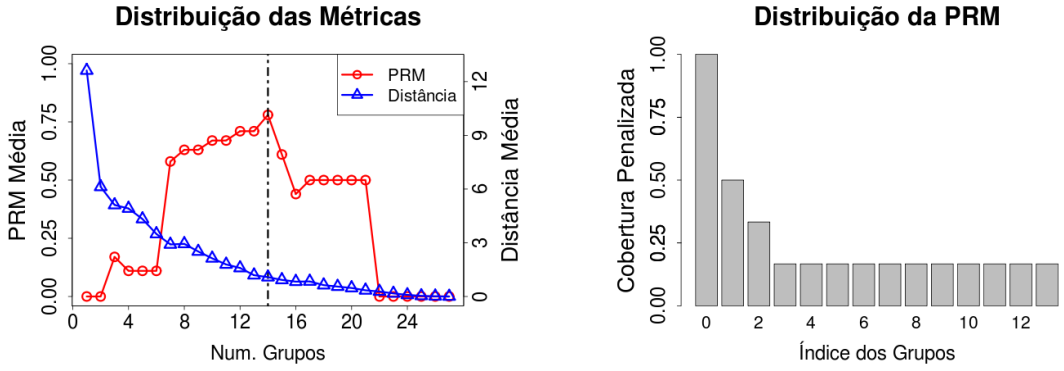


Figura 5.8: Ovomucoide (abordagem SGCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as médias da PRM (gráfico da esquerda). Os valores da métrica PRM para cada grupo estão no gráfico da direita

A tabela 5.6 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos de equiparação dos HP-centroides (AA e MO) podem ser comparados, mesmo que em MO não há valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.6: Análise SGCC para Ovomucoide

COMPARAÇÃO DOS MÉTODOS (SGCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	2.871	10.999	4	1	I	6.869	11.014	4	-
II	2.932	10.999	3	0.5	II	11.737	11.014	4	-
Médias	2.901	10.999	7	0.75	Médias	9.303	11.014	8	-

Granularidade fina: EBCC

Na abordagem EBCC (granularidade fina), 15 grupos estabilizam as distâncias médias intra-grupo e a métrica de avaliação (PRM) aponta somente **2** dos **15** grupos com valores consideráveis de PRM ($> 50\%$), os quais representam *patches* hidrofóbicos mais significativos na interface dos quatro complexos. Neste caso, existem muitas semelhanças entre as abordagens SGCC e EBCC. Comparando os gráficos e as médias da PRM podemos perceber o grau de semelhança. Entretanto, os grupos nas duas abordagens não são coesos, apresentando um comportamento pouco eficiente para detectar padrões conservados na interface de contato.

A figura 5.9 mostra que, neste nível de abstração, nós não podemos identificar um limiar que ressalte claramente grupos de alta qualidade de grupos de baixa qualidade.

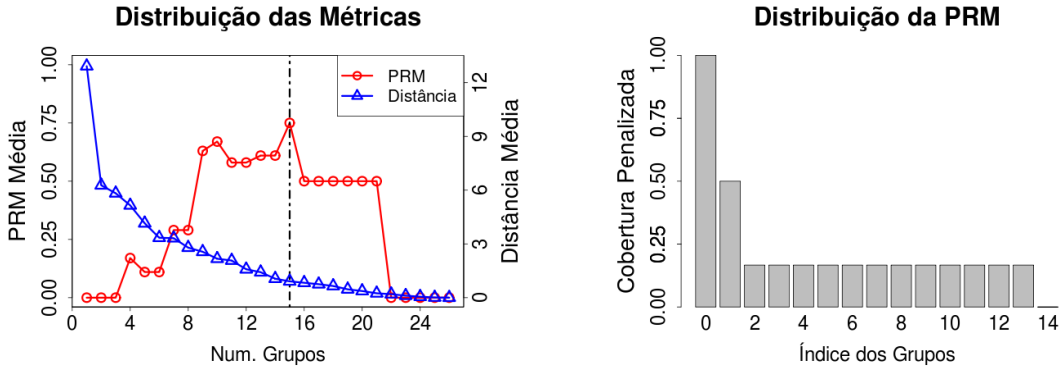


Figura 5.9: Ovomucoide (abordagem EBCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as médias da PRM (gráfico da esquerda). Os valores da métrica PRM para cada grupo estão no gráfico da direita

A tabela 5.7 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos de equiparação de HP-centroides (AA e MO) podem ser comparados, mesmo que em MO não há valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.7: Análise EBCC para Ovomucoide

COMPARAÇÃO DOS MÉTODOS (EBCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	3.605	14.045	4	1	II	3.965	11.997	4	-
II	3.233	14.045	3	0.50	III	8.953	11.997	4	-
Médias	3.419	14.045	7	0.75	Médias	6.459	11.997	8	-

Após a verificação individual de cada abordagem, nós apresentamos na tabela 5.8 uma síntese dos resultados para efeito de comparação entre as três abordagens diferentes. Neste caso, nós calculamos e mostramos as médias gerais de cada uma das abordagens, juntamente com os modelos (AA e MO). O melhor resultado é exposto em negrito, destacando-se o método AA que apresentou melhor resultado com cobertura = **0.94**. A análise de granularidade grossa, em especial, apresentou os melhores resultados. As distâncias atingiram os valores baixos intra e altos inter grupos, que combinado com os valores na métrica possibilitou-nos concluir que a abordagem CCC, em conjunto com o método de agrupamento aglomerativo AA foi a mais apropriada.

Tabela 5.8: Comparação quantitativa entre as abordagens propostas para Ovomucoide

		Dist. Média Intra (Å)	Dist. Média Inter (Å)	Média PRM
CCC	AA	4.803	13.239	**0.94**
	MO	8.009	10.402	-
SGCC	AA	2.901	10.999	0.75
	MO	9.303	11.014	-
EBCC	AA	3.419	14.045	0.75
	MO	6.459	11.997	-

A semântica dos três *patches* hidrofóbicos está representada pelos grupos conservados que são apresentados na figura 5.10 {(a), (b), (c) e (d). Nós podemos ver os grafos atômicos para a interface dos quatro complexos. Os átomos hidrofóbicos (representando os vértices) estão exibidos como pequenas esferas na cor cinza, juntamente com a identificação dos resíduos ao qual eles pertencem. As interações entre os átomos representam as arestas e os HP-centroides são exibidos como esferas coloridas localizadas no ponto central dos componentes conexos. Na última parte da figura (e), nós apresentamos a alça do inibidor (formada por 9 resíduos {KPACTLEYR} - posições 13 a 21) em formato *sticks* na cor cinza, destacando os átomos hidrofóbicos na cor preta. Os três grupos de HP-centroides são sobrepostos em cores. Em escala de vermelho estão os HP-centroides

das enzimas do Tipo Tripsinas e em verde os HP-centroides da enzima Tipo Subtilisina.

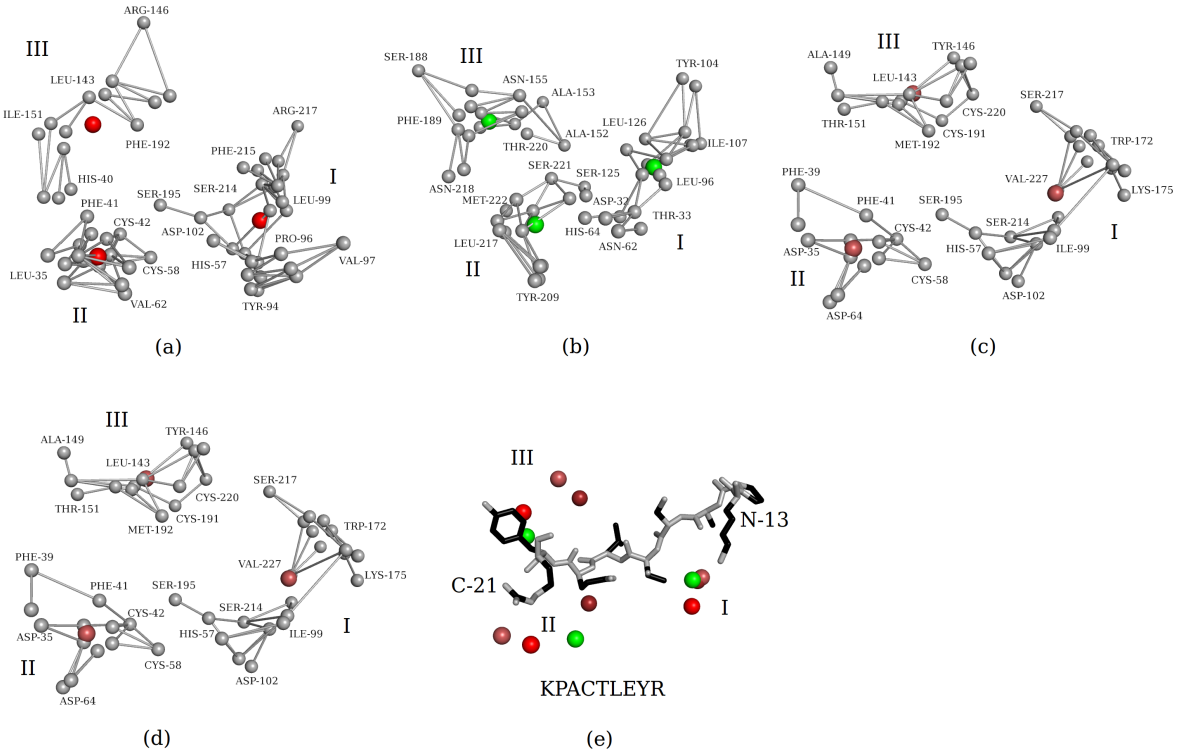


Figura 5.10: *Patches* hidrofóbicos para inibição cruzada pelo inibidor Ovomucoide

O método proposto atinge um nível de abstração que ajuda a detectar HP-centroides conservados nas interfaces em inibição cruzada. Por exemplo, quando nós comparamos os resíduos que compõem o grupo número III, nós podemos ver que na enzima Tipo Subtilisina (a) há presença de ASN-218, PHE-189, SER-188, ANS-155, ALA-153, ALA-152 e THR-220 e, em contraparte, na enzima Tipo Tripsina (b), nós encontramos HIS-40, ILE-151, LEU-143, ARG-146 e PHE-192. Apesar de ter dissimilaridades na composição dos resíduos, no volume e na densidade do *patch*, a nossa metodologia (HydroPaCe) seleciona HP-centroides os quais são espacialmente conservados de acordo com o inibidor.

5.2 Uso de HP-centroides para predição de inibição

Nossas buscas na base de dados do PDB revelaram que complexos com estruturas tridimensionais resolvidas, representando exemplos de inibição cruzada, são escassos. É intrigante presumir a possibilidade de generalizar os padrões de HP-centroides obtidos na inibição cruzada para sítios de ligação em apo enzimas, ou seja, enzimas de estruturas conhecidas, porém, de cadeia simples (sem inibidor).

Nós acreditamos que os HP-centroides encontrados nos complexos, envolvendo inibição cruzada, também podem ser encontrados nas interfaces das apo enzimas de uma mesma família. Em nossas bases de dados, tínhamos dois conjuntos de apo enzimas, do tipo

serino proteases, com identidade de sequência consideravelmente baixa (menor que 50%). Nossas bases de dados são compostas de 35 enzimas Tipo Tripsina e 9 enzimas Tipo Subtilisina não redundantes. É importante ressaltar que em uma simples busca na base de dados do PDB o número de instâncias recuperadas é bem maior, porém, existem muitas redundâncias ou enzimas iguais em organismos diferentes, mas com identidade de sequência alta. Nós entendemos que a diversidade de enzimas é mais importante que a quantidade em nossas análises.

Deste modo, nós tínhamos um desafio. Como determinar os HP-centroides em enzimas sem inibidor e consequentemente sem interface definida? Sabemos que é determinante para uma família de enzimas que haja similaridade nas estruturas tridimensionais de seus membros, independente dos resíduos que compõem as sequências. Assim, nós escolhemos um complexo modelo representativo de cada família e fizemos um alinhamento estrutural, por família, envolvendo o complexo modelo e todos as apo enzimas. Como os resíduos que compõem a interface dos complexos são conhecidos, nós identificamos os resíduos nas apo enzimas que se alinharam aos resíduos da interface dos complexos em um processo que nós chamamos de projeção.

Na figura 5.11 é possível verificar a conservação dos resíduos nas interfaces projetadas por meio dos logos (diagramas de entropia informacional). Observando somente a conservação da interface, a nível de sequência, não é possível entender como a inibição cruzada ocorre, devido a alta entropia informacional dos resíduos. Entretanto, é possível notar a existência de vários resíduos conservados nas duas famílias, porém, estes resíduos são conhecidos por participar no processo de catálise. Esses resíduos podem ser da tríade catalítica (marcados com asterísco (*)), da cavidade do oxianionte (marcados com sinal de mais (+)) ou de sítios de especificidade (marcados com círculo (o)). À parte destes resíduos, não existem conservações significativas que pode ser facilmente observados nos logos.

Após determinar os resíduos que estão na interface das apo enzimas, por meio da projeção, nós aplicamos a metodologia HydroPaCe em cada uma das famílias (Tipo Subtilisina e Tipo Tripsina). Nós observamos uma forte conservação dos HP-centroides encontrados nas análises dos complexos com inibição cruzada. Nossa hipótese é que para a inibição ocorrer nós teríamos que ter conservação dos *patches* em posições específicas para acomodar o inibidor. Por exemplo, o resíduo de PHE-215 em Tipo Tripsina nas figuras 5.11(b) e 5.6(a) é um resíduo hidrofóbico volumoso que é equivalente as porções hidrofóbicas de resíduos nas posições LEU-96, ILE-107 e LEU-126 em Tipo Subtilisina nas figuras 5.11(a) e 5.6(b). Este é um exemplo onde padrões conservados não podem ser inferidos da sequência ou estrutura. Eles são claramente observados em nosso *patch* I.

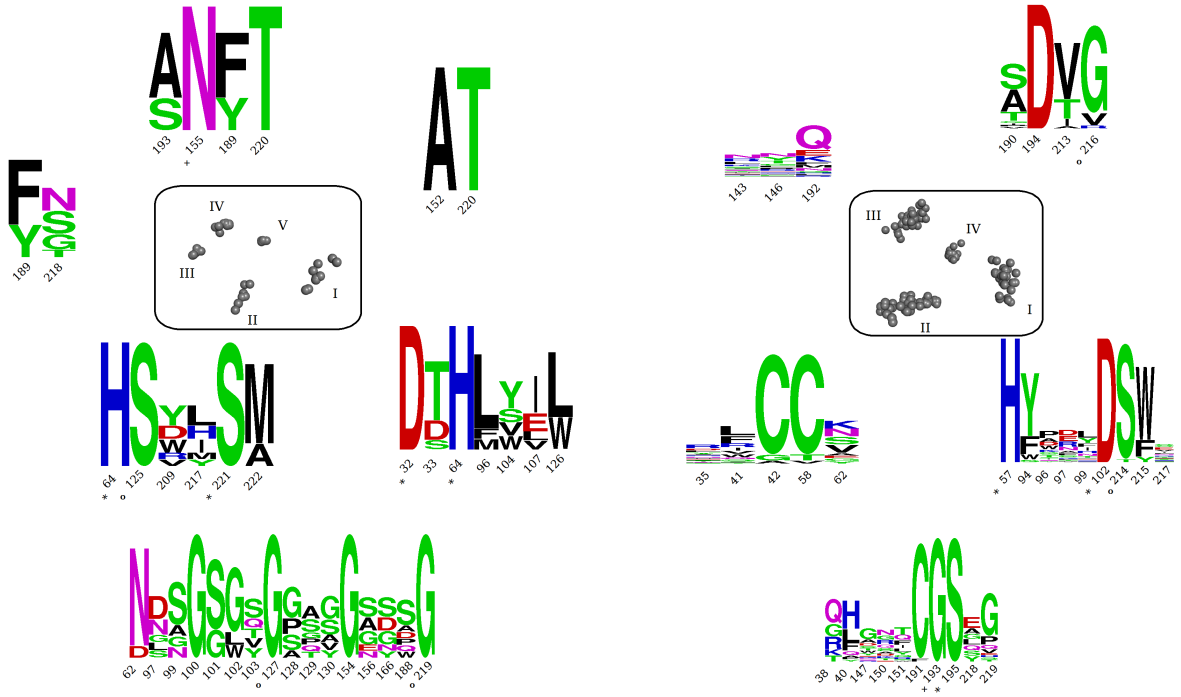


Figura 5.11: Predição da IFRs e HP-centroides encontrados nas inibições cruzadas com inibidor Ovomucoide

Além disso, nós acreditamos que estes padrões possam ser usados para predizer inibições para outras enzimas cujas estruturas tridimensionais estão disponíveis, mas não há evidência experimental de inibição conhecida. Por exemplo, nós usamos 8 exemplares, não redundantes, de apo enzimas do Tipo Subtilisina, tabela 4.4 na seção 4.1, pertencentes a 5 *EC. number* diferentes (3.4.21.62/ 64/ 66/ 75/ 97). Nós consideramos somente enzimas na qual estão com ECs completo com os quatro níveis de anotação. De acordo com a base de dados BRENDA, 3 deles referem-se a inibição por Eglina C (3.4.21.62/ 66/ 75) o que pode ser considerado predições bem sucedidas. Os outros dois restantes (3.4.21.64/ 97) referem as enzimas Proteinase K e Assemblin Protease não são mencionados na literatura, mas apresentam o mesmo padrão das outras enzimas do Tipo Subtilisina. Seria muito interessante verificar experimentalmente se elas podem ser inibidas pelo Eglina C, como eles apresentam os mesmos HP-centroides, nós acreditamos ser possível esta interação.

Todos os resultados da metodologia empregada para encontrar padrões conservados de HP-centroides nas interfaces projetadas das apo enzimas são expostos individualmente nas próximas seções.

5.2.1 Família Tipo Tripsina

Nesta seção, nós apresentamos a análise e a comparação das três abordagens (CCC, SGCC e EBCC) na busca por padrões conservados de HP-centroides nas interfaces projetadas das apo enzimas da família Tipo Tripsina.

Granularidade grossa: CCC

Na análise de granularidade grossa (CCC) os resultados são exibidos na figura 5.12. É possível observar que, no gráfico da esquerda, com **12** grupos, há uma estabilização das distâncias médias intra-grupo e a maior PRM média. Com esta configuração nós obtemos **4** dos **12** grupos mais conservados ($> 50\%$), de acordo com a métrica PRM (gráfico da direita). Entretanto, nem todas as enzimas possuem HP-centroides em todos os grupos têm HP-centroides redundantes. Embora haja um bom comportamento dos HP-centroides, nós devemos investigar os casos negativos.

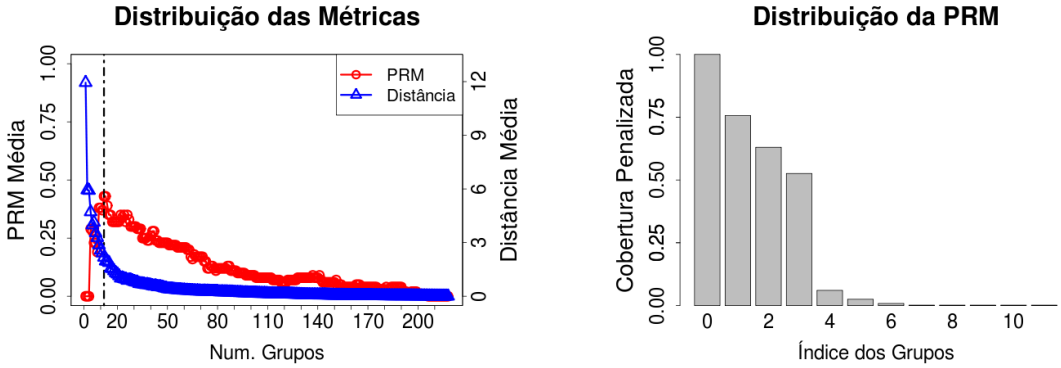


Figura 5.12: Tipo Tripsina (abordagem CCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as médias da PRM (gráfico da esquerda). Os valores da métrica PRM para cada grupo são mostrados no gráfico da direita

A tabela 5.9 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não há valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.9: Análise CCC: predição IFR Tipo Tripsina

COMPARAÇÃO DOS MÉTODOS (CCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	1.312	11.729	35	1	I	5.122	10.547	35	-
II	3.685	10.499	46	0.75	II	2.139	9.498	35	-
III	4.587	13.422	51	0.58	III	6.848	9.955	35	-
IV	5.787	15.267	60	0.52	IV	4.823	11.202	35	-
Médias	3.843	12.729	192	0.71	Médias	4.733	10.300	140	-

Granularidade fina: SGCC

Na análise de granularidade fina (SGCC) os resultados são exibidos na figura 5.13. É possível observar que, no gráfico da esquerda, com **73** grupos há uma estabilização das

distâncias médias intra-grupo. Com esta configuração nós obtemos **5** dos **73** grupos são mais conservados ($> 50\%$), entretanto não é possível identificar um limiar que ressalta claramente a qualidade dos grupos bons dos ruins. Os valores da métrica PRM de cada grupo são vistos no gráfico da direita.

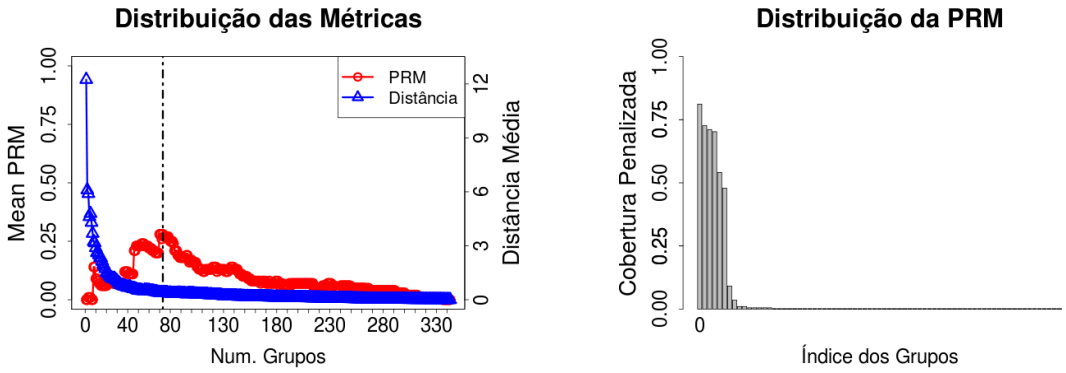


Figura 5.13: Tipo Tripsina (abordagem SGCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as médias da PRM (gráfico da esquerda). Os valores da métrica PRM para cada grupo são demonstrados no gráfico da direita

A tabela 5.10 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não há valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.10: Análise SGCC: predição IFR Tipo Tripsina

COMPARAÇÃO DOS MÉTODOS (SGCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å)		Tam.	PRM	Grupos	Distâncias (Å)		Tam.	PRM
	Intra	Inter				Intra	Inter		
I	1.412	11.146	47	0.81	I	5.161	7.882	35	-
II	1.755	11.915	33	0.72	II	6.649	6.769	35	-
III	2.843	13.382	48	0.71	III	9.542	5.109	35	-
IV	2.669	12.003	45	0.70	IV	9.044	6.725	35	-
V	2.169	11.356	29	0.54	V	9.557	5.759	35	-
Médias	2.169	11.960	202	0.69	Médias	7.990	6.448	175	-

Granularidade fina: EBCC

Na análise de granularidade fina (EBCC) os resultados são exibidos na figura 5.14. É possível observar que, no gráfico da esquerda, com **5** grupos há uma estabilização das distâncias médias intra-grupo. Com esta configuração nós obtemos **3** dos **5** grupos mais conservados ($> 50\%$), entretanto, também não é possível identificar um limiar que

ressalta claramente a qualidade dos grupos bons dos ruins. Os valores da métrica PRM de cada grupo são vistos no gráfico da direita.

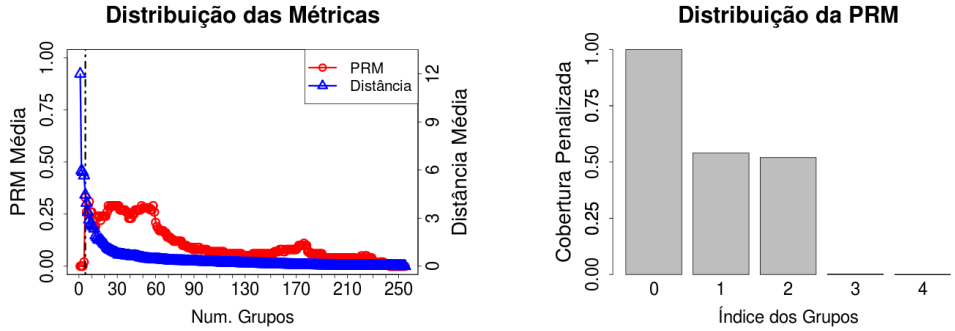


Figura 5.14: Tipo Tripsina (abordagem EBCC. Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as PRM médias (gráfico da esquerda). Os valores da métrica PRM para cada grupo são exibidos no gráfico da direita

A tabela 5.11 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não há valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.11: Análise EBCC: predição IFR Tipo Tripsina

COMPARAÇÃO DOS MÉTODOS (EBCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	0.942	10.623	35	1	I	7.958	6.724	35	-
II	2.384	12.326	28	0.54	II	9.078	4.626	35	-
III	2.292	12.340	60	0.52	III	7.429	6.895	35	-
Médias	1.872	11.763	123	0.68	Médias	8.155	6.082	105	-

Após a verificação individual de cada abordagem, nós apresentamos na tabela 5.12 uma síntese dos resultados comparando as três abordagens diferentes.

Tabela 5.12: Comparação quantitativa entre as abordagens propostas para predição das IFRs das enzimas do Tipo Tripsina

		Dist. Média Intra (Å)	Dist. Média Inter (Å)	Média PRM
CCC	AA	3.843	12.729	**0.71**
	MO	4.733	10.300	-
SGCC	AA	2.169	11.960	0.69
	MO	7.990	6.448	-
EBCC	AA	1.872	11.763	0.68
	MO	8.155	6.082	-

Neste caso, nós calculamos e mostramos as médias gerais de cada uma das abordagens, juntamente com os métodos (AA e MO) para cada uma delas. Os melhores resultados são

expostos em negrito, destacando-se o método AA que apresentou resultado da cobertura penalizada (PRM) = **0.71**. A análise de granularidade grossa, em especial, apresentou os melhores resultados. As distâncias atingiram os valores baixos intra e altos inter grupos, que combinado com os valores na métrica possibilitou-nos concluir que a abordagem CCC, que em conjunto com o método de agrupamento aglomerativo AA foi a mais apropriada.

Para as interfaces projetadas na família Tipo Tripsina as abordagem de grunularidade fina (SGCC e EBCC) mostraram-se pouca eficiência. O comportamento dos grupos indicam que existem poucas conservações, sendo assim bastante penalizados. Porém, a abordagem de grossa granularidade CCC mostrou-se mais apropriada, apresentando mais HP-centroides conservados.

5.2.2 Família Tipo Subtilisina

Nesta seção, nós apresentamos a análise e a comparação das três abordagens (CCC, SGCC e EBCC) na busca por padrões conservados de HP-centroides nas interfaces projetadas para as apo enzimas da família do Tipo Subtilisina.

Granularidade grossa: CCC

Na análise de granularidade grossa (CCC) os resultados são exibidos na figura 5.15. É possível observar que, no gráfico da esquerda, com **17** grupos há uma estabilização das distâncias médias intra-grupo e a maior PRM média. Com esta configuração nós obtemos 4 dos **17** grupos são mais conservados ($> 50\%$), de acordo com a métrica PRM (gráfico da direita). Os HP-centroides estão presentes em 100% nos três primeiros grupos. O quarto, nós consideramos conservado, entretanto, nos demais a cobertura é considerada baixa. Neste caso, podemos dizer que de todos os grupos 3 têm máxima qualidade.

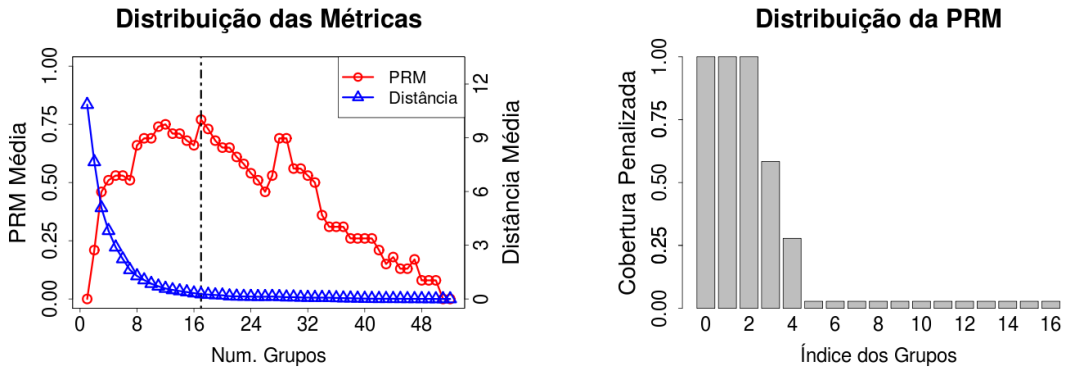


Figura 5.15: Tipo Subtilisina (abordagem CCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as médias da PRM (gráfico da esquerda). Os valores da métrica PRM para cada grupo são exibidos no gráfico da direita

A tabela 5.13 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não existam valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.13: Análise CCC: predição IFR Tipo Subtilisina

COMPARAÇÃO DOS MÉTODOS (CCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	1.106	10.614	9	1	I	3.893	9.983	9	-
II	0.755	9.231	9	1	II	5.570	9.261	9	-
III	0.479	9.251	9	1	III	5.619	8.393	9	-
IV	1.080	13.530	7	0.58	IV	3.126	11.720	9	-
Médias	0.885	10.656	34	0.89	Médias	4.552	9.839	36	-

Granularidade fina: SGCC

Na análise de granularidade fina (SGCC) os resultados são exibidos na figura 5.16. É possível observar que, no gráfico da esquerda, com **15** grupos há uma estabilização das distâncias médias intra-grupo. Com esta configuração nós obtemos **5** dos **25** grupos mais conservados ($> 50\%$), de acordo com a métrica PRM (gráfico da direita). Os HP-centroides estão presentes em 100% somente no primeiro grupo. Nos demais grupos não foi possível identificar um limiar que ressalta os grupos de boa qualidade dos grupos de qualidade ruim.

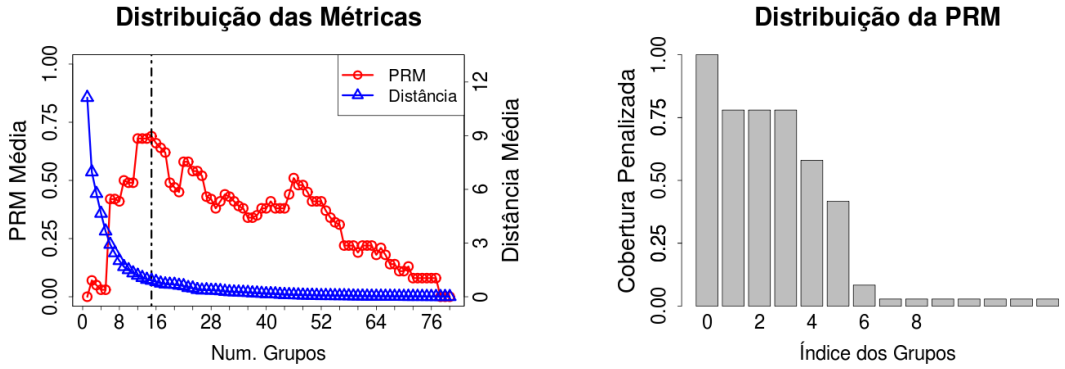


Figura 5.16: Tipo Subtilisina (abordagem SGCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as médias da PRM (gráfico da esquerda). Os valores da métrica PRM para cada grupo são exibidos no gráfico da direita

A tabela 5.14 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não existam valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.14: Análise SGCC: predição IFR Tipo Subtilisina

COMPARAÇÃO DOS MÉTODOS (SGCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	2.103	12.330	9	1	I	4.386	10.968	9	-
II	1.243	10.002	8	0.78	II	8.176	9.881	9	-
III	0.479	11.267	8	0.78	III	5.707	9.816	9	-
IV	1.022	13.927	8	0.78	IV	6.172	8.899	9	-
V	1.144	16.910	7	0.58	V	3.502	9.055	9	-
Médias	1.198	12.887	40	0.78	Médias	5.588	9.723	45	-

Granularidade fina: EBCC

Na análise de granularidade fina (EBCC) os resultados são exibidos na figura 5.17. É possível observar que, no gráfico da esquerda, com **14** grupos há uma estabilização das distâncias médias intra-grupo. Com esta configuração nós obtemos **6** dos **14** grupos têm valores de PRM mais conservados ($> 50\%$) (gráfico da direita). Os HP-centroides estão presentes em 100% somente nos dois primeiros grupos. Para os demais grupos não foi possível identificar um limiar que ressalta os grupos de boa qualidade dos grupos de qualidade ruim.

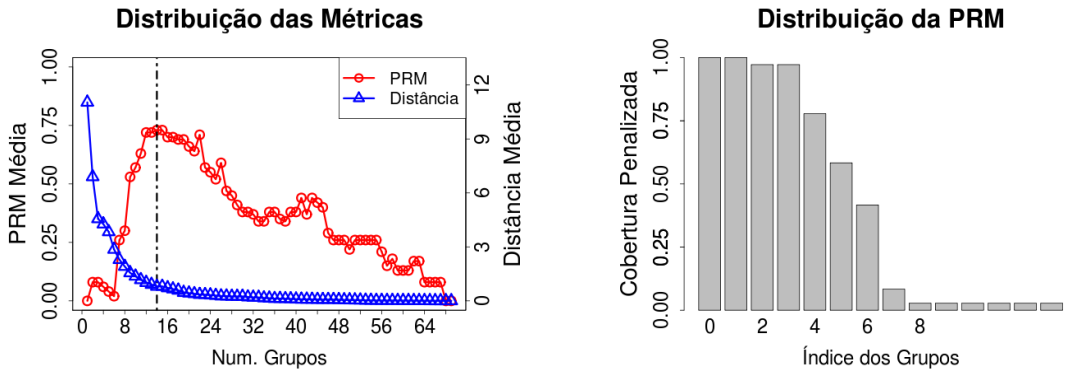


Figura 5.17: Tipo Subtilisina (abordagem EBCC). Relação entre o número de grupos, as distâncias médias dos HP-centroides intra-grupos e as médias da PRM (gráfico da esquerda). Os valores da métrica PRM para cada grupo são exibidos no gráfico da direita

A tabela 5.15 mostra uma análise individual dos valores de cada grupo. Além disso, os métodos AA e MO podem ser comparados, mesmo que em MO não existam valores para PRM. Neste caso, os valores das distâncias médias inter e intra grupos são levados em consideração para efeito de comparação.

Tabela 5.15: Análise EBCC: predição IFR Tipo Subtilisina

COMPARAÇÃO DOS MÉTODOS (EBCC)									
Agrupamento Aglomerativo (AA)					Modelo de Otimização (MO)				
Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM	Grupos	Distâncias (Å) Intra	Inter	Tam.	PRM
I	1.611	12.166	9	1	I	5.608	8.394	9	-
II	0.479	10.160	9	1	II	5.220	9.750	9	-
III	2.728	12.660	11	0.97	III	3.257	10.321	9	-
IV	1.132	11.242	10	0.97	IV	6.772	7.611	9	-
V	1.464	13.213	8	0.78	V	4.456	10.758	9	-
VI	1.419	9.291	7	0.58	VI	6.232	9.671	9	-
Médias	1.472	11.455	54	0.88	Médias	5.257	9.417	54	-

Para as interfaces projetadas na família Tipo Subtilisina as abordagens de granularidade fina (SGCC e EBCC) mostraram-se pouco eficiêntes. O comportamento dos grupos indica que existem poucas conservações, sendo assim bastante penalizados. Entretanto, nota-se uma similaridade entre as abordagens CCC e EBCC. Porém, a abordagem de granularidade grossa CCC mostrou-se mais eficiente, o que pode ser comprovado pela comparação dos valores da métrica de avaliação.

Após a verificação individual de cada abordagem, nós apresentamos na tabela 5.16 uma síntese dos resultados para efeito de comparação entre as três abordagem diferentes.

Tabela 5.16: Comparação quantitativa entre as abordagens propostas para predição das IFRs das enzimas do Tipo Subtilisina

		Dist. Média Intra (Å)	Dist. Média Inter (Å)	Média PRM
CCC	AA	0.885	10.656	**0.89**
	MO	4.552	9.839	-
SGCC	AA	1.198	12.887	0.78
	MO	5.588	9.723	-
EBCC	AA	1.472	11.455	0.88
	MO	5.257	9.417	-

Neste caso, nós calculamos e mostramos as médias gerais de cada uma das abordagens, juntamente com os métodos (AA e MO). Os melhores resultados são expostos em negrito, destacando-se o método AA que apresentou resultado da cobertura = **0.89**. A análise de granularidade grossa, em especial, apresentou os melhores resultados. As distâncias atingiram os valores baixos intra e altos inter grupos, que combinado com os valores na métrica possibilitou-nos concluir que a abordagem CCC, que em conjunto com o método de agrupamento aglomerativo AA foi a mais apropriada.

5.3 Considerações finais

A análise das três abordagens diferentes (CCC, SGCC e EBCC) apresentou comportamento similar nas quatro bases de dados analisadas. Entretanto, a abordagem de

granularidade grossa (CCC) foi mais eficiente para identificar os HP-centroides conservados em todos os casos. Identificamos que a fragmentação dos componentes conectados, premissa da abordagem de granularidade fina, foi menos adequada.

Nós acreditamos que a variabilidade no tamanho dos componentes possa ter influenciado em uma análise de resolução mais baixa. Para os componentes de tamanho menor a fragmentação pode não ter ocorrido, seguindo regras dos algoritmos de detecção de comunidades. Entretanto, em alguns casos, notamos uma certa semelhança entre os algoritmos de granularidade grossa (CCC) e granularidade fina (EBCC). Na análise com as interfaces projetadas das apo enzimas do Tipo Subtilisina as médias globais da PRM apresentaram valores bem próximos (CCC=0.89 e EBCC=0.88). Este fato pode ter ocorrido devido a uma maior similaridade entre o tamanho dos componentes, ocasionando a fragmentação gerando mais grupos que conservaram os HP-centroides em EBCC.

Com base nos resultados apresentados, nós concluímos que as interações hidrofóbicas são muito importantes para o reconhecimento enzima-inibidor, entretanto interações polares podem ser consideradas em trabalhos futuros. Mas interessantemente, nós encontramos um mínimo de três HP-centroides. As proteínas são objetos tridimensionais e nós conjecturamos que, para uma molécula se fixar a outra, devem existir no mínimo três pontos de contatos não colineares. Na verdade, isto é verificado em nossos padrões de inibição cruzada quando nós encontramos um mínimo de três HP-centroides hidrofóbicos essenciais.

Uma vez que o objetivo é encontrar o máximo número de grupos de HP-centroides conservados, a abordagem de granularidade grossa apresentou sistematicamente um resultado melhor. Isto pode indicar que padrões de inibição cruzada é dependente da posição relativa dos HP-centroides e independente de seu volume.

Há vários trabalhos, descritos na literatura sobre interações atômicas em proteínas ou busca por padrões conservados a nível de resíduos (usando métodos tradicionais de alinhamentos). Os trabalhos que estudam as interações atômicas tentam descrever todos os tipos de interações existentes e muitas vezes os padrões de interação não se conservam, devido a sua diversidade.

A metodologia HydroPaCe foi proposta para analisar interações atômicas hidrofóbicas nas interfaces de contatos ou bolsões catalíticos, mapeando-as em redes. Estas redes de interações podem ser consideradas como pontos de interações que determinam a afinidade entre enzima-inibidor. Os resultados indicam que a metodologia proposta atingiu seu objetivo e pode ser estendida a outras análises envolvendo interações em proteínas.

Nós desenvolvemos ainda um *web site* que disponibiliza informações sobre a metodologia. Além disso, todos os códigos, algoritmos e um tutorial explicando todos os passos metodológicos estão disponíveis em www.dcc.ufmg.br/∼raquelcm/hydropace.

Capítulo 6

Conclusão

A conclusão, em relação à proposta central desse trabalho, pode ser respondida positivamente. Constatou-se, através da metodologia proposta, a qual chamamos de HydroPaCe (***Hydro****fobic* ***Pa****tch* ***Ce****ntroid*), que a investigação *in silico*, sobre padrões de reconhecimento molecular entre enzima e inibidor, especialmente dos casos em que ocorre inibição cruzada, foi bem sucedida.

Os métodos tradicionais como alinhamentos de sequências e estruturas não são apropriados para detectar conservação na interface, em casos onde ocorrem inibição cruzada. Os resíduos que estão presentes na interface não são conservados, com exceção dos resíduos que estão envolvidos diretamente na catálise enzimática, como por exemplo, os resíduos que formam a tríade catalítica e sítios de especificidades.

A falta de uma metodologia que pudesse tentar entender como a inibição cruzada ocorre motivou-nos a desenvolver algo novo. Percebemos que as contribuições das mudanças entrópicas é determinante para que o reconhecimento molecular ocorra. Neste sentido, notamos que deveríamos concentrar os esforços nas interações atômicas hidrofóbicas, identificando regiões na interface molecular que interagem com o ligante. Entretanto, notamos que as regiões têm tamanhos e formas diferentes e que era necessário abstrair estas características para efeito de comparação.

A metodologia proposta é formada por um conjunto de abordagens que identifica e analisa as regiões hidrofóbicas em dois níveis de abstração: em granularidade grossa e granularidade fina. Para isso, nós modelamos a interface de contato dos complexos, envolvendo inibição cruzada, através de grafos de interações hidrofóbicas, em nível atômico. O mapeamento das interfaces enzima/inibidor, em grafos, revelou regiões conectadas. Estas regiões foram representadas como centroides geométricos, o que nós chamamos de HP-centroides. Os HP-centroides produzidos a partir das interfaces dos complexos foram comparados a fim de encontrar padrões conservados que pudessem explicar porque a inibição cruzada ocorre.

Nós apresentamos alguns estudos de casos, envolvendo famílias em serino proteases (Tipo Tripsina e Tipo Subtilisina) que são inibidas pelos inibidores Eglina C e Ovomu-

coide. Em seguida, nós mostramos que nossa metodologia foi capaz de detectar regiões hidrofóbicas conservadas que podem explicar a inibição cruzada. Além disso, nós estendemos nossas análises, projetando as interfaces dos complexos para as apo enzimas. Verificamos que as regiões conservadas nos complexos podem ser vistas na maioria das enzimas da família. A existência de dados experimentais conhecidos fortalece nossa hipótese. Ou seja, é possível identificar inibições previstas por nós, por meio da literatura (base de dados BRENDA).

Acreditamos que nosso trabalho pode contribuir para novas pesquisas em bioinformática, tendo como visão a busca por padrões em nível atômico. Além disso, as predições de interações, previstas por nós e que ainda não têm dados experimentais, podem ser testadas.

Capítulo 7

Perspectivas e trabalhos futuros

7.1 Perspectivas

Desde o início das atividades deste trabalho, várias técnicas foram utilizadas na tentativa de recuperar informações que pudessem tentar entender como a inibição cruzada ocorre. Os resíduos que formam a interface molecular ou IFR das famílias de proteases Tipo Tripsina e Tipo Subtilisina foram analisadas. Técnicas de descoberta de padrões foram usadas, como o SVD (*Singular Value Decomposition*) que usa álgebra linear e *fingerprint*, porém, não tiveram um bom desempenho, pois a informação de conservação dos resíduos não caracterizava padrões.

Analisar os resíduos que compõem as interfaces, na tentativa de explicar como as inibições cruzadas ocorrem, se tornou uma árduo trabalho. Nós não conseguimos desenvolver uma técnica robusta que pudesse encontrar padrões conservados. Porém, os resíduos podem ser classificados de acordo com suas características físico-químicas. Por exemplo, o resíduo de lisina é considerado carregado positivamente (em pH neutro). Assim, os resíduos foram agrupados, mas infelizmente, embora os padrões tenham melhorado, ainda sim, eles não eram consenso na maioria das interfaces analisadas.

Deste modo, nós decidimos trabalhar em um nível de abstração atômica. Embora os resíduos tenham uma classificação físico-química na literatura, nós entendemos que todos eles têm porções apolares (hidrofóbicas). Nós baseamos em estudos de interações atômicas e propomos uma tabela que identifica os átomos, de cada resíduo, que são considerados hidrofóbicos. Com esta nova perspectiva, nós desenvolvemos uma metodologia que identifica os átomos hidrofóbicos nas interfaces e os mapeiam em grafos de interações.

Neste nível de abstração, nós conseguimos identificar regiões conectadas e conservadas em interfaces com resíduos muito dissimilares. O que é comum em famílias de enzimas diferentes e que retrata a inibição cruzada.

O nosso desejo é que a metodologia, proposta por nós, possa servir de base para outros trabalhos, contribuindo assim para a comunidades científica.

7.2 Trabalhos futuros

Os resultados apresentados neste trabalho descrevem uma metodologia capaz de identificar padrões de interações em complexos de enzimas de famílias diferentes. Além disso, revelou que estes padrões podem ser estendidos para enzimas que não têm estruturas tridimensionais em complexos no PDB.

Nós entendemos que nossos objetivos foram atingidos, entretanto, existem inúmeros trabalhos que ainda podem ser desenvolvidos, tendo como base a metodologia proposta. A seguir, nós listamos alguns possíveis trabalhos que são pertinentes e que podem ser desenvolvidos pelo nosso grupo.

- Analisar a existência de padrões no *core* hidrofóbico de proteínas enoveladas de uma família, por exemplo mioglobina;
- Analisar a existência de HP-centroides envolvendo interações de ligantes com proteínas;
- Identificar a existência de HP-centroides na interface de diferentes inibidores que se ligam à mesma enzima (por exemplo tripsina);
- Propor um *loop* ou alça consenso que pode ser usada para inibir proteases;
- Analisar a existência de um padrão em proteases que são inibidas por diferentes inibidores. Verificar se existem HP-centroides comuns para diferentes inibidores;
- Testar algoritmos diferentes para analisar os HP-centroides (árvore geradora mínima);
- Analisar possíveis padrões de contatos inter cadeia, ou seja, proteína e inibidor.

Anexo A

Artigo submetido

Os resultados deste trabalho de tese foram submetidos para a revista Bioinformatics (http://bioinformatics.oxfordjournals.org/) intitulado como "HydroPaCe: understanding and predicting cross-inhibition in serine protease through hydrophobic patch centroids"e encontra-se na segunda etapa do processo de revisão. A segunda etapa consiste na adequação e resposta aos revisores.

A revista Bioinformatics é, atualmente, lider mundial no campo da Bioinformática e publica trabalhos científicos de alta qualidade. Possui um fator de impacto de 4.877 e é considerada pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) uma revista com *qualis* A1, sendo considerada uma das principais fontes de pesquisa tanto no meio acadêmico quanto na indústria.

Os autores acreditam no potencial de contribuição que este trabalho possa dar à comunidade científica. Neste sentido, a escolha de uma revista de impacto mundial é importante para divulgar a metodologia empregada e os resultados obtidos. Acredita-se ainda, que a contribuição deste trabalho não limita-se apenas ao uso da metodologia empregada, mas encoraja outros pesquisadores a propor mudanças e melhorias que ajudam a entendender as interações de enzimas com substratos e inibidores.

Alem disso, as predicões teóricas ou *in silico* de interações enzimáticas obtidas nas análises podem ser testadas experimentalmente.

HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids

Journal:	*Bioinformatics*
Manuscript ID:	BIOINF-2011-1214.R1
Category:	Original Paper
Date Submitted by the Author:	17-Nov-2011
Complete List of Authors:	Gonçalves-Almeida, Valdete Maria; Federal University of Minas Gerais Pires, Douglas E.V.; Federal University of Minas Gerais Melo Minardi, Raquel; Federal University of Minas Gerais, Computer Science da Silveira, Carlos Henrique; Federal University of Itajubá Meira Jr., Wagner; Federal University of Minas Gerais Santoro, Marcelo; Federal University of Minas Gerais
Keywords:	Algorithms, Structural bioinformatics, Contacts, Clustering, Graphs, Protein-protein interaction

Dear Editor Dr. Anna Tramontano,

We appreciate the reviewers useful comments and suggestions and agreed that some explanation were missing or not well introduced. Indeed, we did our best to answer to all the criticisms and to write complete and clear explanations to their questions. We hope that our responses meet your expectations and that our manuscript can find a place at this prestigious journal. The main changes are listed bellow:

I) Answering the first referee, we have added a new section called "*Functional role of the patches*" in the supplementary material to clarify the thermodynamic basis of our choice to study conserved hydrophobic patches and to highlight their importance in cross-inhibition phenomenon. We have shown a clear trend of higher apolar / polar accessible surface area ratio toward interface (**Figure 1 of supplementary material**), which is an evidence of the importance of the hydrophobic interactions in protease-inhibitor complex formation. Furthermore, we present new figures to show the shapes and complementarity of the conserved hydrophobic patches and the apolar portions of the inhibitor loop (**Figures 3 and 4 of supplementary material**). We have also presented (**Figure 2 of supplementary material**) a strong linear relationship between the inferred solvation entropy change and the extension of hydrophobic patches, measured in terms of the number of hydrophobic atoms inside them. Finally, we have rewritten the third paragraph of the manuscript to make this point clearer and to make reference to the supplementary material that presents the complete explanation.

II) To answer the second referee suggestions, we have done several changes in figures and tables:

a – We have presented the complete parameter optimization process for the agglomerative clustering algorithm and have included a new curve in the graphs from **Figures 2 and 3**. This new graphs show the distributions of mean PRM and intra-cluster distances. It is now cleared that PRM is maximized and intra-distance values are low and stable.

b – **Figures 4 and 5** were also changed to keep a standard of ordering and coloring what makes easier to inspect and identify the enzyme's family.

c – We have also corrected all the small mistakes pointed out by this referee and they are below mentioned in detail.

III) We have decided to include in the title manuscript the name of the method.: HydroPaCe. We hope this will not be a problem. Thus the title that was:

"U*nderstanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids*"

now is:

"***HydroPaCe:*** *understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids*"

Cordially,
Dr. Valdete Almeida and co-authors.

REVIEWER 1

The topic of this manuscript could be of interest for scientists involved in protease research, because authors were able to identify sequence independent conserved regions on the serine protease surface potentially involved in interaction with inhibitors.

Unfortunately, the manuscript has major deficiencies and the goal of the study is not achieved.

Although the authors were able to solve difficult problem of identification of the conserved hydrophobic patches on the surface of serine proteases, the functional role of these patches remains unknown. Authors could try to demonstrate that the identified patches are functionally important by independent methods. This could be done by estimating the contribution of these patches to the interaction energy, and/or by providing mutagenesis studies, which would demonstrate that residues included into patches are important for protein-inhibitor interaction.

ANSWER

In the 3rd paragraph of introduction, we try to indicate the thermodynamic principles that guided our work. However, we had not been so explicit and we try to make this point clearer in the answer above (also included in the supplementary material as a new section called "*Functional role of the patches*").

Certainly, Kauzmann [1], in the final of 1950's, was the first to point out the influence of hydrophobic interactions of non-polar atoms on the large entropic effects verified experimentally in a myriad of phenomena, as for instance: the dissolution of organic molecules, protein folding, stabilization of ligand-protein complexes, protein aggregations etc.

The degree of exposure of apolar atoms or molecules to the solvent is central to the hydrophobicity concept. Lee and Richards [2] developed a famous method to compute this protein solvent exposition. They created the concept of accessible surface area (known as ASA) as the sum of all the portions of an atom or group of atoms that can be reached by a radius R sphere (representing the solvent) rolled along the protein in close contact with van der Waals surface. Chothia [3] at the beginning of 1970's, showed that there is a linear correlation among ASA and the solubility of amino acids side chain in organic solvents (a hydrophobicity measure). He estimates a free energy contribution between 20 and 30 calories mol^{-1} for each $Å^2$ of non-polar atom area not exposed to solvent.

Chothia and Janin [4] evinced that the apolar / polar ASA ratio tends to be higher for the interface in some protein complexes than for the rest of the surface. This more hydrophobic region at the interface would facilitate, in a thermodynamic perspective, the binding with apolar portions of other chains or molecules. Moreover, a balanced apolar / polar ASA ratio for non-interface regions would collaborate to prevent misaggregations. Remember that the nefarious polymerization of hemoglobin S in sickle-cell disease is induced by a disequilibrium in this ratio, due to a single mutation of one polar residue by another apolar (GLU by VAL) on the surface of beta chain [5].

In this study, we spot a clear trend of higher apolar / polar ASA ratio toward interface (**Figure 1** from this letter), which is an evidence of the importance of the hydrophobic interactions in protease-inhibitor complex formation.

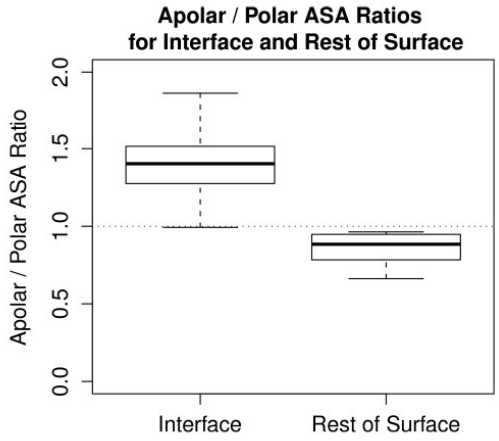


Figure 1 - Boxplot involving the Apolar / Polar ASA ratios of interface and the rest of protein surfaces for the following proteases (PDB codes): 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB

Murply and Freire [6] demonstrated that one can reliably estimate some thermodynamic parameters from apolar and polar ASA calculations. In [7], Baker and Murphy have applied this method to evaluate some of these parameters involving the binding energetics of Turkey Ovomucoid third domain inhibitor (OMTKY3) to Porcine Pancreatic Elastase (PPE). They also have performed a comparison between these empirical calculations with experimental data measured by an ITC (isothermal titration calorimetry). **Table 1** from this letter shows that the correspondence between experimental and calculated data are in agreement, except perhaps by the enthalpy change. Nevertheless, this parameter may be considered negligible in the final free energy change composition (respond only for 4% of it). Hence, we can conclude that the binding of OMTKY3 to PPE is essentially entropically driven. This indicates that the complex formation may be fundamentally guided by hydrophobic interactions.

Table 1: Comparison of thermodynamic parameters from experimental and empirical estimation for OMTKY3 / PPE complex as made in [7]. Experimental data reported at 25 ºC.

Parameter	Experimental	Calculated
ΔC_p^o (kJ K^{-1} mol^{-1})	-1.1 ± 0.1	-1.4
ΔH^o (kJ mol^{-1})	-2.5 ± 1.0	2.3
ΔS^o (J K^{-1} mol^{-1})	195 ± 4	190
ΔG^o (kJ mol^{-1})	-60.6 ± 0.5	-54

If we assume that these premises remain valid for the most of the other protease-inhibitor complex (a reasonable assumption), we can apply the methodology of Murphy and co-authors to estimate the solvation entropy change of the studied complexes through computed ASAs. We have made this choice because the entropy change is the parameter most revealing of hydrophobic interactions and it is the most preponderant entropic factor involving in protein-protein binding [7]. We can make this in two steps:

(1) Determine the heat capacity change as: $\Delta C_p = a \cdot \Delta ASA_{apolar} + b \cdot \Delta ASA_{polar}$

(2) Determine the solvation entropy change as: $$\Delta S = \Delta C_p \cdot ln(\frac{T}{Ts})$$

where, a and b are adjusted parameters estimated in [6] as 1.88 $JK^{-1}mol^{-1}A^{-1}$ e -1.09 $JK^{-1}mol^{-1}A^{-1}$, respectively; T is the room or experimental temperature and Ts is a reference temperature where entropy change is taken to be zero (about 385K) [8].

In **Figure 2** from this letter we can see that there is a strong linear relationship (Pearson correlation coefficient of 0.98) between the solvation entropy change (inferred as described bellow) and the "extension" of our patches, measured in number of hydrophobic atoms inside them. The intercept seems away from zero because the heat capacity change, in Murphy and co-authors methodology, has a polar term.

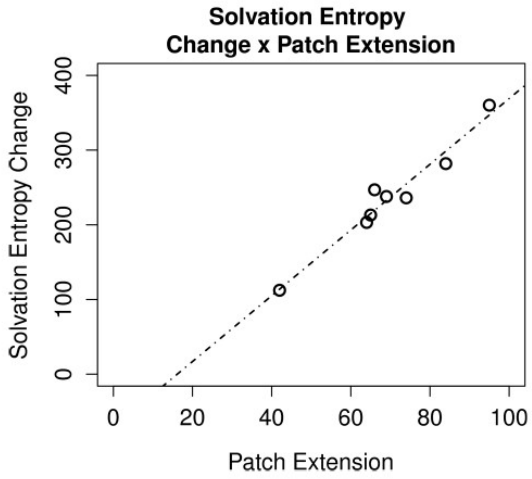


Figure 2 - Plot showing the correlation between solvation entropy change and patch extension for the following proteases (PDB codes): 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB

In conclusion, we believe that the conserved HydroPaCes have a clear functional role in the inhibition. They represent the hydrophobic network that paves (and perhaps help to codify) the active sites of proteases, going beyond the traditional catalytic triads / diads and oxyanion holes conservation. Their interfaces have a hydrophobic bias which is clearly distinct from the rest of the surface, as showed in Figure 01. The energetic dissection of Elastase PPE made by Baker and Murphy [7] indicated that the interaction between proteases and inhibitors may be fundamentally orchestrated by hydrophobic / entropic forces, and that they may represent as high as 96% of free energy change in complex formation. Corroborating these assumptions, our patches also have an evident correlation with the solvation entropy change, as showed in **Figure 2**.

We are aware that our model and algorithms are, as any heuristic, an approximation to the solution. A hydrophobic pattern may be a necessary but not sufficient condition to protease-inhibitor recognition. However, it is important to note that there is no consensus in the literature about the exact functional role of hydrophobic interactions in proteins. While Chothia [4] says that "*as hydrophobic contribution is entirely unspecific, it would lead to all kinds of incorrect interactions in a cell*", Dill [9] defends that "*hydrophobic interactions are not nonspecific glue, but a crucial structure-determining driving force*". Independent of who has the reason, our work has the merit to put empirically in evidence that robust patterns of conservations are emerging when we focus on hydrophobic network of different protease interfaces, even considering those so structurally distinct as Subtilisin-like and Trypsin-like. And we are confident that this patterns could help to explain the experimentally verified cross-inhibition phenomenon.

Even if the enzymes have conserved hydrophobic patches, it is not clear from the manuscript, that the inhibitors have complementary patches that interact with these sites and contribute to cross-inhibition.

If the binding energetics described above for protease-inhibitors is (in fact) entropically driven, the complementarity of hydrophobic sites between enzyme and ligand has to exist in some degree. Indeed, we can see in **Figure 3** from this letter an example of correspondences among the apolar atoms of the inhibitors side and enzymes side.

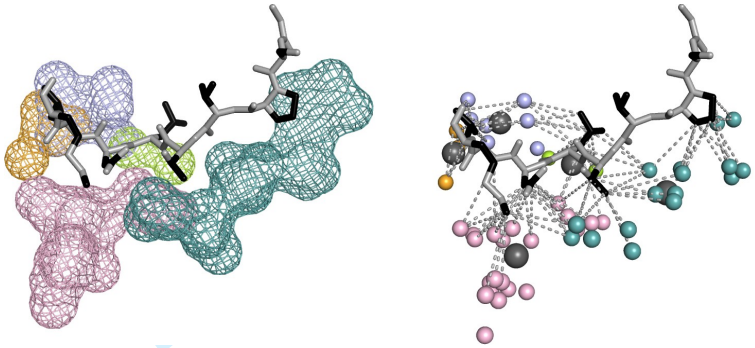


Figure 3 - Hydrophobic interactions between protease Subtilisin-like and inhibitor Eglin C. The loop of inhibitor is in close contact with the protease interface. We present polar atoms in gray and apolar atoms in black. On protease interface, the apolar regions are in meshes and spheres (those between 4 and 6 angstroms from any apolar inhibitor atoms are connected by dashed edges). Big spheres in medium gray represent the centroids.

The complete analysis of complementarity can be found in the supplementary material (**Figures 3 and 4** from supplementary material). Unfortunately, we have to enough space to include them in the manuscript.

References

[1] KAUZMANN, W. Some factors in the interpretation of protein denaturation. Adv. Prot. Chem, 14, 1-63, 1959.

[2] LEE, B.; RICHARDS, F. M. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55, 379-400, 1971.

[3] CHOTHIA, C. Hydrophobic bonding and accessible surface area in proteins. Nature. 248 (5446), 338-339, 1974

[4] CHOTHIA, C.; JANIN, J. Principles of protein-protein recognition. Nature. 256, 705-708, 1975.

[5] DICKERSON, R. E.; GEIS, I.. Hemoglobin. Menlo Park:The Benjamin/Cumming Publishing Co. Inc., 1983.

[6] MURPHY, K. P.; FREIRE, E. Thermodynamics of structural stability and cooperative folding behavior in proteins. Adv Prot Chem. 43, 313-361, 1992.

[7] BAKER, B. M.; MURPHY, K. P. Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase. J. Mol. Biol. 268, 557-569, 1997.

[8] BALDWIN, R. Temperature dependence of the hydrophobic interaction in protein folding. Proc. Natl. Acad. Sci. 83, 8069-8072, 1986.

[9] DILL, K. Polymer principles and protein folding. Protein Science. 8, 1166-1180, 1999.

REVIEWER 2

The manuscript entitled "Understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids" by Gonçalves-Almeida and coworkers describes a methodology called HydroPaCe to detect conserved patterns for understanding and predicting cross-inhibition and to clarify how it happens. To this purpose, protein-protein interfaces are modeled as graphs of atomic apolar interactions, and hydrophobic patches are computed and summarized by centroids, and their conservation is detected.

In my opinion, the method reported in this manuscript is well written and the scientific work done by the authors has been correctly carried out. However, I include below some comments that I believe should be taken into consideration by the authors.

Comments:

Main manuscript:
In the left-hand graph of Fig. 2 and 3, the number of clusters at which the intra-distance values stabilize is graphically determined. However, it is not clear from the plot in the figures if the intra-distance values actually stabilize. It would help to show in the plot the error bars corresponding to the intra-cluster distance distribution. Also, does the PRM distribution values change significantly by choosing a different number of clusters? In Fig. 2, e.g., is 12 the minimum number of clusters for which the PRM values are optimized? (i.e. the best tradeoff between the number of clusters and the PRM value.)

ANSWER

We choose the number of clusters that optimizes the mean PRM. We have computed the metric for the complete range of possible numbers of clusters and analyzed the mean PRM distribution.

Manuscript, page 5, Figures 2 and 3.
Supplementary material, page 9 (Figure 6), page 12 (Figure 8), page 14 (Figure 9) and page 16(Figure 10) and corresponding Tables 5 (page 8), 6 (page 11), 7 (page 13), 8 (page 15) and 9 (page 17).

These graphs and the other complementary analysis were added to the supplementary material.

In Table 1, the quantitative comparison of the proposed algorithms is shown. AC is chosen as the best-performing method in the coarse grained analysis, achieving the lowest intra and highest inter-cluster distances combined with a very high PRM. However, AC achieves the lowest intra and highest inter-cluster distances in the EBCC algorithm and not in the CCC one, although in this case the PRM value is higher (0.98 versus 0.90). Provided that the best performance is the one highlighted in Table 1, I think the sentence describing the algorithm comparison for Eglin C should be rephrased.

ANSWER

We agreed that there was a mistake in this description and changed it to "*AC performs better, especially in the coarse-grained analysis, achieving low intra- and high inter-cluster distances combined with a very high PRM value (0.98).*".

Manuscript, page 5, 4th paragraph.

In Fig. 4 and 5 (and 1 and 3 in the Supplementary Material), the order of the hydrophobic patches and the color of the HP-centroid should be coherent for ease of visual inspection. In Fig. 4 (and 1 in the Supplementary Material), the Trypsin-like examples are reported first followed by the Subtilisin-like ones, and the green shades are the Subtilisin-like HP-centroids while the red ones are the Trypsin-like ones. In Fig. 5 (and 3 in the Supplementary Material) is the other way around, i.e. the Subtilisin-like examples are reported first followed by the Trypsin-like

ones, and the green shades are the Trypsin-like HP-centroids while the red ones are the Subtilisin-like ones. I suggest using the same order of the hydrophobic patches in each enzyme type example and the same color to represent the HP-centroids.

ANSWER

We agreed that the use of different orders and colors for Trypsin-like and Subtilisin-like turned the manuscript a bit confusing. We then defined the shades of red to represent Trypsin-like and the shades of green, Subtilisin-like and kept their order always the same (Trypsin-like always presented first).

Manuscript, page 6, Figures 4 and 5.
Supplementary material, pages 7 (Figure 5) and page 10 (Figures 7).

Also, could the green and red spheres in the inhibitor representation shown in the above-mentioned figures have the same orientation of the respective hydrophobic patches?

ANSWER

In fact, they were already in the same orientation so we did not change the manuscript according to this comment.

The caption of Fig. 2 misses the description of the right-hand graph (as it is in Fig. 3). In the right-hand graphs in Fig. 2 and 3 the x-axis is not shown.

ANSWER

Caption was corrected with the addition of the text: "*In the right-hand graph we present the PRM distribution for the best configuration of clusters.*". Besides, we have added the x-axis in all the similar graphs.

Manuscript, page 5, Figures 2 and 3.
Supplementary material, page 9 (Figure 6), page 12 (Figure 8), page 14 (Figure 9) and page 16(Figure 10)

In the Data Set section, the definition "50% similarity" should be corrected to "50% identity".

ANSWER

Done.

Supplementary material, page 1, 1st paragraph.

"Table 1 shows the PDB ids (single chains) that were used for projections and analyses of the interfaces of the Subtilisin-like family." should be corrected to "Table 1 shows the PDB ids (single chains) that were used for projections and analyses of the interfaces of the Trypsin-like family."

ANSWER

Done.

Supplementary material, page 1, 2nd paragraph.

In the description of Table 1 and 2, the phrase "The first line" should be "The first column".

ANSWER

Done.

Supplementary material, page 1, 2nd and 3rd paragraphs.

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids

V. M. Gonçalves-Almeida [1,2*], D.E.V. Pires [1,2], R. C. de Melo-Minardi [1*], C.H. da Silveira [3], W. Meira Jr. [1] and M.M. Santoro [2*]

[1]Department of Computer Science, Universidade Federal de Minas Gerais, Brazil
[2]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Brazil
[3]Advanced Campus at Itabira, Universidade Federal de Itajubá, Brazil

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Motivation: Protein-protein interfaces contain important information about molecular recognition. The discovery of conserved patterns is essential for understanding how substrates and inhibitors are bound and for predicting molecular binding. When an inhibitor binds to different enzymes (e.g., dissimilar sequences, structures or mechanisms what we call cross-inhibition), identification of invariants is a difficult task for which traditional methods may fail.

Results: To clarify how cross-inhibition happens, we model the problem, propose and evaluate a methodology called HydroPaCe to detect conserved patterns. Interfaces are modeled as graphs of atomic apolar interactions and hydrophobic patches are computed and summarized by centroids (HP-centroids), and their conservation is detected. Despite sequence and structure dissimilarity, our method achieves an appropriate level of abstraction to obtain invariant properties in cross-inhibition. We show examples in which HP-centroids successfully predicted enzymes that could be inhibited by the studied inhibitors according to BRENDA database.

Availability: www.dcc.ufmg.br/∼raquelcm/hydropace

Contact: valdetemg@ufmg.br, raquelcm@dcc.ufmg.br and santoro@icb.ufmg.br

1 INTRODUCTION

Enzyme inhibition occurs when a molecule binds to an enzyme, thus decreasing its activity. Inhibitors may be proteic or non-proteic; they can decrease the enzyme's ability to bind substrates or can lower the enzyme's catalytic activity or a combination of both. Inhibition is an important biochemical mechanism that is involved in metabolism regulation. It controls many intra- and extra-cellular pathways, inflammatory and immunological processes, virus replication and many other biological functions (Barrett *et al.*, 2004) Furthermore, once this natural phenomenon is understood, it might be used for biotechnological purposes including the development of drugs, insecticides, pesticides and disinfectants.

A particular case is the inhibition of peptidases; on this subject, the MEROPS database is currently one of the most important peptidase repositories (Rawlings *et al.*, 2008). The MEROPS database groups both proteases and inhibitors hierarchically into families (sequence-related entities) and clans (structure-related entities). A careful MEROPS search highlighted a well-known but intriguing phenomenon: some protease inhibitors lack specificity and involve different three-dimensional structures and catalytic mechanisms. For instance, Turkey Ovomucoid and Englin C act in different serine peptidase clans such as PA(S) (all beta Trypsin-like folds) and SB (alpha/beta Subtilisin-like folds) and soybean Kunitz trypsin inhibitor decays proteolytic activity as much in serine peptidases as in metallopeptidases (which have very different enzymatic mechanisms). We call this lack of specificity *cross-inhibition*. Our main challenge in this paper is to create a methodology that helps to understand and predict this phenomenon.

Protease-inhibitor recognition and binding are determined by a complex orchestration of interactions and entropic factors that involve the entire protease-inhibitor-solvent system. Fortunately, the experimental binding energetics of many protease-inhibitor complexes have already been thermodynamically determined. It is known, for example, that the binding of Turkey Ovomucoid with Elastase at 25 °C is characterized by a negative Gibbs free energy in which enthalpy change is almost negligible but entropy change is largely positive (Baker and Murphy, 1997). Furthermore, we spot a clear trend of higher apolar / polar accessible surface area ratio toward interface (Figure 1 of supplementary material), which is an evidence of the importance of the hydrophobic interactions in protease-inhibitor complex formation. That said, we particularly focus our attention on the search for conserved hydrophobic interaction patterns. We define these patterns as invariant hydrophobic regions (or patches) that are in contact with the same apolar complementary parts of the inhibitor (Figures 3 and 4 of supplementary material). We show (Figure 2 of supplementary material) a strong linear relationship (Pearson correlation coefficient of 0.98) between the inferred solvation entropy change and the extension of hydrophobic patches, measured in terms of the number of hydrophobic atoms inside them.

Although there are many biochemical studies that analyze diversity in inhibition processes (e.g., (Laskowski and Qasim, 2000; Bode *et al.*, 1986; Qasim *et al.*, 1997; Chakrabarti and Janin,

*To whom correspondence should be addressed

2002a), experimental characterization of inhibition is a labor-intensive process. The large amount of possible inhibitors for a given enzyme can make tests costly; hence, *in silico* methods can contribute to predicting inhibitor-enzyme recognition.

Despite its evident importance, there are few models and algorithms that identify recognition and interaction patterns that could help to clarify how cross-inhibition occurs. In this context, a pattern is a conserved set of interface attributes that is used to explain or predict binding.

Traditionally, sequence comparison and/or structural alignment methods have been used in conservation detection (Zhang *et al.*, 2011; Ribeiro *et al.*, 2010; Melo-Minardi *et al.*, 2007). According to (Tuncbag *et al.*, 2011), structures are more conserved than sequences, and interface-forming residues (IRFs) are even more conserved than the whole structure. However, these classical methods are inappropriate because in cross-inhibition we may deal with very dissimilar sequences and even completely distant folds.

Indeed, in cross-inhibition pattern detection with traditional methods, we identify essentially known conserved residues that directly participate in the catalysis process, such as the catalytic triad, the specificity pocket and oxyanion-binding sites. We note that to correctly assess the eventual hydrophobic contribution of the entire protease-inhibitor interface, we should abstract the residue semantics and should assess patches at the atomic level. A similar approach has been used to characterize the core of protein domains with similar folds but very divergent sequence compositions (Soundararajan *et al.*, 2010). The atomic level is more appropriate because all residues have apolar portions. Lysine, for example, is considered a positively charged residue (at neutral pH), but there are also several hydrophobic methyl groups.

Enzyme-inhibitor recognition is determined by a network of interactions between atoms; hence, graph modeling is a straightforward approach. We model hydrophobic atoms as nodes of a graph and the contacts between them as the edges. We use the graph to obtain conserved hydrophobic patches or, in other words, connected components.

Supposing that the most important property of a hydrophobic patch is where it is positioned to interact with the ligand, we abstract from its composition volume, shape and density, and we represent the patch as a geometric centroid that we call the HP-centroid (hydrophobic patch centroid). In this work, we propose a novel model and algorithms to detect conserved HP-centroids in cross-inhibition.

Finally, we present a qualitative case study that consists of two examples of cross-inhibition, Trypsin-like and Subtilisin-like enzymes, both of which belong to the serine proteases family. They present completely different three-dimensional structures and the sequence identity is as low as 20% (Wallace *et al.*, 1996). However, they possess exactly the same Ser-His-Asp triad on their active sites. In the first case, we have complexes of Trypsin-like and Subtilisin-like enzymes inhibited by Eglin C (Betzel *et al.*, 1993), and in the second case, we have complexes of the same families with Turkey Ovomucoid (Papamokos *et al.*, 1982). We verify that the HP-centroids obtained from the complexes are present in a set of sequence-diverse apo structures that are conserved throughout the family.

2 MATERIALS AND METHODS

Each step of the proposed methodology, called HydroPaCe, is described below. A complete workflow of the methodology is presented in Figure 1.

2.1 Data selection and preparation

As explained previously, we have chosen serine proteases to test our algorithm. We chose them because there are few other examples of cross-inhibition structures in the PDB (Berman *et al.*, 2000). Moreover, this is a well-studied family that presents some peculiarities and similarities in catalytic sites (Page and Di Cera, 2008). Although Trypsin-like and Subtilisin-like have very different three-dimensional structures, they hydrolyze their substrates by the same mechanism (Ekici *et al.*, 2008; Lesk and Fordham, 1996; Siezen and Leunissen, 1997).

Enzyme-inhibitor complexes: We found five non-redundant complexes involving the Eglin C inhibitor: four bond to Subtilisin-like (PDB IDs: 1TEC, 1CSE, 1MEE and 1SBN) and one to Trypsin-like (PDB ID: 1ACB) enzymes. Likewise, we found four complexes involving the Ovomucoid inhibitor: three complexed with Trypsin-like (PDB IDs: 1CHO, 1PPF and 3SGB) and one with Subtilisin-like (PDB ID: 1R0R) enzymes. Despite the large amount of information on enzymatic complexes involving these two families, there is much redundant information regarding the sequence identities, and this leaves only a small number of non-redundant complexes to be analyzed.

Apo enzymes: We selected a set of non-redundant apo enzymes from the two families by removing enzymes that presented greater than 50% of sequence identity. Hence, we use 9 samples from Subtilisin-like and 35 from Trypsin-like families. The complete list of PDB ids is presented in the supplementary material.

All of the structures were submitted to standardization processes using the PDB Enhanced Structures Toolkit (PDBest) (Pires *et al.*, 2007).

2.2 Interface-forming residues

The current analysis is restricted to regions of the molecular interface of the enzyme and its inhibitor. The interface-forming residues (IFRs) can be determined by three different methods. The first defines the interface simply by using a cut-off distance between the residues of the interacting molecules (Chothia and Janin, 1975; Conte *et al.*, 1999). The second approach computes the interactions based on differences in solvent-accessible surface area (ASA) when the monomers are separated (Janin *et al.*, 1990; Chakrabarti and Janin, 2002b). Finally, the last approach defines interfaces through computational geometry using Voronoi diagrams and the alpha shapes theory (Pontius *et al.*, 1996). We used the ASA method because it is the most used method and is therefore more consolidated.

Enzyme-inhibitor interface-forming residues (IFRs): We computed the IFRs in the cross-inhibition complexes using the ASA approach with the STING Millennium Suite platform (SMS) (Neshich *et al.*, 2003).

Projection of IFRs from complexes into apo enzymes: For the apo proteins, the projection was derived by structural alignment using an enzyme-inhibitor complex and the computed IFR. Moreover, the structures were solvated using Gromacs. After applying the treatment to PDB files, all structures, including the complex model, were superimposed using the program MultiProt. Finally, the residues that aligned with the interface of the complex model were considered the interfaces of the apo proteins. This process was performed for analysis of both sets (Trypsin-like and Subtilisin-like) of single-chain proteins in our database.

2.3 Problem modeling

The proposed method is based on the search for conserved hydrophobic patches (HP-centroids). In what follows, we detail each step of our model:

Graph construction: The first step of our model consists of the representation of IFRs as graphs. The nodes are atoms from the IFR residues, and the edges are the presumed contacts. According to our previous work (da Silveira *et al.*, 2009), there are two main approaches to identify contacts in proteins: the first is cut-off dependent (CD), and the other is

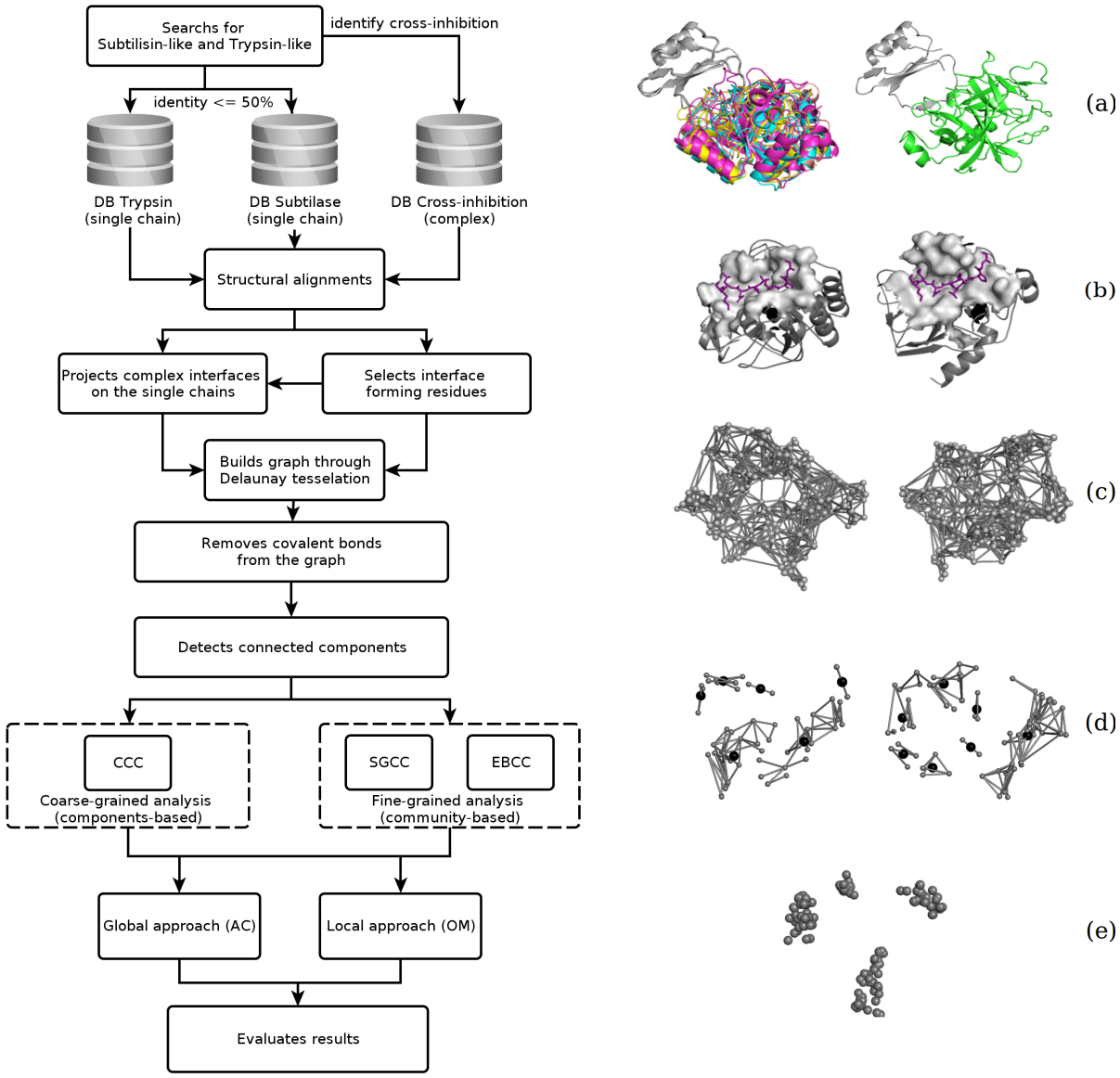


Fig. 1. HydroPaCe workflow: We searched for three-dimensional structures of subtilisin-like and trypsin-like families in the PDB database. The PDB ids were separated into protein-inhibitor complexes and apo proteins. The structures with sequence identities that were more than 50% identical to other selected sequences were discarded. The cross-inhibition complexes were aligned by the inhibitor's chain (a), and the interfaces of contact (also called interface-forming residues, IFRs) were identified using ASA methodology (b). The apo proteins were aligned by their single chains by using an enzyme-inhibitor complex to project the interface. Using the Delaunay Tessellation, possible interatomic contacts were computed, resulting in the edges of a graph where nodes are atoms (c). We considered only the hydrophobic interactions between atoms and removed edges that represented covalent bonds. We then identified the connected components that represent the hydrophobic patches in these graphs (d). We propose two levels of abstractions to represent the hydrophobic patches, both of which are based on geometric centroids (HP-centroids). The first is a coarse-grained analysis that consists of computing a centroid for each connected component, and the second is a fine-grained analysis that searches for dense sub-regions using two different community detection algorithms and calculating the HP-centroids on communities. The obtained HP-centroids were clustered using the OM and AC methods (e). Finally, the HP-centroids were evaluated using the Penalized Recall Metric, which accounts positively for coverage in terms of enzymes and negatively for enzyme redundancy. In a, b, c and d, the left-hand structure is Subtilisin-like and the right-hand structure is Trypsin-like.

independent (CI). Although in the above-mentioned study we found that, at the residue level, the CD approach was a simpler, more complete, and more reliable technique than some CI techniques, here we chose to use a cut-off-independent methodology because we did not find a reliable cut-off value at the atomic level. This paradigm uses classical computational geometry algorithms to compute a Voronoi diagram (VD) (Poupon, 2004) and its dual problem, the Delaunay tessellation (DT) (Dupuis *et al.*, 2005). In the three-dimensional view, the VD decomposes the volume by associating a polyhedron with each site (which is called a Voronoi cell). Each face of these polyhedrons is comprised of a plane that bisects the line and links each site to each of its near sites, thus mapping a neighborhood with the closest (not occluded) contacts (da Silveira *et al.*, 2009).

Deletion of covalent edges: We are interested only in non-covalent interactions; hence, we remove covalent bond edges in a post-processing step.

Deletion of polar edges: Once we have a geometrical inference of non-occluded interactions, we classify them into hydrophobic and polar interactions based on the classification rules proposed in (Sobolev *et al.*, 1999). The complete table with the classifications of all the atoms can be found in the supplementary material. As discussed previously, we restrict our analysis to hydrophobic interactions type by removing polar contact edges. Nevertheless, the analysis can be extended to deal with polar areas.

Computation of hydrophobic patches: We use a depth-first search to efficiently detect the connected components, which are natural representations of the hydrophobic patches.

Abstraction of hydrophobic patches through centroids: Hydrophobic patches may occur in different shapes and volumes; our model considers two levels of abstractions to represent them, both of which are based on geometric centroids (HP-centroid). The first, which we call the coarse-grained analysis, consists of computing a centroid for each connected component. The second is a fine-grained analysis that divides the original connected components into dense subgraphs, or communities. A community is a subgraph in which the nodes are much more connected with the other nodes in the community than with the external nodes. In this approach, the HP-centroids are computed based on communities.

In conclusion, our method is based on the computation of hydrophobic patches and their abstraction through geometric centroids (HP-centroids) that can represent the entire patch (coarse-grained) or communities of these patches (fine-grained). Considering the HP-centroids of a set of cross-inhibition complexes, we propose algorithms to cluster the centroids and to detect those that are conserved across all of them. We describe the algorithms in the next section and then explain how to evaluate the clusters obtained.

2.4 Algorithms

Here, we describe in more detail the different approaches (coarse- and fine-grained) that we propose to abstract from the hydrophobic patches. We briefly describe the paradigms for community detection used in fine-grained decomposition of hydrophobic patches. Finally, we explain the algorithms that we use to cluster the HP-centroids: one attempts to globally match similar centroids and the other locally clustered centroids in an agglomerative manner.

CCC: *Connected Component Centroids* is the name we give to the coarse-grained approach.

EBCC: The *Edge Betweenness Community Centroid (EBCC)* (Newman and Girvan, 2004) is a divisive approach in which the most central edges are broken one after another until the modularity of the graph is maximized. The edge centrality is computed through the edge betweenness, which counts the number of shortest paths that traverse through that edge. The higher the value of edge betweenness, the more the edge is used or the more central it is. In other words, this value indicates when there are no redundant edges to cross between different communities and when the edge joins two different communities.

SGCC: The *Spin Glass Community Centroid (SGCC)* (Reichardt and Bornholdt, 2006) tries to find communities in graphs via a spin-glass model and simulated annealing. That is, it uses simulated annealing to maximize graph modularity. The modularity of a possible division of a graph into communities is defined as the fraction of edges that falls within a given community minus the expected value of this fraction if edges were randomly distributed. Commonly, the randomization of the edges is done in such a way as to preserve the degree of each vertex.

OM: We have developed a linear programming *Optimization Model (OM)* that is based on the transport problem and that attempts to match points by globally minimizing the differences between the edge sizes between all possible pairs of points. The optimization functions that we want to minimize, as well as the associated restrictions, are explained in detail in the supplementary material.

AC: This method is a local strategy based on *Agglomerative Clustering (AC)*. It matches the closest HP-centroids through an iterative bottom-up agglomerative process. In this case, there is an important decision about when to stop the process to ensure that we have high-quality clusters. The strategy for determining this stopping point, and a detailed explanation of the algorithm, are presented in the supplementary material.

2.5 Evaluation

To perform a quantitative evaluation of the clusters formed by the matches, we propose a metric based on the concept of recall that is penalized when different HP-centroids of the same protein (redundant centroids) are grouped together. We have it called the Penalized Recall Metric (PRM) and is formalized below:

$$PRM = \frac{\mathbb{C}\binom{D}{2}}{\mathbb{C}\binom{P}{2}} - \frac{\mathbb{C}\binom{E}{2}}{\mathbb{C}\binom{P}{2}} \quad (1)$$

where $\mathbb{C}\binom{D}{2}$ is the number of pairs of centroids from different enzymes in the same cluster, $\mathbb{C}\binom{E}{2}$ is the number of pairs of HP-centroids from the same protein in the same cluster, $\mathbb{C}\binom{P}{2}$ is the total number of pairs of HP-centroids in the cluster and the values of D and E are limited to P.

The metric produces values in the range of [-1; 1] where -1 is the worst case, with minimum recall and maximum redundancy. It will result in 1 when we have maximum recall and minimum redundancy. When we have similar values for recall and redundancy, the metric approaches 0.

The average of the penalized recall metric of the clusters was used to evaluate the three different approaches (CCC, EBCC and SGCC). It cannot be used to compare between the OM and AC. In the OM, clusters are formed with total variability by definition; in other words, there are no pair of centroids of the same enzyme in a cluster.

In this case, we use traditional intra- and inter-cluster average distances. A priori, a high quality cluster must have low intra-cluster and high inter-cluster distances. That is because, in an ideal clustering, similar elements must be grouped together and dissimilar ones must be separated.

In conclusion, we compare the proposed approaches in the light of the PRM (the closer to 1, the better) and the average intra-and inter-cluster distances (it is better to have low intra-cluster and high inter-cluster distances).

3 RESULTS

In this section, we present and discuss the results of the case study of serine peptidases (Trypsin-like and Subtilisin-like) that are cross-inhibited by Eglin C and Turkey Ovomucoid. We also compare the quality of the conserved HP-centroids that are obtained by the different proposed methods.

3.1 The Eglin C Inhibitor

Eglin C is a small monomeric protein (70 residues) that belongs to the Potato Chymotrypsin Inhibitor I family of serine protease inhibitors that occurs naturally in the Leech *Hirudo medicinalis*.

Functionally, Eglin C can inhibit more than one proteinase family with non-homologous structures (Hyberts *et al.*, 1992). In the BRENDA database (Scheer *et al.*, 2011), we found 12 different EC numbers that are known to be inhibited by this molecule. In this section, we present the analysis with the 5 non-redundant existing experimental complexes, 4 of which are Subtilisin-like and 1 of which is Trypsin-like.

As explained previously, we use different approaches to find the HP-centroids. The OM has no parameters and it clusters all of the centroids. With the AC, we must supply the number of clusters as an input parameter. Figure 2 (left-hand graph) shows the distributions of mean PRM and intra-cluster distances. We observe that PRM is maximized and intra-distance values are stable with 12 clusters. With this configuration, we obtain five high-quality clusters according to the PRM (right-hand graph). This set of conserved HP-centroids presents a very high recall value (i.e., they are present in almost all of the cross-inhibition complexes) and furthermore, there is only one case where two points in a cluster come from the same complex.

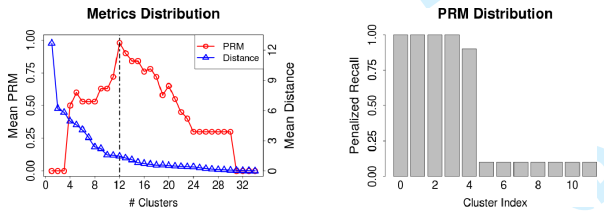


Fig. 2. The CCC approach. In the left-hand graph we present the distribution used to maximize the mean PRM metric as well as the respective mean intra-cluster distance distribution. The right-hand graph shows the PRM distribution for the best configuration achieved, with 12 clusters.

The same experiment was performed using the fine-grained approach, as presented in Figure 3. At this level of abstraction, we could not identify a threshold that clearly distinguishes high-quality clusters from poor-quality ones. Because we aim to find as many conserved HP-centroids as possible, the coarse-grained approach systematically presents better results. This might indicate that the cross-inhibition pattern depends on the inhibitor-relative positions of the conserved HP-centroids regardless of their density.

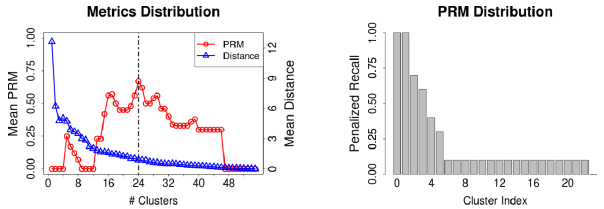


Fig. 3. The SGCC approach. In the left-hand graph we present the distribution used to maximize mean PRM metric as well as the respective mean intra-cluster distance distribution. The right-hand graph shows the PRM distribution for the best configuration achieved, with 24 clusters.

Table 1 shows the complete set of results. AC performs better, especially in the coarse-grained analysis, achieving low intra- and high inter-cluster distances combined with a very high PRM value (0.98).

Table 1. Quantitative comparison of the proposed algorithms for Eglin C cross-inhibition.

		Mean Intra (Å)	Mean Inter (Å)	Mean PRM
CCC	AC	3.435	13.835	**0.98**
	OM	5.460	9.294	-
SGCC	AC	2.450	13.138	0.82
	OM	5.093	11.231	-
EBCC	AC	2.679	12.670	0.90
	OM	5.339	9.986	-

The semantics of the five hydrophobic patches represented by the conserved HP-centroids is presented in Figure 4. We can see why the proposed method reaches an abstraction level that is useful for identifying relevant cross-inhibition patterns. When we compare the residues that compose cluster IV, we can see for a Trypsin-like enzyme the presence of LEU-143, THR-151, ALA-149, TYR-146, CYS-220, CYS-191 and MET-192. At the counterpart cluster in a Subtilisin-like enzyme, we find PHE-193, ASN-163 and THR-224. Despite the very dissimilar residue compositions, patch volumes and densities, the method selects HP-centroids that are spatially conserved according to the inhibitor. Additional graphs for the other 3 samples are presented in the supplementary material.

3.2 The Turkey Ovomucoid Inhibitor

Ovomucoids are the Glycoprotein Protease-inhibitors of avian egg whites. There are several Protease inhibitors in egg white. The Turkey Ovomucoid is from a Kazal-type inhibitor family of serine protease inhibitors that occurs naturally in Meleagris gallopavo. It is a significant contaminant of crude Ovomucoid preparations, and it acts on Bovine Trypsin and Chymotrypsin as well as on Porcine Elastase and Fungal Proteinase (Robertson *et al.*, 1988; Fujinaga *et al.*, 1987).

We analyze the four non-redundant existing complexes, of which three have Trypsin-like enzymes and one has Subtilisin-like enzymes. By conducting similar experiments to those presented in the previous section, and by varying the number of clusters, we observe that the mean intra-distance stabilizes from 4 clusters on. We obtain three high-quality clusters according to the PRM (supplementary material).

Table 2 shows the results for the algorithm comparisons. As in the previous analysis, AC presents a combination of low intra-cluster distances, high inter-cluster distances and the highest PRM value (0.94) indicating a consistent match of the patterns.

According to these results, the coarse-grained approach once more achieved better results than the fine-grained approach.

The three hydrophobic patches that were conserved in the Ovomucoid complexes are presented in Figure 5. Again, we can see a very dissimilar cluster composition and an interesting conservation of position according to the common inhibitor. We

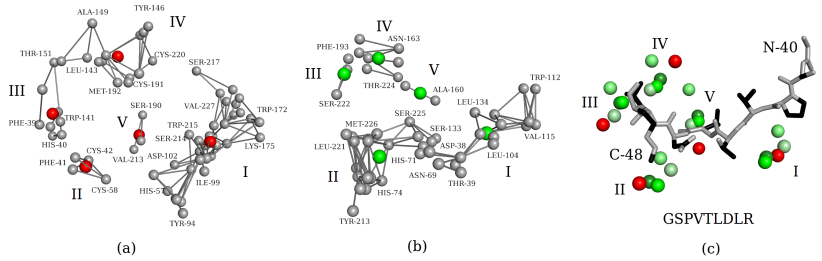


Fig. 4. Hydrophobic patches for cross-inhibition by Eglin C. In (a) PDB id 1ACB:E, we can see a sample with Trypsin-like enzyme and in (b) PDB id 1TEC:E, with Subtilisin-like enzyme (the hydrophobic patches for the five complexes are in the supplementary material). We show an atomic graph in which the residue types and numbers are presented and the red (a) and green (b) spheres are the HP-centroids that represent each of the patches. In the last part of the figure (c), we present the inhibitor (residues from 40 to 48) as gray sticks (in black, the apolar portions), and the five centroids are superposed in colors. The green shades are the Subtilisin-like HP-centroids and the red ones are Trypsin-like.

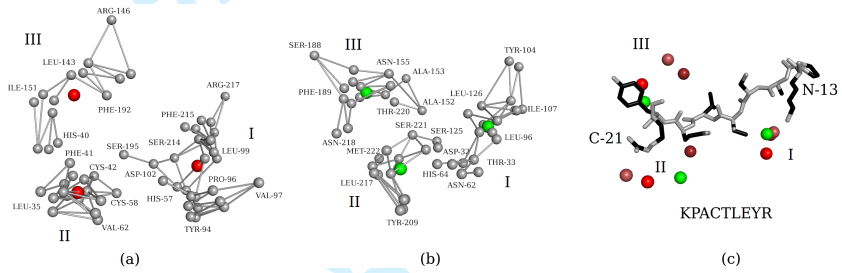


Fig. 5. Hydrophobic patches for cross-inhibition by Turkey Ovomucoid. In (a) PDB id 1R0R:E, we can see a sample with Subtilisin-like enzyme and in (b) PDB id 1PPF:E, a sample with Trypsin-like enzyme (the hydrophobic patches for the four complexes are shown in the supplementary material). We show an atomic graph in which the residue types and numbers are presented and the red (a) and green (b) spheres show the HP-centroids that represent each patch. In the last part of the figure (c), we present the inhibitor (residues from 13 to 21) as gray sticks (apolar portions in black), and the five HP-centroids are superposed in colors. The green shades are the Subtilisin-like centroids and the red ones are the Trypsin-like centroids.

Table 2. Quantitative comparison of the proposed algorithms for Turkey Ovomucoid cross-inhibition.

		Mean Intra (Å)	Mean Inter (Å)	Mean PRM
CCC	AC	4.803	13.239	**0.94**
	OM	8.009	10.402	-
SGCC	AC	2.901	10.999	0.75
	OM	9.303	11.014	-
EBCC	AC	3.419	14.045	0.75
	OM	6.459	11.997	-

present additional graphs for Ovomucoid cross-inhibition in the supplementary material.

According to (Baker and Murphy, 1997), hydrophobic interactions are essential for explaining how inhibition happens in proteases. Our results are in agreement with this hypothesis. Searching for conserved abstractions of hydrophobic patches at the atomic level (HP-centroids) in protease-inhibitor interfaces, we proposed and evaluated a global and a local algorithm to cluster centroids. We aimed to find conserved centroids at coarse- and fine-grained levels. We conclude that the coarse-grained AC local algorithm was able to identify the more complete set of invariant HP-centroids across the protease-inhibitor cross-inhibition examples.

Certainly, the contribution of polar interactions must be studied in more detail in future work; interestingly, however, we have found a minimum of three invariant centroids in all cross-inhibition cases. As proteins are three-dimensional objects, we conjecture that for a molecule to bind and to hold another one, there must exist at least three non-collinear contact points. It is possible that the conserved hydrophobic patches obtained are responsible for binding and holding inhibitors at the enzyme binding sites.

3.3 The use of HP-centroids for inhibition prediction

Once we have the problem of scarcity of experimental complexes representing cross-inhibition examples, it is intriguing to ask whether we can generalize the conserved HP-centroids to binding sites of apo enzymes of the studied families. We extended the analysis to a set of non-redundant apo structures of serine proteases (a list of proteins is in the supplemental material). We project the IFR obtained from the cross-inhibition complexes to the apo enzymes by using structural alignments, and we verify a strong conservation of the HP-centroids found through complex analysis.

Due to the low conservation of residues, it is not possible to understand how inhibition occurs by examining only sequence-level conservation (even when sequence alignments are done by structural alignments as shown in Figure 6). Notice that we can

find some conserved residues (marked with *) that are known to participate in the catalysis (catalytic triad, oxyanion role) or in the specificity binding sites. Apart from these residues, no other interest conservation can be easily identified in these logos.

However, our hypothesis is that for inhibition to occur, we must have very conserved hydrophobic patches in specific positions to accommodate each of the inhibitors. For example, PHE-215 in the Trypsin-like enzymes in Figures 6(b) and 5(b) is a voluminous hydrophobic residue that is equivalent to the hydrophobic portions of residues in positions LEU-96, ILE-107 and LEU-126 in the Subtilisin-like enzymes in Figures 6(a) and 5(a). This is an example in which conserved patterns cannot be inferred from the sequence or structure but are clearly identified in our conserved HP-centroid I.

Going further, we believe that these patterns could be used to predict inhibition for other enzymes for which structures are available but no experimental evidence of inhibition is known. For instance, we used 8 samples of non-redundant Subtilisin-like apo enzymes (listed in the supplemental material) belonging to 5 different EC numbers (3.4.21.62 / 64 / 66 / 75 / 97). We considered only those enzymes for which the ECs are complete with the four levels of annotation. According to the BRENDA database, 3 of these are inhibited by Eglin C (3.4.21.62 / 66 / 75), and we can say that this constitutes successful predictions. As far as we are concerned, the other two enzymes (3.4.21.64 / 97), which represent Proteinase K and Assemblin Protease, are not mentioned in the literature but present the same pattern as do the other Subtilisin-like enzymes. It would be very interesting to verify experimentally whether they can be inhibited by Eglin C, as they present the same HP-centroids as do other complexes with this inhibitor. Similar analyses for Ovomucoid are presented in the supplemental material, and we can also verify successful predictions and several unknown inhibition possibilities.

4 CONCLUSIONS

In this work, we model the problem of understanding and predicting enzyme cross-inhibition. We propose and evaluate algorithms to detect conserved hydrophobic patch centroids (HP-centroids) to clarify how these centroids occur in proteases. Our model is based on the importance of apolar interactions to inhibition in this family and on the fact that these hydrophobic portions should be studied at an atomic level. We model the interfaces between enzymes and inhibitors as graphs of atomic apolar interactions, detect connected components to represent hydrophobic patches, summarize them using centroids and show how to obtain as complete a set of conserved centroids as possible. One of the strengths of the method is that it achieves the appropriate level of abstraction to detect the invariant properties involved in cross-inhibition. One of the main difficulties in the study and understanding of this complex phenomenon through classical methods is that dissimilar sequences and structures might be inhibited by the same inhibitor. Despite the lack of conservation at the sequence and structure levels, the proposed HP-centroids appear to be promising, as they are very conserved across the studied cases of cross-inhibition.

As we have few non-redundant experimental complexes available, we test the generality of HP-centroids with a set of non-redundant apo enzymes representing entire families. By comparing with experimental data available in the BRENDA database, we also show some successful examples of how HP-centroids can be used to predict enzymes that could be inhibited by the studied inhibitors. Finally, we raise some questions about possible enzymes that might be inhibited by Eglic C and/or Turkey Ovomucoid and expose them to further experimental validation.

We believe that this work should be extended to other enzyme families for which entropic changes are known to be important factors in inhibition processes. It would also be interesting to verify whether this method should be used in other problems of protein-protein interaction pattern detection.

ACKNOWLEDGEMENTS

This work was supported by the Brazilian agencies Coordenao de Aperfeioamento de Pessoal de Nvel Superior (CAPES), Conselho Nacional de Desenvolvimento Cientco e Tecnolgico (CNPq), Fundao de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) and Financiadora de Estudos e Projetos (FINEP).

REFERENCES

Baker, B. M. and Murphy, K. P. (1997). Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase. *Journal of Molecular Biology*, **268**(2), 557–69.

Barrett, A. J., Rawlings, N. D., and Woessner, J. F., editors (2004). *Handbook of Proteolytic Enzymes*, volume 1-2. Elsevier, 2 edition.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, **28**(1), 235–242.

Betzel, C., Dauter, Z., Genov, N., Lamzin, V., Navaza, J., Schnebli, H. P., Visanji, M., and Wilson, K. S. (1993). Structure of the proteinase inhibitor eglin c with hydrolysed reactive centre at 2.0 a resolution. *FEBS Letters*, **317**(3), 185–188.

Bode, W., Wei, A. Z., Huber, R., Meyer, E., Travis, J., and Neumann, S. (1986). X-ray crystal structure of the complex of human leukocyte elastase (pmn elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO Journal*, **5**(10), 2453–2458.

Chakrabarti, P. and Janin, J. (2002a). Dissecting protein-protein recognition sites. *Proteins*, **47**(3), 334–343.

Chakrabarti, P. and Janin, J. (2002b). Dissecting protein-protein recognition sites. *Proteins Structure Function and Genetics*, **47**(3), 334–343.

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, **256**(5520), 705–708.

Conte, L. L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, **285**(5), 2177–2198.

da Silveira, C. H., Pires, D. E. V., Melo-Minardi, R. C., Ribeiro, C., Veloso, C. J. M., Lopes, J. C. D., Meira Junior, W., Neshich, G., Ramos, C. H. I., Habesch, R., and Santoro, M. M. (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins*, **74**(3), 727–743.

Dupuis, F., Sadoc, J. F., Jullien, R., Angelov, B., and Mornon, J. P. (2005). Voro3d: 3d voronoi tessellations applied to protein structures. *Bioinformatics*, **21**(8), 1715–1716.

Ekici, O. D., Paetzel, M., and Dalbey, R. E. (2008). Unconventional serine proteases: variations on the catalytic ser/his/asp triad configuration. *Protein Science*, **17**(12), 2023–2037.

Fujinaga, M., Sielecki, A. R., Read, R. J., Ardelt, W., Laskowski, M., and James, M. N. (1987). Crystal and molecular structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 a resolution. *Journal of Molecular Biology*, **195**(2), 397–418.

Hyberts, S. G., Goldberg, M. S., Havel, T. F., and Wagner, G. (1992). The solution structure of eglin c based on measurements of many noes and coupling constants and its comparison with x-ray structures. *Protein Science*, **1**(6), 736–751.

Janin, J., Chothia, C., Shabb, J. B., Ng, L., Corbin, J. D., Butikofer, P., Lin, Z. W., Chiu, D. T., Lubin, B., Kuypers, F. A., *et al.* (1990). The structure of protein-protein recognition sites. *Structure*, **265**(27).

Laskowski, M. and Qasim, M. A. (2000). What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochimica Biophysica Acta*, **1477**(1-2), 324–37.

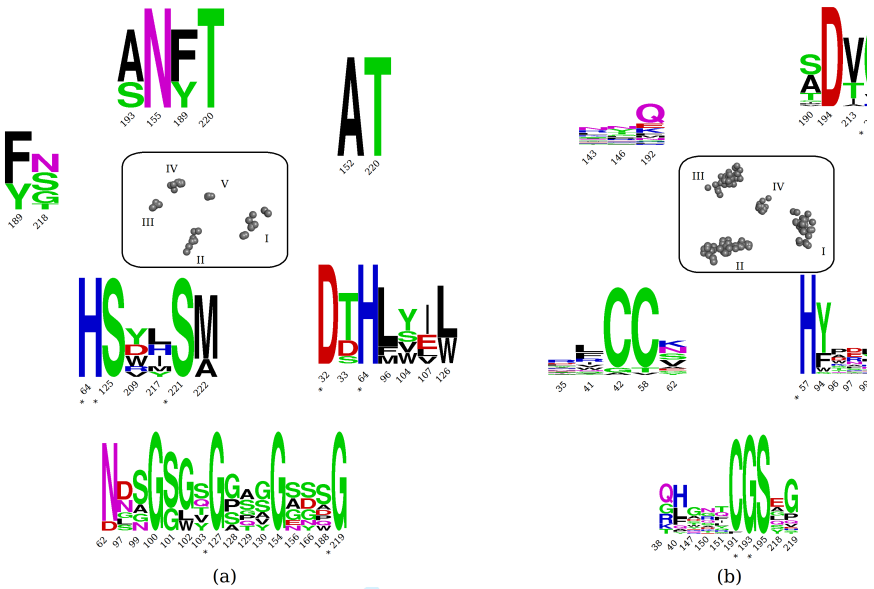


Fig. 6. IFR projections of HP-centroids found in serine proteases that are cross-inhibited by Ovomucoid. In (a), we show results for 9 non-redundant superposed Subtilisin-like enzymes (residue numbers according to PDB id 1R0R:E) and in (b) we show results for 35 non-redundant superposed Trypsin-like enzymes (residue numbers according to PDB id 1PPF:E). On both sides, the bottom logos show the residues that are in the IFR but that are not part of a conserved cluster.

Lesk, A. M. and Fordham, W. D. (1996). Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *Journal of Molecular Biology*, **258**(3), 501–537.

Melo-Minardi, R. C., Ribeiro, C., Murray, C. S., Veloso, C. J. M., da Silveira, C. H., Neshich, G., Meira Junior, W., Carceroni, R. L., and Santoro, M. M. (2007). Finding protein-protein interaction patterns by contact map matching. *Genetics and Molecular Research*, **6**, 946–963.

Neshich, G., Togawa, R. C., Mancini, A. L., Kuser, P. R., Yamagishi, M. E. B., Pappas, G., Torres, W. V., *et al.* (2003). Sting millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Research*, **31**(13), 3386.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**(2), 026113.

Page, M. J. and Di Cera, E. (2008). Serine peptidases: classification, structure and function. *Cellular and Molecular Life Sciences*, **65**(7-8), 1220–1236.

Papamokos, E., Weber, E., Bode, W., Huber, R., Empie, M. W., Kato, I., and Laskowski, M. (1982). Crystallographic refinement of japanese quail ovomucoid, a kazal-type inhibitor, and model building studies of complexes with serine proteases. *Journal of Molecular Biology*, **158**(3), 515–537.

Pires, D. E. V., da Silveira, C. H., Santoro, M. M., and Meira Junior, W. (2007). Pdbest: Pdb enhanced structures toolkit. *In Proceedings of the 3rd International Conference of Brazil Association for Bioinformatics. São Paulo: AB3C Publishing*, page 39.

Pontius, J., Richelle, J., and Wodak, S. J. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of Molecular Biology*, **264**(1), 121–136.

Poupon, A. (2004). Voronoi and voronoi-related tessellations in studies of protein structure and interaction. *Current Opinion in Structural Biology*, **14**(2), 233–241.

Qasim, M. A., Ganz, P. J., Saunders, C. W., Bateman, K. S., James, M. N., and Laskowski, M. (1997). Interscaffolding additivity. association of p1 variants of eglin c and of turkey ovomucoid third domain with serine proteinases. *Biochemistry*, **36**(7), 1598–1607.

Rawlings, N. D., Morton, F. R., Kok, C. Y., Kong, J., and Barrett, A. J. (2008). Merops: the peptidase database. *Nucleic Acids Research*, **36**(Database issue), D320–325.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, **74**(1), 016110.

Ribeiro, C., Togawa, R. C., Neshich, I. A. P., Mazoni, I., Mancini, A. L., Melo-Minardi, R. C., da Silveira, C. H., Jardine, J. G., Santoro, M. M., and Neshich, G. (2010). Analysis of binding properties and specificity through identification of the interface forming residues (ifr) for serine proteases in silico docked to different inhibitors. *BMC Structural Biology*, **10**, 36.

Robertson, A. D., Westler, W. M., and Markley, J. L. (1988). Two-dimensional nmr studies of kazal proteinase inhibitors. 1. sequence-specific assignments and secondary structure of turkey ovomucoid third domain. *Biochemistry*, **27**(7), 2519–2529.

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., and Schomburg, D. (2011). BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, **39**, 670–676.

Siezen, R. J. and Leunissen, J. A. (1997). Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Science*, **6**(3), 501–523.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**(4), 327–332.

Soundararajan, V., Raman, R., Raguram, S., Sasisekharan, V., and Sasisekharan, R. (2010). Atomic interaction networks in the core of protein domains and their native folds. *PLoS One*, **5**(2), e9391.

Tuncbag, N., Gursoy, A., and Keskin, O. (2011). Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. *Physical Biology*, **8**(3), 035006.

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to ser-his-asp catalytic triads in the serine proteinases and lipases. *Protein Science*, **5**(6), 1001–1013.

Zhang, Z., Li, Y., Lin, B., Schroeder, M., and Huang, B. (2011). Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**(15), 2083–2088.

Anexo B

Material Suplementar

O material suplementar é um complemento das informações contidas no artigo. Diversos gráficos, figuras e tabelas são exibidos para que o leitor possa acompanhar todos os detalhes.

É possivel notar uma seção dedicada para explicar a escolha da buscar por padrões hidrofóbicos o qual evidencia a importância das interações hidrofóbicas para formação dos complexos protease e inibidor.

A complementaridade dos *patches* são exibidas para todos os complexos. É possivel constatar que todos os complexos analisados possui correspondência com átomos hidrofóbicos da alça do inibidor. A grande maioria dos átomos envolvidos estão à uma distância entre 4 e 6 angstroms.

A figura B.1 mostra um exemplo desta complementaridade, destacando os *patches* em cores distintas.

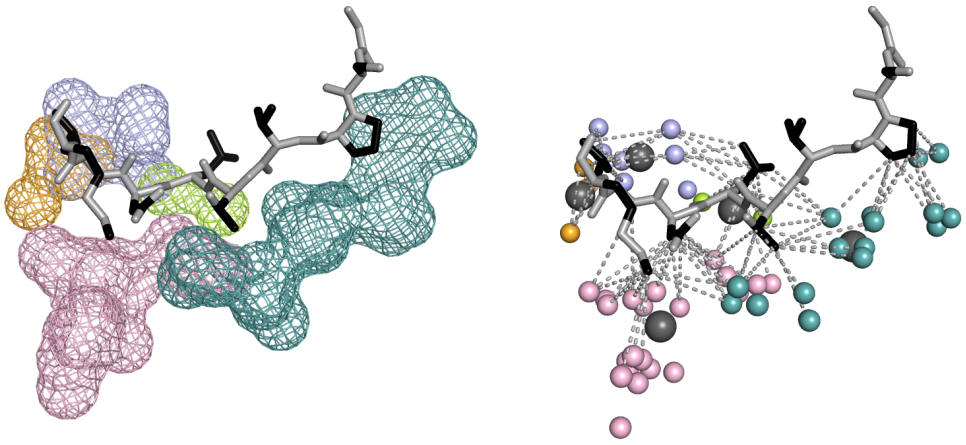


Figura B.1: Exemplo de complementaridade- complexo 1TEC

SUPPLEMENTARY MATERIAL

1 DATA SET

The data set was selected using the PDB advanced search interface and the expressions Trypsin-like and Subtilisin-like. We restricted the data to sequences presenting no more than 50% sequence identity. The structures are grouped into complexes with cross-inhibition occurrences and single chains (apo proteins).

Table 1 shows the PDB ids (single chains) that were used for projections and analyses of the interfaces of the Trypsin-like family. The first column shows the ids of the complex used as the model in the projection (a total of 35 structures). Enzymes with ids followed by (*) have experimental data in the BRENDA database, which confirms inhibition by the Turkey Ovomucoid inhibitor. This fact reinforces the capability of inhibition prediction of the obtained patterns in the projected interfaces and enables the prediction of novel complexes.

Table 1. Table of PDB ids (Trypsin-like family) with projected interfaces using the Elastase/Turkey Ovomucoid inhibitor complex (1PPF).

PDB	Chain	Molecule	EC. Number
1PPF	E	Human L. Elastase	3.4.21.37
1AUT	C	Activated Protein C	3.4.21.46
1BIO	A	Comp. Factor D	3.4.21.46
1DLE	A	Comp. Factor B	3.4.21.-
1EAX	A	Tumorigenicity 14	3.4.21.-
1EKB	B	Enteropeptidase	3.4.21.9
1ELV	A	Comp C1S	3.4.21.42
1EQ9	A	Chymotrypsin	3.4.21.1 *
1FIW	A	β-Acrosin	3.4.21.-
1GG6	B	γ chymotrypsin	3.4.21.1 *
1GJ7	B	Urokinase	3.4.21.73
1GVK	B	Elastase 1	3.4.21.-
1GVZ	A	Kallikrein	3.4.21.-
1HAO	H	α-Thrombin	3.4.21.5
1HAP	H	α-Thrombin	3.4.21.5
1HJ9	A	β-Trypsin	3.4.21.4 *
1HUT	H	α-Thrombin	3.4.21.5
1LO6	A	Kallikrein 6	3.4.21.-
1M9U	A	Earthworm Fibrinolytic	3.4.21.-
1MD8	A	C1R Comp.	3.4.21.41
1NPM	A	Neuropsin	3.4.21.-
1OP0	A	Venom serine proteinase	3.4.21.-
1ORF	A	Granzyme A precursor	3.4.21.78
1P57	B	Serine protease hepsin	3.4.21.- *
1PQ7	A	Trypsin	3.4.21.4
1RFN	A	Coagulation Factor IX	3.4.21.-
1RTF	B	Plasminogen Activator	3.4.21.68
1T32	A	Cathepsin G	3.4.21.20
1Z8G	A	Serine protease hepsin	3.4.21.-
2BZ6	H	Blood Coag. Factor VIIA	3.4.21.21
2F91	A	hepatopancreas trypsin	3.4.21.4 *
2FPZ	A	Tryptase beta-2	3.4.21.59 *
2HLC	A	Collagenase	3.4.21.-
2PKA	B	Kallikrein A	3.4.21.35
2QY0	B	Comp. C1r subcomponent	3.4.21.41

Table 2 shows PDB ids (single chains) used in the projection and analysis of the interfaces of the Subtilisin-like family. The first column shows the PDB id of the complex used as the model for projection (a total of 9 structures). Enzymes with ids followed by (*) are those that are experimentally proved to be inhibited by Eglic C according to the BRENDA database.

Table 2. Table of PDB ids (Subtilisin-like family) with projected interfaces using the Thermitase/Eglin C inhibitor complex (1TEC).

PDB ids	Chain	Molecule	EC. Number
1TEC	E	Thermitase	3.4.21.66 *
1CGI	A	Sutilisin	3.4.21.62 *
1P8J	A	Furin precursor	3.4.21.75 *
1THM	A	Thermitase	3.4.21.66 *
1V6C	A	alkaline serine protease	3.4.21.-
1WMD	A	protease	3.4.21.-
1ID4	A	Assemblin protease	3.4.21.97
2IXT	A	36KDA protease	3.4.21.-
2PWK	A	Proteinase K	3.4.21.64

2 MATERIALS AND METHODS

Table 3 shows residue atoms that we classify as hydrophobics. The data were obtained through the study of atomic interactions by Sobolev and colleagues and in the methodology used by the Blue Star STING suite:
$http://www.cbi.cnptia.embrapa.br/SMS/index.html$.

Table 3. Table of hydrophobic classifications of the atoms.

Residue	Atom	Residue	Atom	Residue	Atom
ALA	CB	ILE	CD1	SER	CB
ARG	CB	LEU	CB	THR	CB
ARG	CG	LEU	CG	THR	CG2
ARG	CD	LEU	CD1	TRP	CB
ARG	CZ	LEU	CD2	TRP	CG
ASN	CB	LYS	CB	TRP	CD1
ASN	CG	LYS	CG	TRP	CD2
ASP	CB	LYS	CD	TRP	CE2
ASP	CG	LYS	CE	TRP	CE3
CYS	CB	MET	CB	TRP	CZ2
GLN	CB	MET	CG	TRP	CZ3
GLN	CG	MET	SD	TRP	CH2
GLN	CD	MET	CE	TYR	CB
GLU	CB	PHE	CB	TYR	CG
GLU	CG	PHE	CG	TYR	CD1
GLU	CD	PHE	CD1	TYR	CD2
HIS	CB	PHE	CD2	TYR	CE1
HIS	CG	PHE	CE1	TYR	CE2
HIS	CD2	PHE	CE2	TYR	CZ
HIS	CE1	PHE	CZ	VAL	CB
ILE	CB	PRO	CB	VAL	CG1
ILE	CG1	PRO	CG	VAL	CG2
ILE	CG2	PRO	CD		

2.1 Algorithms

2.1.1 Optimization model Linear programming is a technique for optimization of a linear objective function, given a list of requirements represented as linear equations (or in equations). The geometric interpretation of these constraints is a convex polytope that is called the feasible region. These linear program models can be solved by the simplex algorithm that constructs a feasible solution at a vertex of the polytope and then by walking along a path to vertices with non-decreasing values of the objective function until an optimum is reached.

Now we present how these mathematical models can be used to represent the problem of matching the graphs that represent the hydrophobic patches. In this graph $G = (V, E)$, V are the vertices representing the centroids and E is the set of edges of the clique composed by all the centroids. We want to match the edges of the graph such that:

$$min \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} x_{ij}$$

where n is the number of centroids from a first protein and m is the number of centroids from a second protein to be matched. x_{ij} are binary variables that encode possible matches between edges i from the first graph and j from the second graph and d_{ij} is the difference in length of edges i and j.

We optimize this equation subject to the following constraints:

$$\forall i \sum_{j=1}^{m} x_{ij} = 1 (1)$$

$$\forall j \sum_{i=1}^{n} x_{ij} \leq 1 (2)$$

which means that (1) every edge from the first graph must be matched with an edge from the second graph and (2) each edge in the second graph must be matched with an edge from the first graph. This model was built using a perl script and solved using Glpsol.

2.1.2 Agglomerative clustering In this section, we explain how the clustering algorithm works.

Input:
A set of centroids (i.e., a set of three-dimensional coordinates);
The number of clusters.
Output:
A mapping of centroids to clusters.
Each step of the Algorithm 1 is described in the following:

- The first step is to calculate the pairwise distance between the centroids (Line 1).
- Next, the centroid pairs are sorted in ascending order by distance (Line 2).
- The current number of clusters is set to the number of centroids, and each centroid is assigned to a separate cluster (Lines 3-5).
- Subsequently, while the desired number of clusters is not achieved, the clusters are merged according to the closest centroid pairs (Lines 6-7).
- If the selected centroid pair belongs to a different cluster, the two clusters are merged, the mapping is properly modified and the current number of clusters is reduced (Lines 8-10).
- Finally, the centroid-to-cluster mapping is returned (Line 11).

The asymptotic time complexity of the algorithm $O(n^2)$, where n is the number of centroids, and it is due to the pairwise distance calculation between the centroids.

Algorithm 1 Agglomerative Clustering

Input: $CentroidSet, NumberOfClusters$
Output: $MapClusters$
1: Distances ← $calculateDistances$(CentroidSet)
2: SortedPairs ← $sortAscendDist$(CentroidSet, Distances)
3: CurrentNumClusters ← $size$(CentroidSet)
4: **for all** centroid $i \in (CentroidSet)$ **do**
5: MapClusters[i] ← i
6: **while** CurrentNumClusters < NumberOfClusters **do**
7: (A, B) ← $getNextPair$(SortedPairs)
8: **if** MapClusters[A] ≠ MapClusters[B] **then**
9: $mergeClusters$(MapClusters, A, B)
10: CurrentNumClusters ← CurrentNumClusters - 1
11: **return** $MapClusters$

3 FUNCTIONAL ROLE OF THE PATCHES

In this section, we discuss the thermodynamic principles that guided our work. In other words, why we study conserved hydrophobic patches in the interface between enzymes and its inhibitors.

Certainly, Kauzmann (1959) was the first to point out the influence of hydrophobic interactions of non-polar atoms on the large entropic effects verified experimentally in a myriad of phenomena, as for instance: the dissolution of organic molecules, protein folding, stabilization of ligand-protein complexes, protein aggregations etc.

The degree of exposure of apolar atoms or molecules to the solvent is central to the hydrophobicity concept. Lee and Richards (1971) developed a famous method to compute this protein solvent exposition. They created the concept of accessible surface area (known as ASA) as the sum of all the portions of a atom or group of atoms that can be reached by a radius R sphere (representing the solvent) rolled along the protein in close contact with van der Waals surface. Chothia (1974) showed that there is a linear correlation among ASA and the solubility of side chain amino acids in organic solvents (a hydrophobicity measure). He estimates a free energy contribution between 20 and 30 calories mol^{-1} for each $Å^2$ of non-polar atom area not exposed to solvent.

Chothia and Janin (1975) evinced that the apolar / polar ASA ratio tends to be higher for the interface in some protein complexes than for the rest of the surface. This more hydrophobic region at the interface would facilitate, in a thermodynamic perspective, the binding with apolar portions of other chains or molecules. Moreover, a balanced apolar / polar ASA ratio for non-interface

regions would collaborate to prevent misaggregations. Remember that the nefarious polymerization of hemoglobin S in sickle-cell disease is induced by a disequilibrium in this ratio, due to a single mutation of one polar residue by another apolar (GLU by VAL) on the surface of beta chain (Dickerson and Geis (1983)).

In this study, we spot a clear trend of higher apolar / polar ASA ratio toward interface (Figure 1), which is an evidence of the importance of the hydrophobic interactions in protease-inhibitor complex formation.

Murphy and Freire (1992) demonstrated that one can reliably estimate some thermodynamic parameters from apolar and polar ASA calculations. Baker and Murphy (1997) have applied this method to evaluate some of these parameters involving the binding energetics of Turkey Ovomucoid third domain inhibitor (OMTKY3) to Porcine Pancreatic Elastase (PPE). They also have performed a comparison between these empirical calculations with experimental data measured by an ITC (isothermical titration calorimetry). Table 4 below shows that the correspondence between experimental and calculated data are in agreement, except perhaps by the enthalpy change. Nevertheless, this parameter may be considered negligible in the final free energy change composition (responding only for 4% of it). Hence, we can conclude that the binding of OMTKY3 to PPE is essencially entropic driven. That means that the complex formation is fundamentally guided by hydrophobic interactions.

Table 4. Comparison of thermodynamic parameters from experimental and empirical estimation for OMTKY3 / PPE complex as made in Baker and Murphy (1997). Experimental data reported at 25^oC.

Parameter	Experimental	Calculated
$\Delta C_p^o(kJK^{-1}mol^{-1})$	-1.1 ± 0.1	-1.4
$\Delta H^o(kJmol^{-1})$	-2.5 ± 1.0	2.3
$\Delta S^o(JK^{-1}mol^{-1})$	195 ± 4	190
$\Delta G^o(kJmol^{-1})$	-60.6 ± 0.5	-54

If we assume that these premises remain valid for the most of the other protease-inhibitor complexes (a reasonable assumption), we can apply the methodology of Murphy and co-authors to estimate the solvation entropy change of the studied complexes through computed ASAs. We have made this choice because the entropy change is the parameter most revealing of hydrophobic interactions and it is the most preponderant factor involving in protein-protein binding (Baker and Murphy (1997)). We can make this in two steps:

(1) Determine the heat capacity change as:

$$\Delta Cp = a\Delta ASA_{apolar} + b\Delta ASA_{polar} \quad (1)$$

(2) Determine the solvation entropy change as:

$$\Delta S = \Delta C_p.ln(T/Ts) \quad (2)$$

where, a and b are adjusted parameters estimated in Murphy and Freire (1992) as $1.88JK^{-1}mol^{-1}A^{-1}$ e $-1.09JK^{-1}mol^{-1}A^{-1}$, respectivelly; T is the room or experimental temperature and Ts is a reference temperature where entropy change is taken to be zero (about $385K$) Baldwin (1986).

In Figure 2 we can see that there is a strong linear relationship (Pearson correlation coefficient of 0.98) between the solvation entropy change (inferred as described above) and the extension of our patches, measured in number of hydrophobic atoms inside them. The intercept seems away from zero because the heat capacity change, in Murphy and co-authors methodology, has a polar term.

In conclusion, we believe that the conserved HydroPaCes have a clear functional role in the inhibition. They represent the hydrophobic network that paves (and perhaps help to codify) the active sites of proteases, going beyond the traditional catalytic triads / diads and oxyanion holes conservation. Their interfaces have a hydrophobic bias which is clearly distinct from the rest of the surface, as showed in Figure 1. The energetic dissection of Elastase PPE made by Baker and Murphy (1997) indicated that the interaction between proteases and inhibitors may be fundamentally orchestrated by hydrophobic / entropic forces, and that they may represent as high as 96% of free energy change in complex formation. Corroborating these assumptions, our patches also have an evident correlation with the solvation entropy change, as showed in Figure 2.

We are aware that our model and algorithms are, as any heuristic, an approximation to the solution. A hydrophobic pattern may be a necessary but not sufficient condition to protease-inhibitor recognition. However, it is important to note that there is no consensus in the literature about the exact functional role of hydrophobic interactions in proteins. While Chothia and Janin (1975) says that as hydrophobic contribution is entirely unspecific, it would lead to all kinds of incorrect interactions in a cell, Dill (1999) defends that hydrophobic interactions are not nonspecific glue, but a crucial structure-determining driving force. Independently of who has the reason, our work has the merit to put empirically in evidence that robust patterns of conservations are emerging when we focus on hydrophobic network of different protease interfaces, even considering those so structurally distinct as Subtilisin-like and Trypsin-like. We are confident that this patterns could help to explain the experimentally verified cross-inhibition phenomenon. Indeed, we can see in Figures 3 and 4 some correspondences among the apolar atoms of the inhibitors and enzymes interfaces.

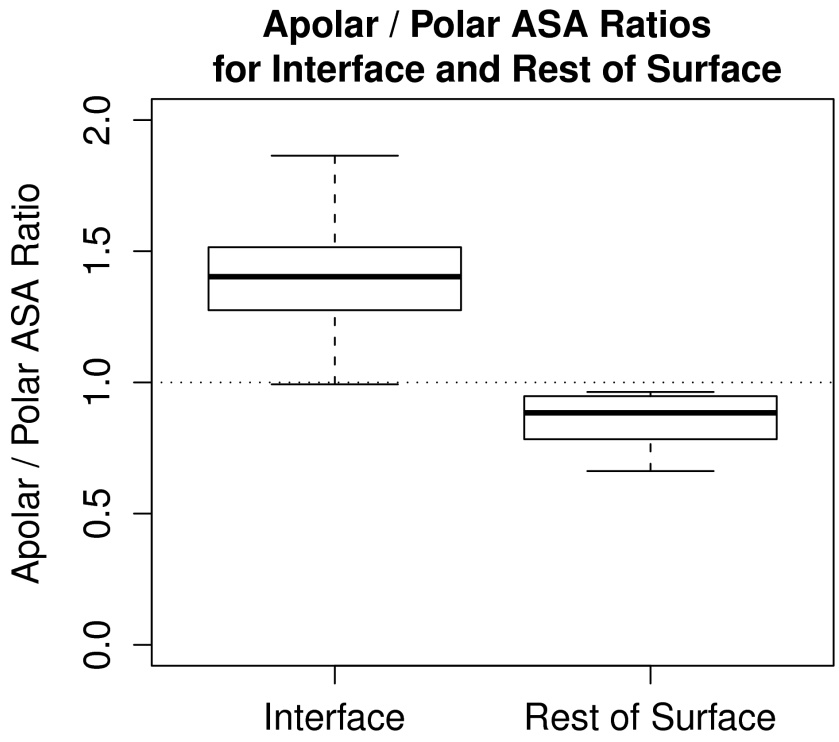


Fig. 1. Apolar / polar ASA ratios of interface and the rest of protein surfaces for the following proteases (PDB codes): 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB

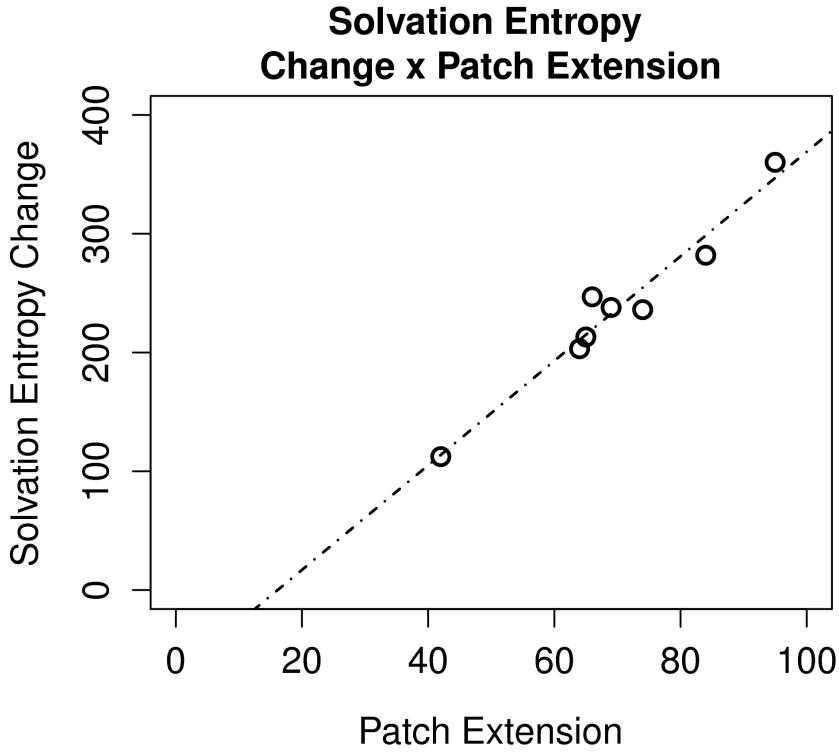


Fig. 2. Correlation between solvation entropy change and patch extension for the following proteases (PDB codes): 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB

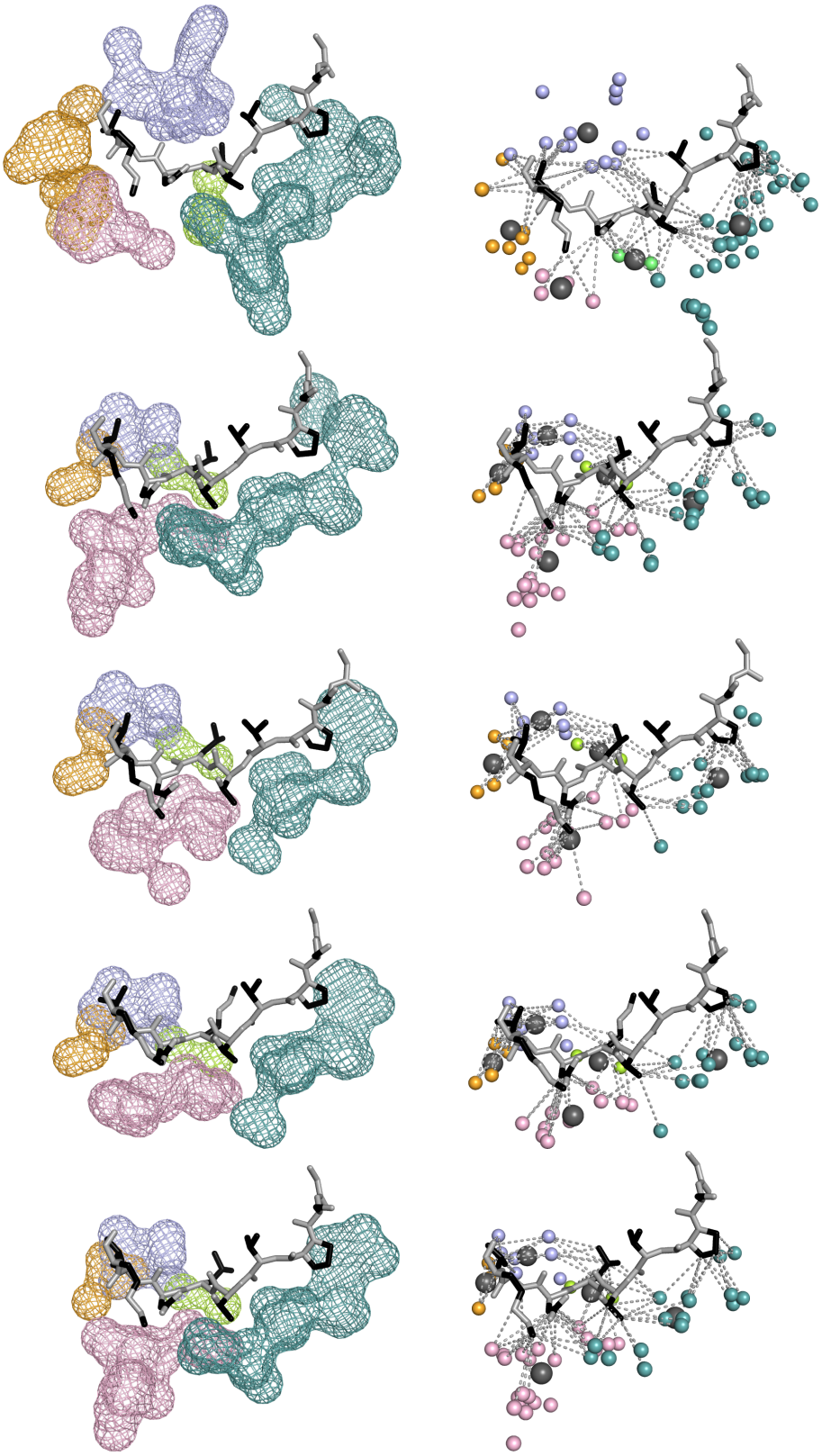


Fig. 3. Complementarity between the patches in enzymes and inhibitor Eglin C interface. We present polar atoms in gray and apolar atoms in black. On proteases interfaces, the apolar regions are presented with meshes (left column) and spheres (right column). Atoms between 4 and 6 angstrons from any apolar inhibitor atom are connected through dashed edges. Big spheres in medium gray represent the centroids.

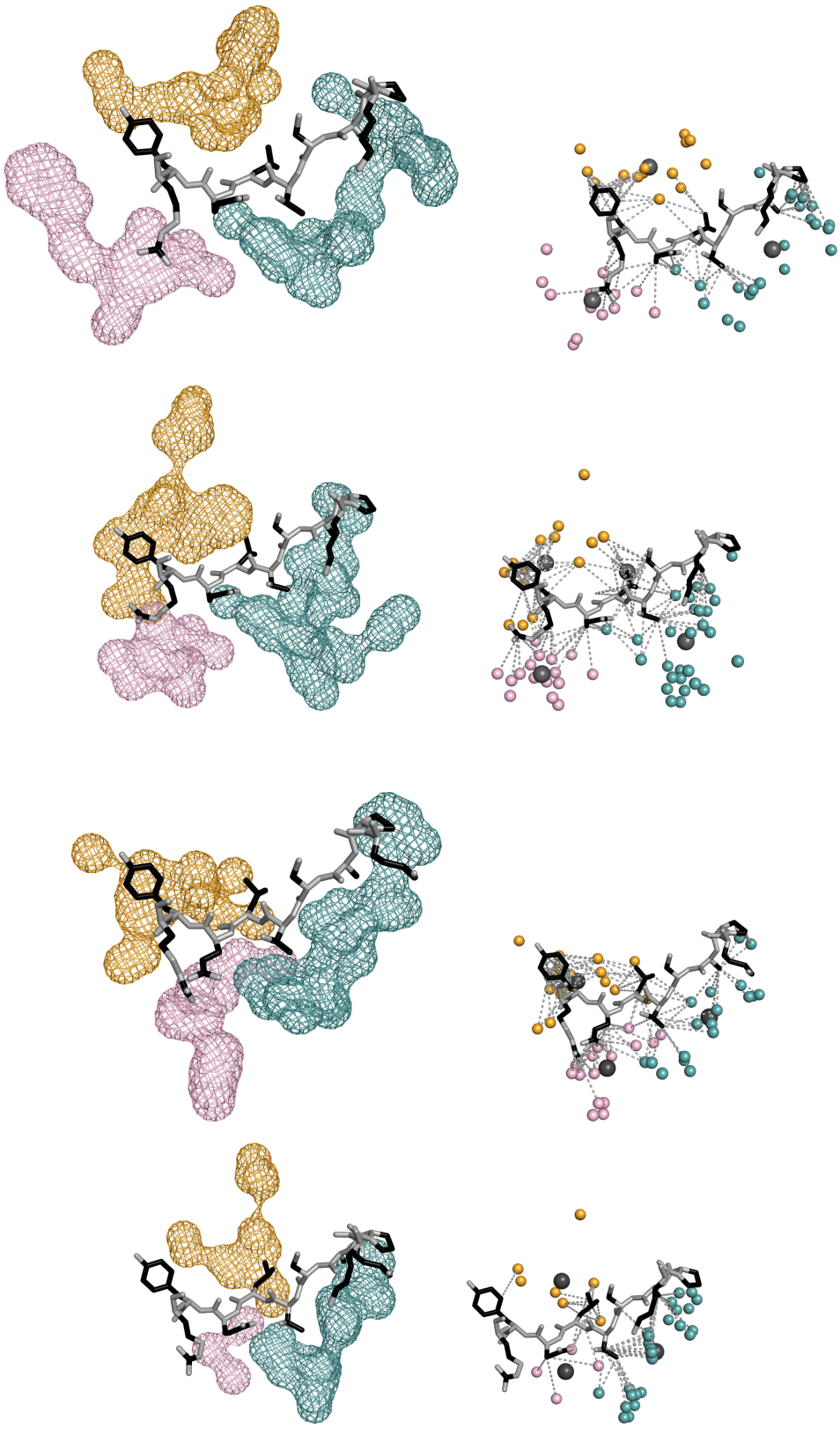


Fig. 4. Complementarity between the patches in enzymes and inhibitor Ovomucoid interface. We present polar atoms in gray and apolar atoms in black. On proteases interfaces, the apolar regions are presented with meshes (left column) and spheres (right column). Atoms between 4 and 6 angstrons from any apolar inhibitor atom are connected through dashed edges. Big spheres in medium gray represent the centroids.

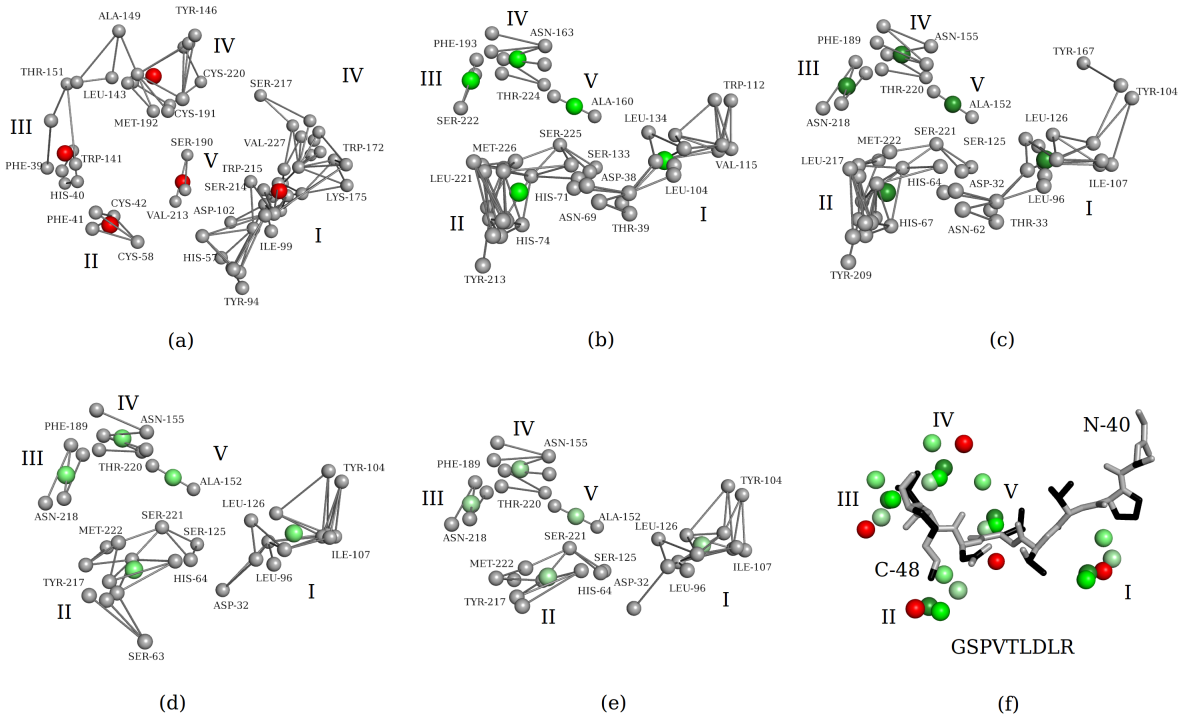


Fig. 5. Connected components of cross-inhibition: hydrophobic patches for 5 Serine Proteases inhibited by Eglin C. In (a), we can see a sample with Trypsin-like, and in (b), (c), (d) and (e) with Subtilisin-like enzymes. We show an atomic graph in which residue types and numbers are presented. Green shade balls show the centroids that represent each of the patches in the Subtilisin-like enzymes and the red ones are the Trypsin-like. In the last part of the figure (f), we present the inhibitor (residues from 40 to 48) as gray sticks (highlighting the hydrophobic atoms in black), and the five centroids are superposed in colors.

Table 5. Eglin C cross-inhibition analysis: Evaluation of the proposed approaches in Eglin C cross-inhibition.

METHOD COMPARISON (CCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	4.735	14.563	5	1	I	2.923	10.092	5	-
II	3.125	11.669	5	1	II	6.154	8.348	5	-
III	2.941	17.408	5	1	III	2.785	9.189	5	-
IV	2.844	12.220	5	1	IV	5.793	8.729	5	-
V	3.528	13.314	6	0.9	V	9.647	10.109	5	-
Means	3.435	13.835	26	0.98	Means	5.460	9.294	25	-

METHOD COMPARISON (SGCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	2.840	11.485	5	1	I	7.363	16.348	5	-
II	2.361	17.882	5	1	II	5.227	10.236	5	-
III	2.371	13.504	8	0.70	III	2.823	9.964	5	-
IV	2.228	9.680	4	0.60	IV	4.960	8.375	5	-
Means	2.450	13.138	22	0.82	Means	5.093	11.231	20	-

METHOD COMPARISON (EBCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	2.448	16.456	5	1	I	9.444	10.736	5	-
II	3.076	11.255	5	1	II	2.176	11.300	5	-
III	2.421	11.629	5	1	III	5.793	9.007	5	-
IV	3.290	13.319	7	0.90	IV	3.076	9.852	5	-
V	2.160	10.692	4	0.6	V	6.205	9.035	5	-
Means	2.679	12.670	26	0.90	Means	5.339	9.986	25	-

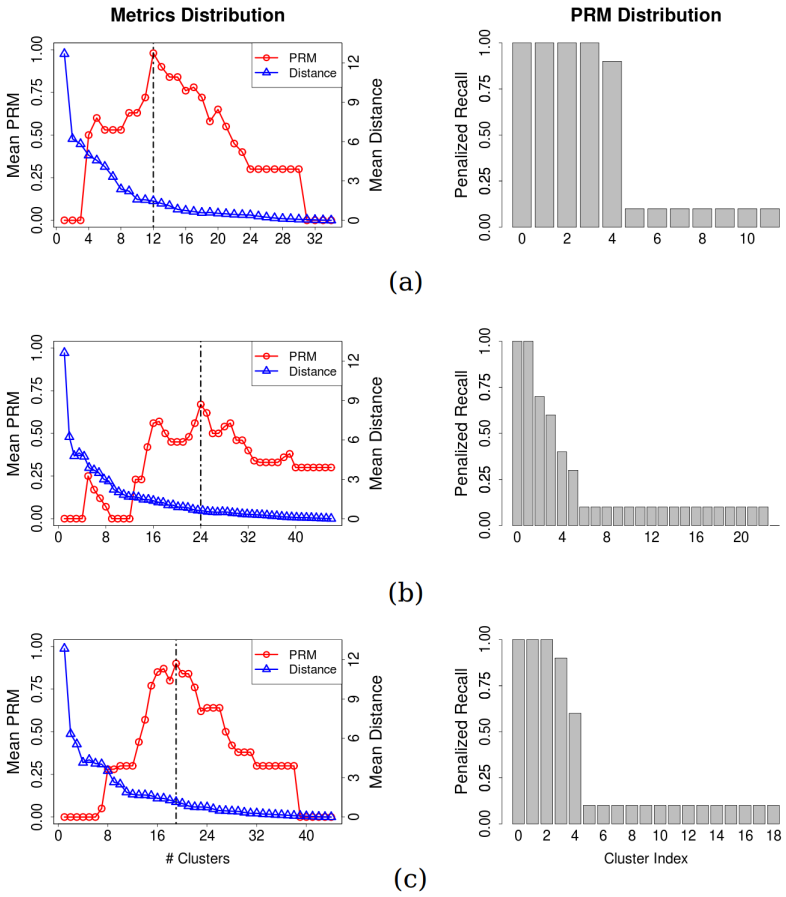


Fig. 6. The left-hand graphs show the distributions of PRM means and intra-cluster distance means obtained. Notice that we use the number of clusters that maximizes the PRM and this always implies in low intra-cluster distances means. The right-hand graphs show the PRM distribution for the best clustering configuration or in other words the highest number of conserved clusters. Here we can see the results for the five complexes with the inhibitor Eglin C. In (a), the use of the CCC (with 12 clusters) shows the 5 most conserved regions at the interfaces. In (b), the use of the SGCC (with 24 clusters) shows the 4 most conserved regions at the interfaces. In (c), the use of the EBCC (with 19 clusters) shows the 5 most conserved regions at the interfaces.

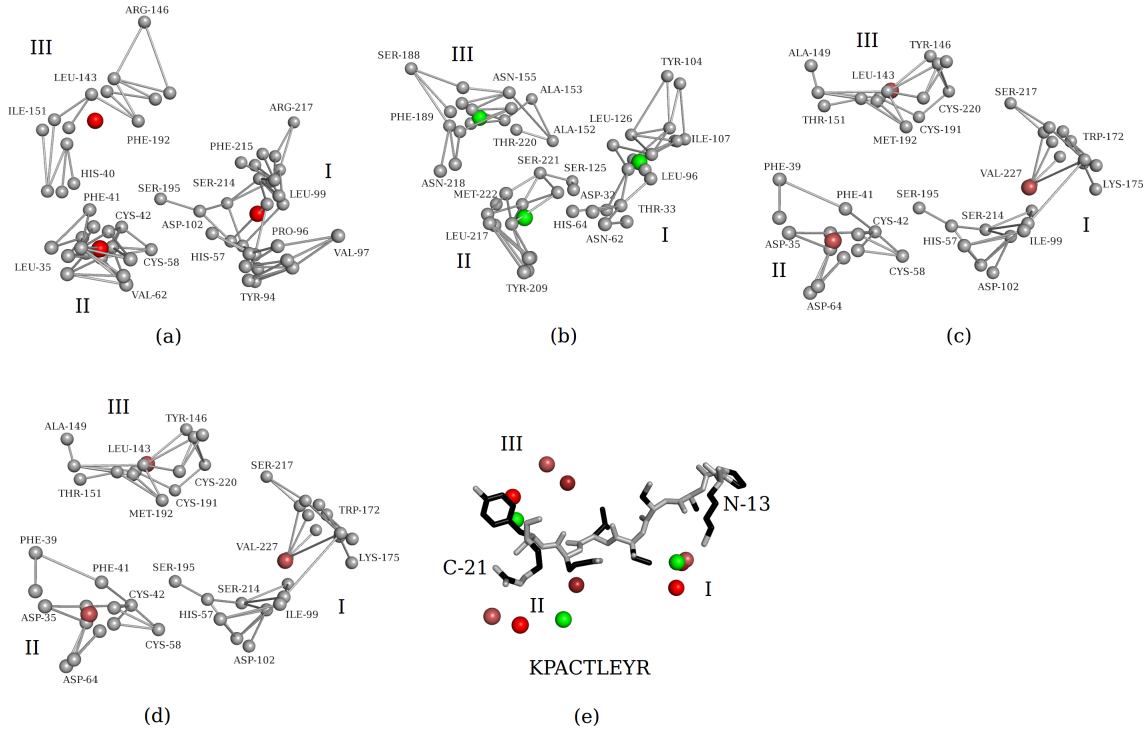


Fig. 7. Connected Components of Cross-Inhibition: Hydrophobic patches for 4 Serine Proteases inhibited by Turkey Ovomucoid. In (a),(c) and (d) we can see a sample with Trypsin-like, and in (b) with Subtilisin-like, enzymes. We show an atomic graph in which residue types and numbers are presented. Red shades balls show the centroids that represent each of the patches in the Trypisin-like enzymes and the green ones are the Subtilisin-like. In the last part of the figure (e), we present the inhibitor (residues from 13 to 21) as gray sticks (highlighting the hydrophobic atoms in black), and the three centroids are superposed in colors.

Table 6. Turkey Ovomucoid cross-inhibition analyses: an exhaustive comparison of the proposed approaches in Turkey Ovomucoid cross-inhibition.

METHOD COMPARISON (CCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	6.392	12.757	4	1	I	10.884	9.387	4	-
II	2.136	14.042	4	1	II	5.892	11.531	4	-
III	5.880	12.917	5	0.83	III	7.251	10.288	4	-
Means	4.803	13.239	13	0.94	Means	8.009	10.402	12	-

METHOD COMPARISON (SGCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	2.871	10.999	4	1	I	6.869	11.014	4	-
II	2.932	10.999	3	0.5	II	11.737	11.014	4	-
Means	2.901	10.999	7	0.75	Means	9.303	11.014	8	-

METHOD COMPARISON (EBCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	3.605	14.045	4	1	II	3.965	11.997	4	-
II	3.233	14.045	3	0.50	III	8.953	11.997	4	-
Means	3.419	14.045	7	0.75	Means	6.459	11.997	8	-

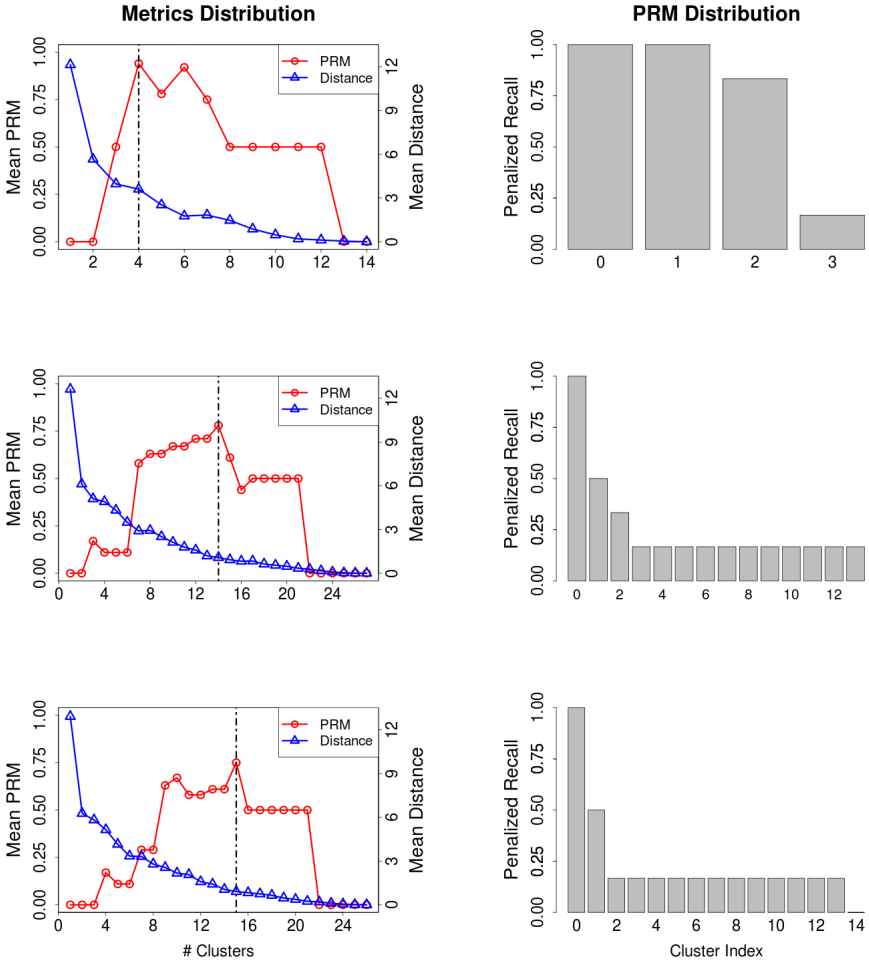


Fig. 8. The left-hand graphs show the distributions of PRM means and intra-cluster distance means obtained. Notice that we use the number of clusters that maximizes the PRM and this always implies in low intra-cluster distances means. The right-hand graphs show the PRM distribution for the best clustering configuration or in other words the highest number of conserved clusters. Here we can see the results for the fiour complexes with the inhibitor Turkey Ovomucoid. In (a), the use of the CCC (with 4 clusters) shows the 3 most conserved regions at the interfaces. In (b), the use of the SGCC (with 14 clusters) shows the 2 most conserved regions at the interfaces. In (c), the use of the EBCC (with 15 clusters) shows the 2 most conserved regions at the interfaces.

Table 7. Trypsin-like projection analysis: Exhaustive comparison of the proposed approaches for Trypsin-like projections.

METHOD COMPARISON (CCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	1.312	11.729	35	1	I	5.122	10.547	35	-
II	3.685	10.499	46	0.75	II	2.139	9.498	35	-
III	4.587	13.422	51	0.58	III	6.848	9.955	35	-
IV	5.787	15.267	60	0.52	IV	4.823	11.202	35	-
Means	3.843	12.729	192	0.71	Means	4.733	10.300	140	-

METHOD COMPARISON (SGCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	1.412	11.146	47	0.81	I	5.161	7.882	35	-
II	1.755	11.915	33	0.72	II	6.649	6.769	35	-
III	2.843	13.382	48	0.71	III	9.542	5.109	35	-
IV	2.669	12.003	45	0.70	IV	9.044	6.725	35	-
V	2.169	11.356	29	0.54	V	9.557	5.759	35	-
Means	2.169	11.960	202	0.69	Means	7.990	6.448	175	-

METHOD COMPARISON (EBCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	0.942	10.623	35	1	I	7.958	6.724	35	-
II	2.384	12.326	28	0.54	II	9.078	4.626	35	-
III	2.292	12.340	60	0.52	III	7.429	6.895	35	-
Means	1.872	11.763	123	0.68	Means	8.155	6.082	105	-

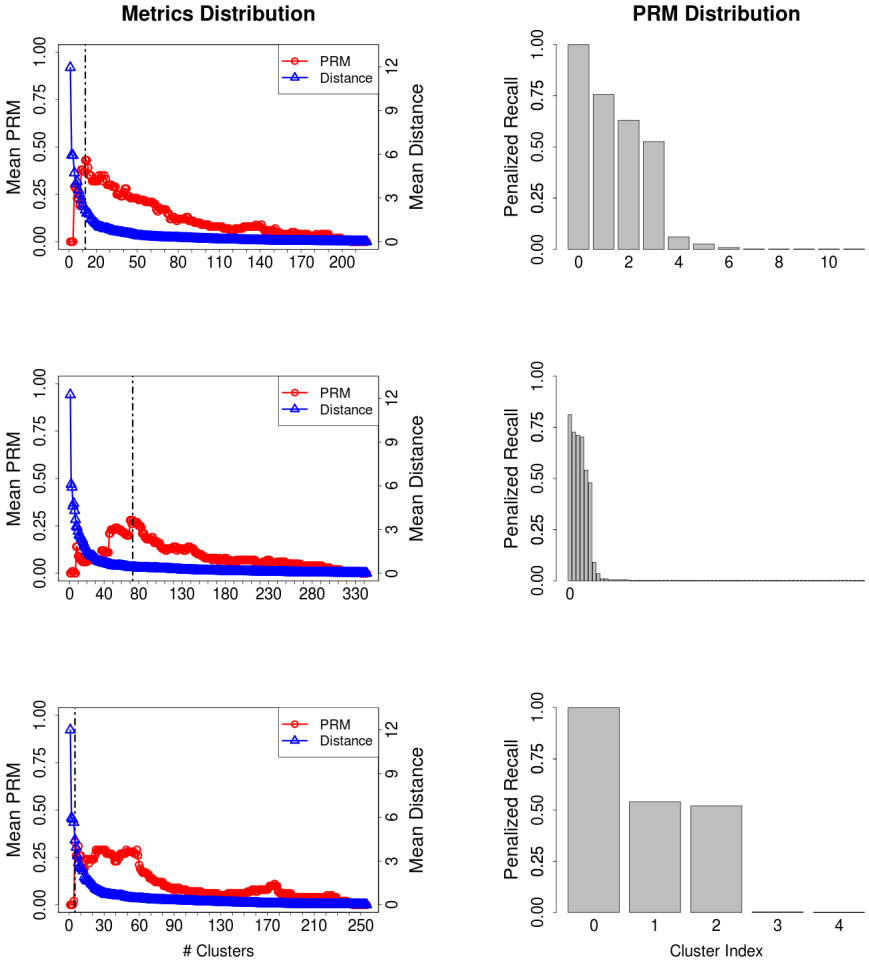


Fig. 9. The left-hand graphs show the distributions of PRM means and intra-cluster distance means obtained. Notice that we use the number of clusters that maximizes the PRM and this always implies in low intra-cluster distances means. The right-hand graphs show the PRM distribution for the best clustering configuration or in other words the highest number of conserved clusters. Here we can see the results for the interfaces of the 35 Trypsin-like enzymes projected. In (a), the use of the CCC (with 12 clusters) shows the 4 most conserved regions at the interfaces. In (b), the use of the SGCC (with 73 clusters) shows the 5 most conserved regions at the interfaces. In (c), the use of the EBCC (with 5 clusters) shows the 3 most conserved regions at the interfaces.

Table 8. Subtilisin-like projection analysis: Exhaustive comparison of the proposed approaches for Subtilisin-like projections.

METHOD COMPARISON (CCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	1.106	10.614	9	1	I	3.893	9.983	9	-
II	0.755	9.231	9	1	II	5.570	9.261	9	-
III	0.479	9.251	9	1	III	5.619	8.393	9	-
IV	1.080	13.530	7	0.58	IV	3.126	11.720	9	-
Means	0.885	10.656	34	0.89	Means	4.552	9.839	36	-

METHOD COMPARISON (SGCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	2.103	12.330	9	1	I	4.386	10.968	9	-
II	1.243	10.002	8	0.78	II	8.176	9.881	9	-
III	0.479	11.267	8	0.78	III	5.707	9.816	9	-
IV	1.022	13.927	8	0.78	IV	6.172	8.899	9	-
V	1.144	16.910	7	0.58	V	3.502	9.055	9	-
Means	1.198	12.887	40	0.78	Means	5.588	9.723	45	-

METHOD COMPARISON (EBCC)									
Agglomerative Cluster (AC)					Optimization Model (OM)				
Clusters	Distances (Å)		Size	PRM	Clusters	Distances (Å)		Size	PRM
	Intra-cluster	Inter-cluster				Intra-cluster	Inter-cluster		
I	1.611	12.166	9	1	I	5.608	8.394	9	-
II	0.479	10.160	9	1	II	5.220	9.750	9	-
III	2.728	12.660	11	0.97	III	3.257	10.321	9	-
IV	1.132	11.242	10	0.97	IV	6.772	7.611	9	-
V	1.464	13.213	8	0.78	V	4.456	10.758	9	-
VI	1.419	9.291	7	0.58	VI	6.232	9.671	9	-
Means	1.472	11.455	54	0.88	Means	5.257	9.417	54	-

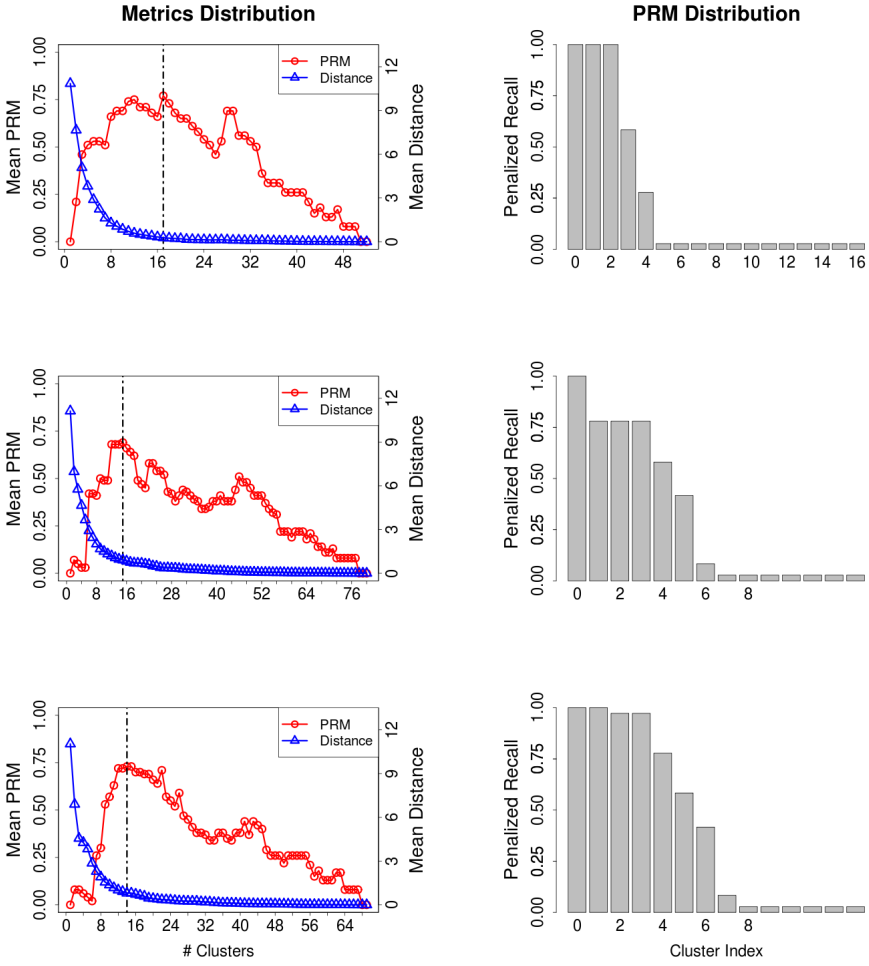


Fig. 10. The left-hand graphs show the distributions of PRM means and intra-cluster distance means obtained. Notice that we use the number of clusters that maximizes the PRM and this always implies in low intra-cluster distances means. The right-hand graphs show the PRM distribution for the best clustering configuration or in other words the highest number of conserved clusters. Here we can see the results for the interfaces of the 9 Subtilisin-like enzymes projected. In (a), the use of the CCC (with 17 clusters) shows the 4 most conserved regions at the interfaces. In (b), the use of the SGCC (with 15 clusters) shows the 5 most conserved regions at the interfaces. In (c), the use of the EBCC (with 14 clusters) shows the 6 most conserved regions at the interfaces.

Table 9. Summary comparison of the different approaches and experiments: In this figure, we compare all of the proposed approaches for Eglin C and Turkey Ovomucoid cross-inhibition and for the respective projections.

			Mean Intra-cluster (Å)	Mean Inter-cluster (Å)	PRM
Eglin C Cross Inhibition	CCC	AC	3.435	13.835	**0.98**
		OM	5.460	9.294	-
	SGCC	AC	2.450	13.138	0.82
		OM	5.093	11.231	-
	EBCC	AC	2.679	12.670	0.90
		OM	5.339	9.986	-
			Mean Intra-cluster (Å)	Mean Inter-cluster (Å)	PRM
Turkey Ovomucoid Cross Inhibition	CCC	AC	4.803	13.239	**0.94**
		OM	8.009	10.402	-
	SGCC	AC	2.901	10.999	0.75
		OM	9.303	11.014	-
	EBCC	AC	3.419	14.045	0.75
		OM	6.459	11.997	-
			Mean Intra-cluster (Å)	Mean Inter-cluster (Å)	PRM
Trypsin-like Projection	CCC	AC	3.843	12.729	**0.71**
		OM	4.733	10.300	-
	SGCC	AC	2.169	11.960	0.69
		OM	7.990	6.448	-
	EBCC	AC	1.872	11.763	0.68
		OM	8.155	6.082	-
			Mean Intra-cluster (Å)	Mean Inter-cluster (Å)	PRM
Subtilisin-like Projection	CCC	AC	0.885	10.656	**0.89**
		OM	4.552	9.839	-
	SGCC	AC	1.198	12.887	0.78
		OM	5.588	9.723	-
	EBCC	AC	1.472	11.455	0.88
		OM	5.257	9.417	-

REFERENCES

Baker, B. M. and Murphy, K. P. (1997). Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase. *J. Mol. Biol.*, **268**, 557–569.

Baldwin, R. (1986). Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl. Acad. Sci.*, **83**, 8069–8072.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, **248**(5446), 338–339.

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, **256**, 705–708.

Dickerson, R. E. and Geis, I. (1983). *Hemoglobin*. Menlo Park:The Benjamin/Cumming Publishing Co. Inc.

Dill, K. (1999). Polymer principles and protein folding. *Protein Science*, **8**, 1166–1180.

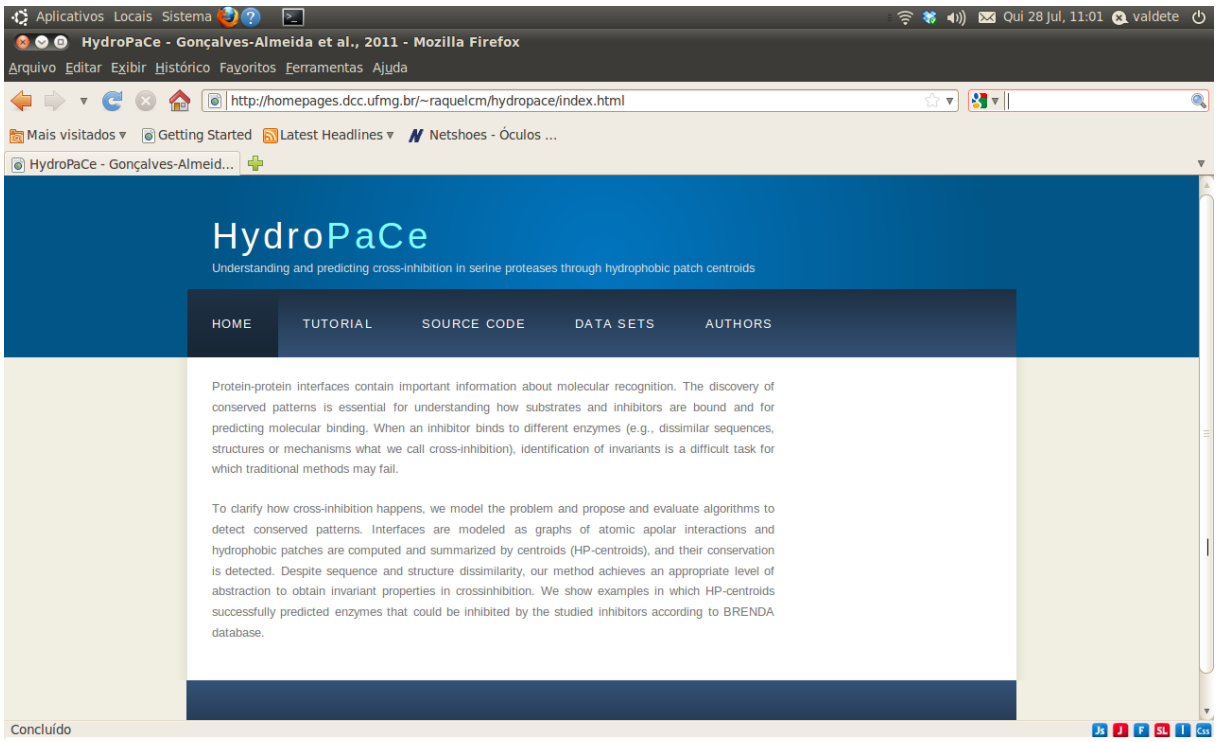
Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Prot. Chem*, **14**, 1–63.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Murphy, K. P. and Freire, E. (1992). Thermodynamics of structural stability and cooperative folding behavior in proteins. *Adv Prot Chem.*, **43**, 313–361.

Anexo C

Site (www.dcc.ufmg.br/∼raquelcm/hydropace)



Anexo D

Nossa Equipe

Valdete Maria Gonçalves de Almeida
Programa de Doutorado em Bioinformática
Universidade Federal de Minas Gerais/UFMG
Doutoranda em Bioinformática

Prof. Marcelo Matos Santoro
Departamento de Bioquímica e Imunologia
Universidade Federal de Minas Gerais/UFMG
Orientação do doutorado

Profa. Raquel Cardoso de Melo Minardi
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais/UFMG
Co-orientadora do doutorado

Prof. Carlos Henrique da Silveira
Departamento de Engenharia da Computação
Universidade Federal de Itajubá/UNIFEI - campus avançado de Itabira-MG
Co-orientador do doutorado

Douglas Eduardo Valente Pires
Doutorando em Bioinformática
Universidade Federal de Minas Gerais
Colaborador

Prof. Wagner Meira Jr.
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais/UFMG
Colaborador

Referências Bibliográficas

Angelov, B.; Sadoc, J.-F.; Jullien, R.; Soyer, A.; Mornon, J.-P. e Chomilier, J. (2002). Nonatomic solvent-driven voronoi tessellation of proteins: an open tool to analyze protein folds. *Proteins*, 49(4):446–56.

Bachman, P. e Liu, Y. (2009). Structure discovery in ppi networks using pattern-based network decomposition. *Bioinformatics*, 25(14):18–1821.

Bahadur, R. e Zacharias, M. (2008). The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cellular and Molecular Life Sciences*, 65(7):1059–1072.

Baker, B. e Murphy, K. (1997). Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase1,* 1. *Journal of molecular biology*, 268(2):557–569.

Berg, J. M.; Tymoczko, J. L. e Stryer, L. (2005). *Biochimestry.* W. H. Freeman and Company, 5 edição.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. e Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28(1):235–42.

Betzel, C.; Dauter, Z.; Genov, N.; Lamzin, V.; Navaza, J.; Schnebli, H. P.; Visanji, M. e Wilson, K. S. (1993). Structure of the proteinase inhibitor eglin c with hydrolysed reactive centre at 2.0 a resolution. *FEBS Lett*, 317(3):185–8.

Beynon, R. e Bond, J. S. (2001). *Proteolytic enzymes: nomenclature and classification.* Oxford University Press.

Bode, W.; Wei, A. Z.; Huber, R.; Meyer, E.; Travis, J. e Neumann, S. (1986). X-ray crystal structure of the complex of human leukocyte elastase (pmn elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO J*, 5(10):2453–8.

Caffrey, D.; Somaroo, S.; Hughes, J.; Mintseris, J. e Huang, E. (2004). Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13(1):190–202.

Chakrabarti, P. e Janin, J. (2002a). Dissecting protein-protein recognition sites. *Proteins Structure Function and Genetics*, 47(3):334–343.

Chakrabarti, P. e Janin, J. (2002b). Dissecting protein-protein recognition sites. *Proteins Structure Function and Genetics*, 47(3):334–343.

Chothia, C. e Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 256(5520):705–708.

Colizza, V.; Serrano, M. A. e Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nat. Phys.*, (2):110–115.

Conte, L.; Chothia, C. e Janin, J. (1999). The atomic structure of protein-protein recognition sites1. *Journal of molecular biology*, 285(5):2177–2198.

Csermely, P. (2008). Creative elements: network-based predictions of active centres in proteins and cellular and social networks. *Trends Biochem Sci*, 33(12):569–76.

da Silveira, C. H.; Pires, D. E. V.; Minardi, R. C.; Ribeiro, C.; Veloso, C. J. M.; Lopes, J. C. D.; Meira, W.; Neshich, G.; Ramos, C. H. I.; Habesch, R. e Santoro, M. M. (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins*, 74(3):727–43.

Dupuis, F.; Sadoc, J.-F.; Jullien, R.; Angelov, B. e Mornon, J.-P. (2005). Voro3d: 3d voronoi tessellations applied to protein structures. *Bioinformatics*, 21(8):1715–6.

Ekici, O. D.; Paetzel, M. e Dalbey, R. E. (2008). Unconventional serine proteases: variations on the catalytic ser/his/asp triad configuration. *Protein Sci*, 17(12):2023–37.

Erik, L.; Berk, H. e van der Spoel, D. (2001). Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model*, 7:306–317.

Finn, R.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J.; Gavin, O.; Gunesekaran, P.; Ceric, G.; Forslund, K.; Holm, L.; Sonnhammer, E.; Eddy, S. e Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Research*, (38):211–222.

Fujinaga, M.; Sielecki, A. R.; Read, R. J.; Ardelt, W.; Laskowski, M. e James, M. N. (1987). Crystal and molecular structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 a resolution. *J Mol Biol*, 195(2):397–418.

Gent, D. V.; Sharp, P.; Morgan, K. e Kalsheker, N. (2003). Serpins: Structure, function and molecular evolution. *The International Journal of Biochemistry and Cell. Biology*, (35):1536–1547.

Huntington, J. A. e Carrell, R. W. (2001). The serpins: nature's molecular mousetraps. *Sci Prog*, 84(2):125–36.

Hyberts, S. G.; Goldberg, M. S.; Havel, T. F. e Wagner, G. (1992). The solution structure of eglin c based on measurements of many noes and coupling constants and its comparison with x-ray structures. *Protein Sci*, 1(6):736–51.

IUBMB (1999). Enzyme nomenclature 1999. recommendations of the nomemclature committee of the international union of biochemistry and moleculare biology on the nomenclature and classification of enzimes. *European Journal Biochemistry*.

Janin, J.; Chothia, C.; Shabb, J.; Ng, L.; Corbin, J.; Butikofer, P.; Lin, Z.; Chiu, D.; Lubin, B.; Kuypers, F. et al. (1990). The structure of protein-protein recognition sites. *structure*, 265(27).

Krowarsch, D.; Cierpicki, T.; Jelen, F. e Otlewski, J. (2003). Canonical protein inhibitors of serine proteases. *Cell Mol Life Sci*, 60(11):2427–44.

Laskowski, M. e Qasim, M. A. (2000a). What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim Biophys Acta*, 1477(1-2):324–37.

Laskowski, M. J. e Qasim, M. A. (2000b). What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim Biophys Acta*, (1477):324–337.

Lesk, A. M. e Fordham, W. D. (1996). Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *J Mol Biol*, 258(3):501–37.

López-Otín, C. e Bond, J. (2008). Proteases: multifunctional enzymes in life and disease. *Journal of Biological Chemistry*, 283(45):30433.

Mangan, M.; Kaiserman, D. e Bird, P. I. (2008). The role of serpins in vertebrate immunity. *Tissue Antigens*, (72):1–10.

Melo, R.; Ribeiro, C.; Murray, C.; Veloso, C.; Silveira, C.; Neshich, G.; Meira Junior, W.; Carceroni, R. e Santoro, M. (2007). Finding protein-protein interaction patterns by contact map matching. *Genetics and Molecular Research*, 6:946–963.

Murzin, A. G.; Brenner, S. E.; Hubbard, T. e Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40.

Neshich, G.; Togawa, R.; Mancini, A.; Kuser, P.; Yamagishi, M.; Pappas, G.; Torres, W. et al. (2003). Sting millennium: a web-based suite of programs for comprehensive

and simultaneous analysis of protein structure and sequence. *Nucleic acids research*, 31(13):3386.

Newman, M. e Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Page, M. J. e Di Cera, E. (2008). Serine peptidases: classification, structure and function. *Cell Mol Life Sci*, 65(7-8):1220–36.

Papamokos, E.; Weber, E.; Bode, W.; Huber, R.; Empie, M. W.; Kato, I. e Laskowski, M. (1982). Crystallographic refinement of japanese quail ovomucoid, a kazal-type inhibitor, and model building studies of complexes with serine proteases. *J Mol Biol*, 158(3):515–37.

Pires, D.; Silveira, C.; Santoro, M. e Meira, W. J. (2007). Pdbest-pdb enhanced structures toolkit. *In Proceedings of the 3rd International Conference of Brazil Association for Bioinformatics. São Paulo: AB3C Publishing*, p. 39.

Pontius, J.; Richelle, J. e Wodak, S. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of molecular biology*, 264(1):121–136.

Poupon, A. (2004). Voronoi and voronoi-related tessellations in studies of protein structure and interaction. *Current Opinion in Structural Biology*, 14(2):233–241.

Qasim, M. A.; Ganz, P. J.; Saunders, C. W.; Bateman, K. S.; James, M. N. e Laskowski, M. (1997). Interscaffolding additivity. association of p1 variants of eglin c and of turkey ovomucoid third domain with serine proteinases. *Biochemistry*, 36(7):1598–607.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, M.; Tanksale, A. P.; Ghatge, M. e Deshpande, V. (1998). Molecular and biotechnological aspects of microbial protease, microbiology and molecular biology reviews. *Microbiol Mol. Biol. Rev.*, (62):597–635.

Rawlings, N.; Barrett, A.; Barrett, A.; Rawlings, N. e Woessner, J. (2004). Handbook of proteolytic enzymes. *Handbook of Proteolytic Enzymes*, 1.

Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J. e Barrett, A. J. (2008). Merops: the peptidase database. *Nucleic Acids Res*, 36(Database issue):D320–5.

Reichardt, J. e Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1):016110.

Ribeiro, C.; Togawa, R. C.; Neshich, I. A. P.; Mazoni, I.; Mancini, A. L.; de Melo Minardi, R. C.; da Silveira, C. H.; Jardine, J. G.; Santoro, M. M. e Neshich, G. (2010). Analysis of binding properties and specificity through identification of the interface forming residues (ifr) for serine proteases in silico docked to different inhibitors. *BMC Struct Biol*, 10:36.

Richards, F. (1974). The interpretation of protein structures: Total volume, group volume distributions and packing density* 1. *Journal of Molecular Biology*, 82(1):1–14.

Robertson, A. D.; Westler, W. M. e Markley, J. L. (1988). Two-dimensional nmr studies of kazal proteinase inhibitors. 1. sequence-specific assignments and secondary structure of turkey ovomucoid third domain. *Biochemistry*, 27(7):2519–29.

Roland, J. S.; Jack, A. e Rao, M. (1997). Subtlases: The superfamily of subtilisin-like serine proteases. *Protein Science*, (6):501–523.

Sales-Pardo, M. e Amaral, L. A. N. (2007). Classes of complex networks defined by role-to-role connectivity profiles. *Nat. Phys.*, (3):63–69.

Scheer, M.; Grote, A.; Chang, A.; Schomburg, I.; Munaretto, C.; Rother, M.; Söhngen, C.; Stelzer, M.; Thiele, J. e Schomburg, D. (2011). BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, 39:670–676.

Shatsky, M.; Nussinov, R. e Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–56.

Shulman-Peleg, A.; Shatsky, M.; Nussinov, R. e Wolfson, H. (2007). Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC biology*, 5(1):43.

Siezen, R. J. e Leunissen, J. A. (1997). Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Sci*, 6(3):501–23.

Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E. e Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–32.

Soundararajan, V.; Raman, R.; Raguram, S.; Sasisekharan, V. e Sasisekharan, R. (2010a). Atomic interaction networks in the core of protein domains and their native folds. *PLoS One*, 5(2):e9391.

Soundararajan, V.; Raman, R.; Raguram, S.; Sasisekharan, V. e Sasisekharan, R. (2010b). Atomic interaction networks in the core of protein domains and their native folds. *PLoS One*, 5(2):e9391.

Stubbs, M. T.; Huber, R. e Bode, W. (1995). Crystal structures of factor xa specific inhibitors in complex with trypsin: structural grounds for inhibition of factor xa and selectivity against thrombin. *FEBS Lett*, 375(1-2):103–7.

Tuncbag, N.; Gursoy, A. e Keskin, O. (2011). Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. *Phys Biol*, 8(3):035006.

Turk, B. (2006). Targeting proteases: successes, failures and future prospects. *Nature Reviews*, (5):785–799.

Voet, D. e Voet, J. G. (2002). *Fundamentals of Biochemistry*. ArtMed, 2 edição.

Wallace, A.; Laskowski, R. e Thornton, J. (1996). Derivation of 3d coordinate templates for searching structural databases: application to ser-his-asp catalytic triads in the serine proteinases and lipases. *Protein Science*, 5(6):1001–1013.

Wang, Y.; Luo, W. e Reiser, G. (2008). Trypsin and trypsin-like proteases in the brain: Proteolysis and cellular functions. *Cell. Mol. Life Sci*, pp. 237–252.

Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M. e Huang, B. (2011). Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, 27(15):2083–88.