

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

Nilma Rodrigues Alves

**MAPEAMENTO DE CORRESPONDÊNCIAS HIDROFÓBICAS EM
COMPLEXOS SERINO PEPTIDASES E INIBIDORES PROTEICOS
ATRAVÉS DA VARREDURA DE AGRUPAMENTO ESPECTRAL**

Belo Horizonte
2015

Nilma Rodrigues Alves

**MAPEAMENTO DE CORRESPONDÊNCIAS HIDROFÓBICAS EM
COMPLEXOS SERINO PEPTIDASES E INIBIDORES PROTEICOS
ATRAVÉS DA VARREDURA DE AGRUPAMENTO ESPECTRAL**

Versão Final

Tese apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutora em Bioinformática.

Orientador: Dr. Carlos Henrique da Silveira

Belo Horizonte
2015

043

Alves, Nilma Rodrigues.

Mapeamento de correspondências hidrofóbicas em complexos serino peptidases e inibidores proteicos através da varredura de agrupamento espectral [manuscrito] / Nilma Rodrigues Alves. – 2015.

216 f. : il. ; 29,5 cm.

Orientador: Dr. Carlos Henrique da Silveira.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Interações Hidrofóbicas e Hidrofílicas. 3. Inibidores da Tripsina. 4. Subtilisina. I. Silveira, Carlos Henrique da. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004



ATA DA DEFESA DE TESE

Nilma Rodrigues Alves

66/2015
entrada
1º/2011
CPF:
003.184.116-36

Às quatorze horas do dia **29 de outubro de 2015**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado de Programa, para julgar, em exame final, o trabalho intitulado: "**Mapeamento de correspondências hidrofóbicas em complexos de rino peptidases e inibidores proteicos através da varredura de agrupamento espectral**", requisito para obtenção do grau de Doutora em **Bioinformática**. Abrindo a sessão, o Presidente da Comissão, **Dr. Carlos Henrique da Silveira**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dr. Carlos Henrique da Silveira	UNIFEI	58794549672	APROVADA
Dra. Raquel Cardoso de Melo Minardi	UFMG	046454366-51	aprovado
Dr. Lucas Bleicher	UFMG	813.528.593-05	APROVADO
Dr. Marcos Augusto dos Santos	UFMG	274585106-44	APROVADO
Dr. José Maurício Schneedorf Ferreira da Silva	UNIFAL	683584206-68	APROVADO
Dra. Valdete M. Gonçalves de Almeida	UNIBH	036675-226-06	aprovado

Pelas indicações, a candidata foi considerada: APROVADA
O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 29 de outubro de 2015.

Dr. Carlos Henrique da Silveira - Orientador _____
Dra. Raquel Cardoso de Melo Minardi _____
Dr. Lucas Bleicher _____
Dr. Marcos Augusto dos Santos _____
Dr. José Maurício Schneedorf Ferreira da Silva _____
Dra. Valdete Maria Gonçalves de Almeida _____

Dr. Vasco Ariston de C. Azevedo
Prof. Titular e Coordenador do Programa de
Pós Graduação em Bioinformática
ICB/UFMG

Agradecimentos

Em tudo dai graças! (Tes 5, 18)

À Deus, fonte de todo amor e bem.

Aos meus queridos pais José Rodrigues da Silva e Terezinha de Jesus Alves Silva. Aos meus irmãos, Geraldo, Afonso, Dulce, Wânia, Gisele, Luciene, Gizelda. Aos sobrinhos, Daniel, Daniela, Ana, Ângelo, Sara. Sempre presentes.

Ao querido professor Marcelo Matos Santoro (*In memoriam*).

Ao meu orientador Carlos Henrique da Silveira, por tudo.

Ao colegiado do curso.

A todos do Laboratório de Bioinformática e Sistemas (LBS), em particular à professora Raquel Cardoso de Melo-Minardi, Valdete Maria Gonçalves de Almeida e Sabrina Azevedo.

Aos professores João Romanelli (UNIFEI-Itabira) e Leonardo Lima (UFSJ-Sete Lagoas).

Ao professor Lucas Bleicher (UFMG).

À Sheila Santana pela simpatia e profissionalismo na secretaria do curso.

À Moema Monteiro, Elisa Boari, Lucas Saraiva, Sandro Isidoro, Wellisson, Tiago Mendes, Rodrigo Kato, Ângelo Bruno, Fábio Mendes.

Às amigas de sempre Roberta Ribeiro, Ieda Reis, Laura Castro, Gleice, aos amigos Pe. Douglas Arão e Roger Gomes.

À Marcelus Virgílius e colegas de trabalho José Muniz, Monique, Wladimir Pavão.

O que faz o mundo não são as coisas, são as relações. (Rubem Alves)

Resumo

Na formação de complexos proteína-proteína, as interações hidrofóbicas desempenham um papel importante no reconhecimento molecular. Neste trabalho, foram analisadas as correspondências hidrofóbicas entre superfícies de serino peptidases e seus inibidores protéicos. Cada *patch* hidrofóbico foi determinado em nível atômico como sendo constituído de átomos previamente classificados como apolares e polares, e organizados em uma rede de contatos por meio de grafos. Tanto o mapeamento das interfaces quanto os pesos das interações entre átomos das peptidases com respectivos inibidores foram definidos pela metodologia Silveira-Romanelli (SR) com base nas áreas de contatos. Para a identificação das correspondências, foi aplicado o agrupamento espectral utilizando matrizes Laplacianas normalizadas, representativas de 36 complexos não-redundantes de enzimas do tipo tripsina e tipo subtilisina, e respectivos inibidores, filtrados do PDB, com informações do MEROPS e PFAM. Por meio de uma varredura realizada variando-se o número de agrupamentos e considerando-se o coeficiente de silhueta como medida de qualidade dos mesmos, foi possível chegar a um modelo geral, onde cada complexo apresenta de 2 a 6 regiões hidrofóbicas complementares entre enzima-inibidor. Essa hidrofobicidade inerente a todas as regiões ajudaria a explicar o sucesso do trabalho pioneiro de SCHECHTER & BERGER - 1967, em usar polialaninas (um peptídeo todo hidrofóbico) como sondas para os primeiros mapeamentos das interfaces em peptidases. Nos 36 complexos analisados, as regiões também parecem formar uma superestrutura anelar ou semianelar hidrofóbica, algo semelhante aos *O-rings* identificados por BORGAN & THORN - 1998, em seu clássico artigo sobre a organização de *hot spots* em interfaces proteína-proteína. Essa superestrutura anelar hidrofóbica em enzimas poderia ter efeitos sobre as águas de solvatação, de modo a interferir no padrão das flutuações de densidade local do solvente, algo que CHANDLER - 2005 e outros autores vêm demonstrando ser fundamental para as complexações proteína-proteína.

Palavras-chave: região hidrofóbica, inibidores de tripsina e subtilisina, agrupamento espectral, estrutura anelar hidrofóbica

Abstract

In the formation of protein-protein complexes, hydrophobic interactions play an important role in molecular recognition. In this study, the hydrophobic correspondences between serine peptidases surfaces and their protein inhibitors were analyzed. Each hydrophobic patch was determined at atomic level featuring atoms previously classified as polar and non-polar, and organized in a graph contact network. Both the mapping of the interfaces and the weight of the interactions between peptidase atoms with their respective inhibitors were set by the Silveira-Romanelli (SR) methodology, based on the areas of contacts. In order to identify correspondences, the spectral clustering using normalized Laplacian matrices was applied. These matrices represented 36 non-redundant complexes of trypsin-like and subtilisin-like enzymes and their respective inhibitors, PDB filtered, with information from MEROPS and PFAM. By means of a sweep carried out altering the number of clustering and considering the silhouette coefficient as a measure of their quality, it was possible to reach a general model, in which each complex shows from 2 to 6 complementary hydrophobic regions between enzyme-inhibitor. Such hydrophobicity inherent to all regions could help explain the success of SCHECHTER & BERGER - 1967's pioneering work, which used polyalanines (an entirely hydrophobic peptide) as probes for the first mappings of the interfaces in peptidases. In the 36 analyzed complexes, the regions also seem to form a hydrophobic ring or semi-annular superstructure, resembling the O-rings identified by BORGAN & THORN - 1998, in their classic paper on the organization of hot spots in protein-protein interfaces. This hydrophobic ring-like superstructure in enzymes might affect solvation water, interfering with the pattern of fluctuation of the solvent local density, a fundamental aspect for the protein-protein complexes, as shown by CHANDLER - 2005 and other authors.

Keywords: hydrophobic patch, trypsin and subtilisin inhibitors, spectral clustering, hydrophobic ring

Lista de Figuras

1.1	Interações hidrofóbicas entre tipo subtilisina e o inibidor Eglina C. O <i>loop</i> do inibidor está em contato próximo com a interface da serino peptidase. Os átomos polares estão em cinza e os apolares em preto. Na interface da enzima, as regiões apolares estão em <i>meshes</i> e esferas (aquelas que estão entre 4 e 6 Å de qualquer átomo apolar do inibidor estão conectadas por arestas pontilhadas). Esperas maiores, em cinza médio, representam os centroides.	23
1.2	Gráfico mostrando as razões ASA apolar/polar da interface e do resto da superfície para as peptidases: 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB (ID's PDB).	25
1.3	Gráfico mostrando a correlação entre mudanças na entropia de solvatação e a extensão do <i>patch</i> para as peptidases: 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB (ID's PDB).	27
3.1	Nomenclatura dos subsítios de uma peptidase e os resíduos complementares de seu substrato - nomenclatura de Schechter e Berger.	32
3.2	Tipos de peptidases e representação do modo de ação. Os círculos azuis representam os resíduos de aminoácidos na cadeia polipeptídica. Os círculos amarelos representam os resíduos terminais e as setas indicam o local da clivagem.	33
3.3	Estrutura tridimensional da Quimotripsina (ID PDB: 1ACB). Tríade catalítica: ASP102 (amarelo), HIS57 (vermelho) e SER195 (azul).	37
3.4	Estrutura tridimensional da subtilisina (ID PDB: 1GCI). Tríade catalítica: ASP32 (amarelo), HIS64 (vermelho) e SER221 (azul).	39
3.5	Alinhamento das sequências do tipo Tripsina (complexo 1PPF contendo enzima <i>Human Leukocyte Elastase de Homo sapiens</i>) e do tipo Subtilisina (complexo 1SBN contendo enzima <i>Subtilisin NOVO BPN</i>). <i>Score</i> de identidade de 19,61% - ClustalO.	39
3.6	Semelhança topológica dos resíduos da tríade catalítica de enzimas tipo Tripsina e tipo Subtilisina. Em (a), uma enzima do tipo Tripsina e sua tríade (complexo 1PPF) e em (b) do tipo Subtilisina e sua tríade (complexo 1SBN).	40
3.7	Mecanismo catalítico das serino peptidases.	41
3.8	Estrutura tridimensional da hirudina extraída da saliva do <i>Hirudo medicinalis</i> (PDB 2HIR). Pontes dissulfeto representadas em verde.	44

3.9	Estrutura tridimensional da Serpina α_1 -antitripsina humana (PDB 1PSI). Os inibidores do tipo Serpina são formado por folhas β (em magenta), oito ou nove α -hélices (cyan) e um <i>loop</i> na porção superior que é a RCL (vermelho).	45
4.1	Trajectoria reativa em vários pontos ao longo do tempo. Cada figura (A, B, C, D) está rotulada na trajetória plotada no espaço unidimensional de r_{12} , indicando a distância entre as partículas.	49
5.1	Exemplos de grafos. Em (a), o grafo pode ser representado por $V = \{a, b, c, d, e\}$ e $E = \{ab, ac, be, cd, de\}$ (ou $E = \{ba, ca, eb, dc, ed\}$, visto que o grafo é não dirigido. A cardinalidade deste grafo ou número de vértices é dada por $n = V = 5$ e o número de arestas por $m = E(G) = 5$. Em (b), $V = \{a, b, c, d, e\}$ e $E = \{aa, ab, be, ca, cd, dc, de\}$, $n = 5$ e $m = 7$. Neste grafo dirigido, também é exemplificada a ocorrência de laço, representado pela aresta <i>aa</i> .	53
5.2	Visão, baseada em grafo, de coesão (à esquerda) e separação (à direita) de clusteres.	62
6.1	Área acessível ao solvente.	65
7.1	Fluxograma para construção da base de dados	76
7.2	Ac para dois átomos vizinhos de raios r_i e r_j .	82
7.3	Condição limite mínimo para Ac	85
7.4	Condição limite máximo para Ac	87
7.5	Matriz de adjacências feita quadrada e simétrica para grafos bipartidos de contatos entre átomos (nós) enzima (ENZ) e inibidor (INB).	89
7.6	Representação simplificada do complexo 1PPF	92
7.7	Grafo da representação simplificada do complexo 1PPF	93
7.8	Matriz de distâncias do grafo da representação simplificada do complexo 1PPF	93
7.9	Matriz de contatos <i>AC</i> representativa das áreas de contatos do modelo simplificado 1PPF	94
7.10	Grafo representativo da matriz de contatos <i>AC</i>	94
7.11	Matriz diagonal <i>D</i> calculada para o modelo simplificado 1PPF	95
7.12	Cálculo da matriz Laplaciana <i>ACL</i> do modelo simplificado 1PPF	96
7.13	Cálculo da matriz Laplaciana Normalizada <i>ACL_{rw}</i> do modelo simplificado 1PPF	97
7.14	Autovalores da matriz Laplaciana Normalizada <i>ACL_{rw}</i> do modelo simplificado 1PPF	98
7.15	Matriz de autovetores obtida a partir da matriz Laplaciana Normalizada <i>ACL_{rw}</i> do modelo simplificado 1PPF	99
7.16	Seleção da submatriz P_k à partir da matriz de autovetores	99
7.17	Processo de obtenção dos grupos com aplicação do <i>k</i> -medoides	100

7.18	Gráfico com coeficientes de silhueta - grupos do modelo simplificado 1PPF . . .	101
7.19	Agrupamento com dois grupos para o modelo simplificado 1PPF	101
7.20	Agrupamento com três grupos para o modelo simplificado 1PPF	102
8.1	Distribuição de densidade para as áreas de contatos dos complexos 1PPF, 1ACB, 1R0R e 1TEC em função da classificação dos átomos. (A) e (B): perfil apolar para o experimento inicial e o final, respectivamente. (C) e (D): perfil polar para o experimento inicial e o final, respectivamente.	104
8.2	Gráficos de distribuição das áreas de contato apolar e polar (36 complexos). . .	109
8.3	1PPF (Elastase de leucócitos humanos - ELY e OMTKY3): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	111
8.4	1ACB (α -Quimotripsina e Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	112
8.5	Exemplo do modelo geral para os complexos 1PPF e 1ACB.	113
8.6	Visão 3D de 6 (k=6) regiões hidrofóbicas correspondentes (enzima-inibidor) para complexos 1PPF e 1ACB.	113
8.7	4B2B (Tripsina com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	114
8.8	1ACB e 4B2B: comparação de clusters.	114
8.9	1F2S (β -tripsina e MCTI-II): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	115
8.10	1PPF e 1F2S: comparação de clusters.	116
8.11	1ACB e 1F2S: comparação de clusters.	116
8.12	1PPF, 1ACB e 1F2S: comparação de clusters (k=4).	117
8.13	1PPF, 1ACB e 1F2S: visão de grafos e estrutura tridimensional (k=4).	117
8.14	1MCT (Tripsina e MCTI-II): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	118
8.15	1F2S e 1MCT: comparação de clusters.	119
8.16	1PPF e 1MCT: comparação de clusters.	119
8.17	1PPE (Tripsina e CMTI-I): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	120
8.18	1F2S e 1PPE: comparação de clusters.	121
8.19	1MCT e 1PPE: comparação de clusters.	121
8.20	1MCT, 1F2S e 1PPE: sobreposição de clusters - Inibidor	122
8.21	1HJA (Quimotripsina C e Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	123
8.22	1PPF e 1HJA: comparação de clusters (k=6 e k=7, respectivamente).	123

8.23	3SGB (<i>Streptomyces griseus proteinase B</i> e OMTKY3): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	124
8.24	1PPF e 3SGB: comparação de clusters considerando apenas o maior componente conexo.	125
8.25	1PPF e 3SGB: comparação considerando todos os componentes conexos.	125
8.26	1PPF e 3SGB: sobreposição - Inibidores.	126
8.27	4SGB (<i>Streptomyces griseus proteinase B</i> e Potato Inhibitor I (PCI-1)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	127
8.28	3SGB e 4SGB: comparação de clusters.	127
8.29	3SGB e 4SGB: sobreposição de clusters - Inibidor	128
8.30	1Z7K (β -Tripsina suína e OMTKY2): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	129
8.31	1PPF e 1Z7K: comparação de clusters.	130
8.32	1PPF e 1Z7K: sobreposição de clusters - Inibidores.	130
8.33	4H4F (Quimotripsina C e Eglina C): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	131
8.34	1ACB e 4H4F: comparação de clusters.	132
8.35	1ACB e 4H4F: comparação de clusters - Inibidores.	132
8.36	1T8O (Quimotripsina A e BPTI): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	133
8.37	1ACB e 4H4F: comparação de clusters.	134
8.38	1ACB e 1T8O: sobreposição de clusters - Inibidores.	134
8.39	1HJA e 1T8O: comparação de clusters - Inibidores.	135
8.40	1HJA e 1T8O: sobreposição de clusters	135
8.41	2FI5 (Tripsina A e BPTI): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	136
8.42	1T8O e 2FI5: comparação de clusters	137
8.43	1T8O e 2FI5: sobreposição de clusters - Inibidores.	137
8.44	1YC0 (<i>Hepatocyte growth factor activator</i> (HGFA) e <i>Kunitz-type protease inhibitor 1</i> (HAI-1)): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	138
8.45	1T8O e 1YCO: comparação de clusters.	139
8.46	1YCO: 3 alças desconexas. 1a:V(G)RCR; 2a: VY(GG)CL; 3a: FP.	139
8.47	1T8O e 1YCO: estruturas tridimensionais com regiões hidrofóbicas. Inibidor em <i>sticks</i>	140
8.48	1T8O e 1YCO: comparação de clusters - Inibidores.	141
8.49	4DG4 (Tripsina III-mesotripsina e BPT): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	142

8.50	2FI5 e 4DG4: comparação de clusters.	142
8.51	1YCO e 4DG4: comparação de clusters.	143
8.52	2FI5 e 4DG4: sobreposição de clusters.	143
8.53	1TAW (Tripsina e APP): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	144
8.54	1T8O e 1TAW: comparação de clusters.	145
8.55	1T8O e 1TAW: sobreposição de clusters.	145
8.56	1FI8 (Granzima B e ECO): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	146
8.57	1PPF e 1FI8S: comparação de clusters	147
8.58	1EZS (Tripsina II e ECO): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	148
8.59	1FI8 e 1EZS: comparação de clusters.	148
8.60	1FI8 e 1EZS: sobreposição de clusters - Inibidores	149
8.61	1XX9 (FXIa e ECO): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	150
8.62	1PPF e 1XX9: comparação de clusters (k=6).	151
8.63	1PPF e 1XX9: comparação de clusters (k=7).	151
8.64	1PPF e 1XX9: comparação de clusters (k=7.	152
8.65	1EZS e 1XX9: sobreposição de clusters - Inibidores.	152
8.66	1FLE (Elastase Pancreática - PPE e Elafin): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters. Na estrutura tridimensional, a região marrom é composta somente por átomos de resíduos <i>noloop</i> (Pro29 e Arg31).	153
8.67	1PPF e 1FLE: comparação de clusters (k=6).	154
8.68	1PPF e 1FLE: comparação de clusters (k=6 e k=8, respectivamente).	154
8.69	2Z7F (Elastase de leucócitos humanos e 1/2SLPi): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	155
8.70	1PPF e 2Z7F: comparação de clusters.	156
8.71	1PPF e 2Z7F: sobreposição de clusters - Inibidores.	156
8.72	4DOQ (Tripsina e 1/2SLPi): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	157
8.73	2Z7F e 4DOQ: comparação de clusters.	158
8.74	Sobreposição de clusters - Inibidores: (A) 2Z7F e 4DOQ, (B)1PPF e 4DOQ.	158
8.75	1FLE e 4DOQ: comparação de clusters.	159
8.76	1FLE e 4DOQ: sobreposição de clusters dos inibidores.	159
8.77	1K9O (Tripsina e ALASERPIN): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	160
8.78	1PPF e 1K9O: comparação de clusters.	161

8.79	1PPF e 1K9O: sobreposição de clusters contendo apenas resíduos dos inibidores. Destaque para o cluster da 1K9O que se interpõe a dois clusters da 1PPF.	161
8.80	1OPH (Tripsina e α 1-antitripsina): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	162
8.81	1PPF e 1OPH: comparação de clusters.	163
8.82	1K9O e 1OPH: comparação de clusters.	163
8.83	1K9O e 1OPH: sobreposição de clusters.	164
8.84	3MYW (Tripsina e BBI): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	164
8.85	1PPF e 3MYM: comparação de clusters.	165
8.86	1PPF e 3MYM: sobreposição de clusters - Inibidores.	165
8.87	3MYW (Tripsina e BBI): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	166
8.88	1PPF e 3VEC: comparação de clusters.	166
8.89	1PPF e 3VEC: sobreposição de clusters.	167
8.90	1KIG (Fator XA e TAP: visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters. Na estrutura tridimensional, os átomos em stiks esverdeados são noloops. O cluster marrom é formado apenas por átomos noloops.	168
8.91	1PPF e 1KIG: comparação de clusters.	169
8.92	1PPF e 1KIG: comparação de clusters lado enzima.	169
8.93	1PPF e 1KIG: comparação de clusters lado inibidor.	170
8.94	Estrutura 3D de 3 regiões hidrofóbicas de 1KIG e comparação das alças inibitórias de 1PPF E 1KIG. A região em marrom além dos resíduos da enzima é formada somente por resíduos fora da alça inibitória.	170
8.95	1R0R (Subtilisina Carlsberg e OMTKY3): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	172
8.96	Visão 3D de 6 (k=6) regiões hidrofóbicas correspondentes (enzima-inibidor) para complexos 1PPF e 1R0R.	172
8.97	1PPF e 1R0R: comparação de clusters.	173
8.98	1TEC (Termitase com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	174
8.99	1R0R e 1TEC: comparação de clusters	175
8.100	1R0R e 1TEC: comparação de clusters no Pymol	175
8.101	2SEC (Carlsberg com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	176
8.102	1TEC e 2SEC: comparação de clusters	177

8.1031SBN (Subtilisina BPN com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	178
8.1041TEC e 1SBN: comparação de clusters	178
8.1051LW6 (BPN e Inibidor de Quimotripsina 2 (CI-2)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	179
8.1061LW6: clusters para $k=6$ e $k=5$	180
8.1071R0R e 1LW6: comparação de clusters.	180
8.1081TM1 (BPN e Inibidor de Quimotripsina 2A (CI-2A)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	181
8.1091LW6 e 1TM1: comparação de clusters	182
8.1101OYV (CAN e Wound-induced proteinase inhibitor-II (WIP)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	183
8.1111R0R e 10YV: comparação de clusters	183
8.1121V5I (Subtilisina BPN e POIA1): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	184
8.1131V5I: (A) e (B) visão da estrutura tridimensional com destaque para região hidrofóbica fora do sítio de especificidade. (C) destaque do sítio de especificidade.	185
8.1141R0R e 1V5I: comparação de clusters.	185
8.1151R0R e 1V5I: sobreposição de clusters.	186
8.1161ROR e 1V5I: sobreposição simplificada dos clusters de 1V5I, ou seja, os clusters de <i>noloops</i> foram excluídos.	186
8.1174GI3 (CAN e Greglina): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.	187
8.1181R0R e 4GI3: comparação de clusters.	188
8.1191F2S e 4GI3: comparação de clusters.	188
8.1201F2S e 4GI3: sobreposição de clusters somente dos inibidores.	188
8.121Relação entre o número de clusters e o Coeficiente médio de Silhueta para todos os 36 complexos. Ponto preto: $s_m < 0$, ou seja, algum átomo ou conjunto de átomos não foram clusterizados adequadamente. Ponto cinza: algum átomo ou conjunto de átomos estão num ponto intermediário entre dois clusters. Ponto branco: os átomos têm valores próximos a 1, portanto bem definidos nos clusters onde se encontram.	190

8.122	Modelo Geral dos Clusteres - composição lado inibidor. (A) Clusterização para enzimas tipo tripsina com diferentes inibidores. (B) Clusterização tipo subtilisina com diferentes inibidores. [] indicam um cluster, {} indicam um componente desconexo, indicam um <i>noloop</i> , ou seja, região no inibidor fora da alça inibitória. Resíduos de aminoácidos com letras minúsculas correspondem a resíduos de fronteira, ou seja, que estão numa posição intermediária entre os clusters.	191
8.123	Modelo clássico de mapeamento de sítios, definido por SCHECHTER e BERGER de 1967.	192
8.124	Exemplo do formato anelar do modelo geral (1F2S e 1PPF).	194

Lista de Tabelas

1.1	Comparação de parâmetros termodinâmicos e estimação empírica para o complexo OMTKY3/PPE. Dado experimental a 25°C.	26
3.1	Famílias de inibidores de serino peptidases com estrutura cristalina resolvida. Inibidores de origem ^a animal, ^v vegetal e de ^m microorganismos.	43
5.1	<i>Coeficiente de silhueta e interpretação de agrupamentos</i>	63
7.1	Tabela de inibição cruzada com Eglina C	75
7.2	Tabela de inibição cruzada com Ovomucoide (OMTKY3)	75
7.3	Classificação dos inibidores segundo o Merops. O inibidor WCI da família I3 pertence a subfamília A. Os demais inibidores selecionados não possuem subfamílias.	78
7.4	Classificação das enzimas segundo o Merops. As enzimas tipo tripsina selecionadas pertencem ao subclã PA(S), enquanto as tipo subtilisinas não são agrupadas em subclãs. Todas as enzimas estão classificadas na subfamília A, exceto a tipo tripsina SGY ou SGPB dos complexos 3SGB e 4SGB, cuja subfamília é E.	79
7.5	Relação de inibidores e enzimas em complexo	80
7.7	Raios de van der Waals (R_{vdw})	89
7.8	Classificação dos átomos - subclasses	91
8.1	Perfil Apolar	105
8.2	Perfil Polar	105
8.3	Classificação dos átomos - classes	108
8.4	Sequências P4 a P3' da região do sítio ativo de OMTKY2 e OMTKY3. P3 é a posição conservada entre esses domínios.	129
8.5	Alça dos inibidores dos complexos 1FLE,4DOQ, 2Z7F	159

Lista de abreviaturas e siglas

APP	Protease Inhibitor Domain Of Alzheimer's Amyloid Beta-Protein Precursor
ASA	Accessible Surface Area - Área da superfície acessível ao solvente
BBI	Bowman-Birk Inhibitor
BPN	Subtilisina BPN'
BPT	Pancreatic Trypsin Inhibitor
CAN	Subtilisina Carlsberg
CHY	Quimotripsina
CL2	Chymotrypsin Inhibitor 2
EC	Number Enzyme Classification
ECO	Ecotina
EGL	Eglina
ELY	Human Leukocyte Elastase
GRG	Greglin
GRY	Granzym B
HGY	Hepatocyte growth factor activator
OMTKY	Turkey Ovomuroid
OVO	Ovomucoide
PPI	Protein-protein interaction
POA	POIA1 , IA-1, Serine proteinase inhibitor
POT	Potato
PPE	Porcine Pancreatic Elastase
RCL	Reactive Centre Loop
SER	Serpina
SGY	Streptomyces griseus proteinase B
SKY	Skin-derived antileukoprotease
SLP	Secretory Leucoprotease Inhibitor
SQF	Squash Family
THN	Thermitase
TRY	Tripsina
WAP	Whey Acidic Protein
WCI	Chymotrypsin inhibitor 3
WIP	Wound-induced proteinase inhibitor-II
XAY	FXIa

Sumário

1	Introdução	21
1.1	Motivação Teórica e Experimental	24
2	Objetivos	29
2.1	Objetivo Geral	29
2.1.1	Objetivos Específicos	29
3	Serino peptidases e seus inibidores proteicos	30
3.1	Peptidases	30
3.1.1	Classificação das Peptidases	32
3.1.1.1	Peptidases agrupadas pelo tipo de reação química	32
3.1.1.2	Peptidases agrupadas pelo mecanismo de ação da catálise	34
3.1.1.3	Peptidases agrupadas pela estrutura e homologia	35
3.2	Serino Peptidases	36
3.2.1	Enzimas Tipo Tripsinas	37
3.2.2	Enzimas Tipo Subtilisinas	38
3.2.3	Mecanismo catalítico clássico	38
3.3	Inibidores de Serino Peptidases	41
3.3.1	Inibidores canônicos	42
3.3.2	Inibidores não canônicos	44
3.3.3	Serpinas	45
4	Hidrofobicidade e formação de complexos proteína-proteína	46
4.1	Hidrofobicidade	46
4.1.1	Influência do tamanho das partículas hidrofóbicas	47
4.1.2	Flutuações da densidade do solvente	48
5	Agrupamento espectral em grafos	51
5.1	Notação matemática	51
5.2	Apresentação	51
5.3	Grafos: conceitos básicos	52
5.4	Grafos: particionamento	54
5.4.1	Corte de grafos (Cuts)	55
5.5	Agrupamento Espectral de Grafos	56

5.5.1	Propriedades das matrizes Laplacianas	57
5.5.1.1	Matriz Laplaciana não Normalizada	58
5.5.1.2	Matriz Laplaciana Normalizada	59
5.5.2	Algoritmo para agrupamento espectral	60
5.5.2.1	Algoritmo K -Medoides	61
5.5.2.2	Avaliação dos grupos formados	61
6	Trabalhos correlacionados	64
6.1	Cálculo da superfície de proteínas	64
6.2	Cálculo da superfície de contato entre proteínas	66
6.3	Identificação de regiões hidrofóbicas em proteínas	66
6.4	Agrupamento espectral	68
7	Materiais e Métodos	71
7.1	Bases de dados utilizadas	71
7.1.1	PDB - <i>Protein Data Bank</i>	71
7.1.2	MEROPS	72
7.1.3	PFAM	73
7.2	Scripts utilizados	74
7.3	Construção da base de dados	74
7.3.1	Classificação das enzimas e inibidores	77
7.3.2	Normalização dos dados	77
7.4	Cálculo da área de contato entre átomos da enzima e inibidor	81
7.4.1	Cálculo da Ac utilizando a ASA	81
7.4.2	Cálculo da Ac utilizando SR	83
7.4.2.1	Limite inferior para Ac	84
7.4.2.2	Limite superior para Ac	86
7.5	Cálculo das distâncias e das matrizes de similaridades	88
7.6	Classificação dos átomos	90
7.7	Agrupamento espectral	91
7.7.1	Exemplo	92
7.7.1.1	Passo 1: Geração da matriz de distâncias	93
7.7.1.2	Passo 2: Geração da matriz de áreas de contatos AC	93
7.7.1.3	Passo 3: Criação da matriz Laplaciana Normalizada $ACLrw$	94
7.7.1.4	Passo 4: Decomposição espectral da matriz Laplaciana Normalizada $ACLrw$	97
7.7.1.5	Passo 5: Seleção de uma submatriz P_k contendo apenas os k últimos autovetores correspondentes aos k menores autovalores	99

7.7.1.6	Passo 6: Aplicação de k -medoides para rotular k grupos em P_k	100
8	Resultados e discussão	103
8.1	Análises da Inibição Cruzada	103
8.2	Definição dos tipos de átomos	107
8.3	Distribuição estendida para os 36 complexos	108
8.4	Regiões hidrofóbicas em cada complexo	109
8.4.1	Complexos Tipo Tripsinas e Inibidores	110
8.4.1.1	Complexo 1PPF	111
8.4.1.2	Complexo 1ACB	112
8.4.1.3	Complexo 4B2B	114
8.4.1.4	Complexo 1F2S	115
8.4.1.5	Complexo 1MCT	118
8.4.1.6	Complexo 1PPE	120
8.4.1.7	Complexo 1HJA	122
8.4.1.8	Complexo 3SGB	124
8.4.1.9	Complexo 4SGB	126
8.4.1.10	Complexo 1Z7K	128
8.4.1.11	Complexo 4H4F	131
8.4.1.12	Complexo 1T8O	133
8.4.1.13	Complexo 2FI5	136
8.4.1.14	Complexo 1YCO	138
8.4.1.15	Complexo 4DG4	141
8.4.1.16	Complexo 1TAW	144
8.4.1.17	Complexo 1FI8	146
8.4.1.18	Complexo 1EVS	147
8.4.1.19	Complexo 1XX9	150
8.4.1.20	Complexo 1FLE	153
8.4.1.21	Complexo 2Z7F	155
8.4.1.22	Complexo 4DOQ	157
8.4.1.23	Complexo 1K9O	160
8.4.1.24	Complexo 1OPH	162
8.4.1.25	Complexo 3MYW	164
8.4.1.26	Complexo 3VEQ	166
8.4.1.27	Complexo 1KIG	168
8.4.2	Subtilisinas	171
8.4.2.1	Complexo 1R0R	171
8.4.2.2	Complexo 1TEC	174

8.4.2.3	Complexo 2SEC	176
8.4.2.4	Complexo 1SBN	177
8.4.2.5	Complexo 1LW6	179
8.4.2.6	Complexo 1TM1	181
8.4.2.7	Complexo 1OYV	182
8.4.2.8	Complexo 1V5I	184
8.4.2.9	Complexo 4GI3	187
8.5	Clusteres - Modelo Geral	189
8.5.1	Varredura de Agrupamentos	189
8.5.2	Alinhamento por Agrupamento	190
8.5.3	Superestrutura Hidrofóbica Anelar ou Semianelar	193
9	Conclusões	196
9.1	Perspectivas	197
9.1.1	Padrões de Distribuição de Contatos	197
9.1.2	Estatística das Métricas de Agrupamento	198
9.1.3	Alinhamento dos Agrupamentos	198
9.1.4	Visualização dos Modelos	198
9.1.5	Aspectos Dinâmicos	199
9.1.6	Superestrutura Hidrofóbica Anelar	199
9.1.7	Outras Peptidases	199
	Referências	200
	Apêndice A Método Silveira-Romanelli	211

Capítulo 1

Introdução

O reconhecimento entre moléculas biológicas para formação de associações que desempenham papéis importantes na manutenção da vida é um processo fundamental na biologia.

O reconhecimento molecular e a associação proteína-proteína têm sido estudados em termos de vários aspectos, tais como a natureza das interações envolvidas (forças eletrostáticas, van der Waals etc), as correlações entre entropia e hidrofobicidade (em certos casos, indiretamente mapeadas pelo tamanho da área acessível ao solvente (ASA, do inglês *Accessible Surface Area*)), a forma e a complementaridade entre as superfícies, a composição das interfaces, efeitos dinâmicos na formação dos complexos, as redes de interações intrasolvente, especialmente das águas de solvatação etc [[Chothia and Janin \(1975\)](#), [Korn and Burnett \(1991\)](#)].

Neste trabalho, foram analisadas as correspondências hidrofóbicas entre superfícies de peptidases e seus inibidores proteicos, visto que nessa associação a hidrofobicidade em suas superfícies parece desempenhar papel fundamental [[Young et al. \(1994\)](#)].

As peptidases constituem um grupo numeroso de proteínas que representam aproximadamente 2% do número total de proteínas presentes em todos os tipos de organismos [[Polgar \(2005\)](#)]. São constituintes essenciais em todas as formas de vida, incluindo microorganismos como vírus, bactérias, protozoários, leveduras, fungos, as plantas e os animais.

Desde os primeiros estudos envolvendo a composição química das proteínas em termos de seus aminoácidos, ficou evidente que boa parte de suas cadeias laterais eram compostas por átomos apolares. Átomos nessa condição tendem a ter uma densidade eletrônica mais homogênea ou isotrópica, ao contrário dos átomos polares, em que essa densidade pode variar, criando subregiões mais densas em elétrons que outras (formando “polos”, daí o termo “polarizado” ou polar). Apesar dessa anisotropia de densidade eletrônica, tais átomos polares ainda estão neutros, com números equivalentes de prótons e elétrons. Podemos considerar um extremo de polarização quando os átomos apresentam carga formal diferente de zero, ou seja, estão ionizados ou carregados com elétrons a mais ou a menos em relação ao número de prótons.

Entre 20% a 30% dos resíduos em uma proteína compreendem aminoácidos como valina, leucina, isoleucina ou fenilalanina, cujas cadeias laterais são arranjos de grupos

metis essencialmente apolares [Kauzmann (1959)]. Com exceção da glicina, que não tem cadeia lateral, mesmo aminoácidos tipicamente tidos como polares, apresentam algum átomo apolar ou pouco polar, como os carbonos betas em todos os demais aminoácidos, e outros grupos metis, como os que antecedem o grupo amino na lisina, classificada como um resíduo carregado positivamente (em pH neutro). Isso sem contar enxofres pouco polares, como o tioeter na metionina e o tiol oxidado nas pontes dissulfetos [Kauzmann (1959)].

Como esses grupos não-polares tendem a ter baixa afinidade pela água (uma molécula polar), há uma tendência natural deles se agregarem entre si, criando uma fase própria não-aquosa, o que pode levar à formação de complexas estruturas em proteínas e outras biomoléculas. Chama-se essa tendência de “efeito hidrofóbico” ou “interação hidrofóbica” [Kauzmann (1959)], e falaremos bastante sobre esse tema ao longo desta tese. Por outro lado, os átomos polares (ou carregados) tendem a interagir bem com a água, e por conta disso, são tidos como hidrofílicos.

[Korn and Burnett (1991)], ao estudarem as colaborações das forças hidrofóbicas e hidrofílicas na ligação entre interfaces de proteínas oligoméricas e multiméricas (com várias cadeias agregadas entre si), apontaram quantitativamente a complementaridade hidropática entre as proteínas estudadas e mostraram qualitativamente, por meio de imagens coloridas com uma escala hidropática, que essa complementaridade significa que as proteínas interagem colocando seus centros hidrofóbicos da superfície contra os centros hidrofóbicos de outra superfície e o mesmo acontece com os centros hidrofílicos.

Nosso grupo de pesquisa, em trabalho publicado em 2012, Gonçalves-Almeida et al. (2012) aprofundou o trabalho de Korn e coautores, mas com foco especial no fenômeno de inibição cruzada, que ocorre tanto quando um mesmo inibidor proteico inibe proteases com estruturas dimensionais bem diferentes e com baixa identidade de sequência, como tripsinas e subtilisinas, como também quando uma mesma protease é inativada ou bloqueada por diferentes inibidores. Estabelece-se, pois, uma relação n para n entre proteases e inibidores. Exemplos clássicos envolvem os inibidores ovomucoide e eglinas, ambos capazes de inibir tanto proteases tipo tripsinas quanto tipo subtilisinas [Bode et al. (1986), Horn et al. (2003), McPhalen and James (1988), Frigerio et al. (1992)].

Para aquela investigação, foi construída a metodologia *Hydropace* para identificar padrões conservados de interação hidrofóbica em serino peptidases, em nível atômico. Estudos anteriores, por trabalharem no nível de resíduos, não evidenciavam padrões relevantes. Na metodologia *Hydropace*, ao abstrair os resíduos em prol de uma granulosidade atômica, vários padrões emergiram de forma inesperada. Esses padrões foram associados à regiões invariantes (ou *patches*) no lado enzima. Essas regiões poderiam compreender, numa peptidase, parte de um único volumoso resíduo aromático; mas em outras, átomos de dois ou mais pequenos resíduos com cadeias laterais alifáticas. Uma análise que não fosse por regiões mas por resíduos teria dificuldades de destacar tais correspondências.

Os autores em [Gonçalves-Almeida et al. \(2012\)](#) também pressuporam que essas regiões estariam em contato com regiões apolares complementares no lado do inibidor, sem no entanto, comprová-la.

Após a submissão do artigo à *Bioinformatics*, um dos revisores questionou essa pressuposição, apontando que não estava evidente na metodologia que os inibidores tinham de fato *patches* complementares que interagissem com os respectivos sítios hidrofóbicos da enzima, e qual seria (se existisse) sua contribuição para a inibição cruzada. Abaixo, são transcritos a indagação do revisor e a resposta dos autores.

Revisor: *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.*

Argumentação do autores: *If the binding energetics described above for protease-inhibitors are (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 1.1 an example of correspondences among the apolar atoms of the inhibitors side and enzymes side.*

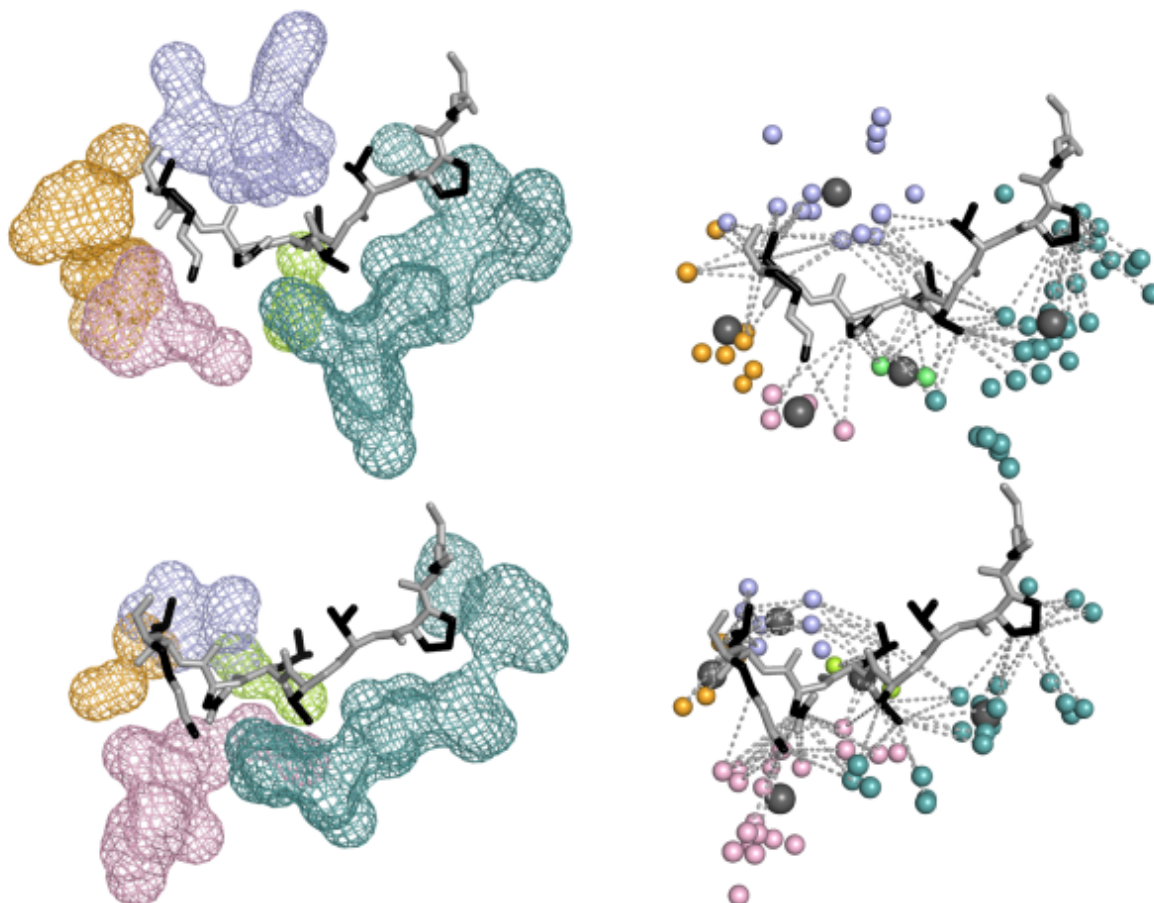
Mesmo sendo um único exemplo sem nenhuma análise computacional e estatística mais elaborada, os revisores aceitaram a publicação. É provável que tenham entendido que não era o objetivo maior daquele trabalho aprofundar tal correspondência. Seja como for, para os autores do artigo (o que inclui o orientador desta tese), a questão não havia sido satisfatoriamente respondida. Ela foi temporariamente colocada na “gaveta das ideias”, algo a ser enfrentado num trabalho futuro. Futuro este agora presente na forma desta tese.

Então, com o objetivo de identificar as regiões hidrofóbicas do lado do inibidor e estabelecer as referidas correspondências com o lado enzima, este trabalho consistiu em tratar esse tema como objeto de pesquisa. Inicialmente foi aplicada a própria metodologia Hydropace, mas devido ao menor número de átomos do lado do inibidor, o método não se mostrou suficiente. Com isso, foi escolhido o agrupamento espectral acoplado a um grafo de áreas de contatos, algo que será descrito em detalhes ao longo desta tese.

1.1 Motivação Teórica e Experimental

O trecho a seguir foi baseado na resposta dada aos questionamentos dos revisores do referido artigo da *Bioinformatics* [Gonçalves-Almeida et al. \(2012\)](#). Dá destaque a alguns dos conceitos que embasaram a busca por correspondências hidrofóbicas nas interfaces entre peptidases e inibidores proteicos. As reflexões e análises prévias feitas neste excerto

Figura 1.1: Interações hidrofóbicas entre tipo subtilisina e o inibidor Eglina C. O *loop* do inibidor está em contato próximo com a interface da serino peptidase. Os átomos polares estão em cinza e os apolares em preto. Na interface da enzima, as regiões apolares estão em *meshes* e esferas (aquelas que estão entre 4 e 6 Å de qualquer átomo apolar do inibidor estão conectadas por arestas pontilhadas). Esferas maiores, em cinza médio, representam os centroides.



Fonte: [Gonçalves-Almeida et al. (2012)].

motivaram grande parte dos aprofundamentos teóricos e dos desenhos experimentais *in silico* que estruturaram esta tese.

“As áreas das superfícies acessíveis estão correlacionadas com energias livres [Chothia and Janin (1975)], portanto, com princípios termodinâmicos. Certamente que [Kauzmann (1959)], no final da década de 1950, foi um dos primeiros a indicar de forma convincente as correlações entre as interações hidrofóbicas (envolvendo átomos não-polares) e a entropia, em especial do solvente, fato verificado experimentalmente em diversos fenômenos, como por exemplo: a dissolução das moléculas orgânicas, enovelamento proteico, estabilização dos complexos proteína-ligante, a agregação proteica e outros.

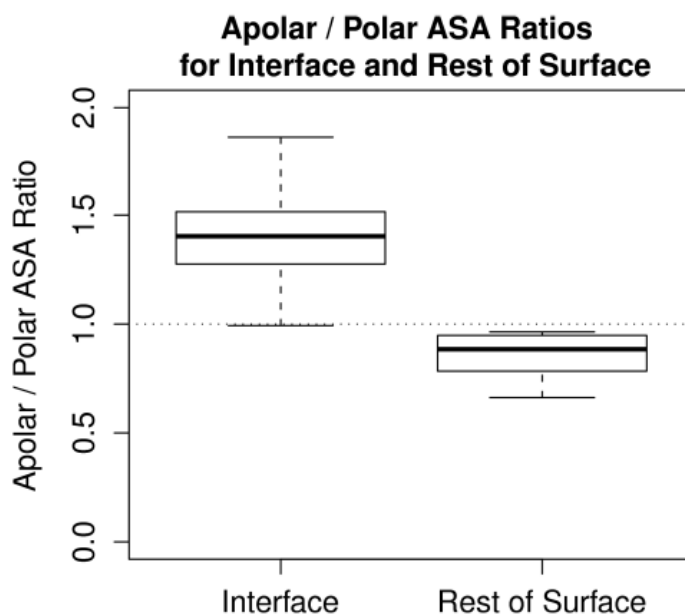
O grau de exposição dos átomos apolares ou moléculas ao solvente é fundamental para o conceito de hidrofobicidade. [Lee and Richards (1971)] desenvol-

veram um método (hoje clássico) para calcular essa exposição da proteína ao solvente. Eles criaram o conceito de área de acessibilidade ao solvente (ASA) como a soma de todas as regiões de um átomo ou grupo de átomos que podem ser alcançadas por uma esfera (a probe) de raio R (representando o solvente) em contato com a proteína através da superfície de Van der Waals. No início da década de 1970, [Chothia (1974)] demonstrou que existe uma correlação linear entre a ASA e a solubilidade dos aminoácidos em solventes orgânicos (uma medida de hidrofobicidade). Ele estimou a contribuição da energia livre entre 20 e trinta calorias por mol^{-1} para cada Å^2 da área dos átomos apolares que não estão expostos aos solvente.

[Chothia and Janin (1975)] evidenciaram que a razão apolar/polar da ASA aparenta ser maior para a interface em alguns complexos proteicos que para o resto da superfície. Esta maior hidrofobicidade da região na interface facilitaria, em uma perspectiva termodinâmica, a agregação com regiões apolares de outras cadeias ou moléculas. Além disso, uma proporção balanceada da ASA apolar/polar em regiões fora da interface colaboraria para a prevenção de aglutinações disfuncionais. Sempre bom lembrar que, a polimerização da hemoglobina S na anemia falciforme é induzida pelo desequilíbrio nessa razão apolar/polar, devido a uma simples mutação em um resíduo polar por outro apolar (Glu pela Val) na superfície da cadeia beta [Dickerson and Geis (1983)].

[Gonçalves-Almeida et al. (2012)] apontaram uma tendência de aumento da razão apolar/polar da ASA na interface em detrimento do resto da superfície (Figura 1.2), uma evidência da importância das interações hidrofóbicas na formação do complexo proteína-inibidor.

Figura 1.2: Gráfico mostrando as razões ASA apolar/polar da interface e do resto da superfície para as peptidases: 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB (ID's PDB).



Fonte: [Gonçalves-Almeida et al. (2012)].

[Murphy and Freire (1992)] demonstraram que pode ser confiável estimar alguns parâmetros termodinâmicos através de cálculos da ASA apolar e polar, a partir das coordenadas atômicas. Isso pressupõe que os entes nos complexos comportam-se como corpos rígidos, com presença desprezível de rearranjos dinâmicos ou alostéricos.

[Baker and Murphy (1997)] aplicaram esse método para estimar alguns desses parâmetros envolvendo a variação da energia livre de ligação (ΔG de *binding*) do inibidor ovomucoide (OMTKY3 - *Turkey Ovomucoid Third Domain*) em complexo com elastase pancreática de porco (PPE - *Porcine Pancreatic Elastase*). Eles também compararam os cálculos empíricos da ASA com os dados experimentais medidos por ITC (*Isothermal Titration Calorimetry*). A Tabela 1.1 abaixo contrasta cálculos empíricos e experimentais, todos em boa concordância, exceto, talvez, a variação de entalpia. Mesmo assim, esse parâmetro tem peso pequeno na variação da energia livre total (somente 4%). Portanto, pode-se concluir que a formação do complexo OMTKY3 - PPE é entrópica dirigida, pondo em destaque a importância dos casamentos ou pareamentos hidrofóbicos na interface.

Se se assume que essas premissas se mantêm válidas para a maioria dos outros complexos peptidases-inibidores (uma hipótese plausível), a metodologia de Murphy e co-autores pode ser aplicada para estimar a variação de entropia

Tabela 1.1: Comparação de parâmetros termodinâmicos e estimação empírica para o complexo OMTKY3/PPE. Dado experimental a 25°C.

Parâmetro	Experimental	Calculado
$\Delta C_p^o (kJK^{-1}mol^{-1})$	-1.1 ± 0.1	-1.4
$\Delta H^o (kJK^{-1}mol^{-1})$	-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

Fonte: [Baker and Murphy (1997)].

de solvatação dos complexos estudados por meio dos cálculos da ASA. Essa escolha, então foi feita, porque a mudança da entropia de solvatação é o parâmetro mais relevante das interações hidrofóbicas e é o fator mais predominante envolvido na ligação proteína-proteína.

[Baker and Murphy (1997)]. Dois passos podem ser realizados:

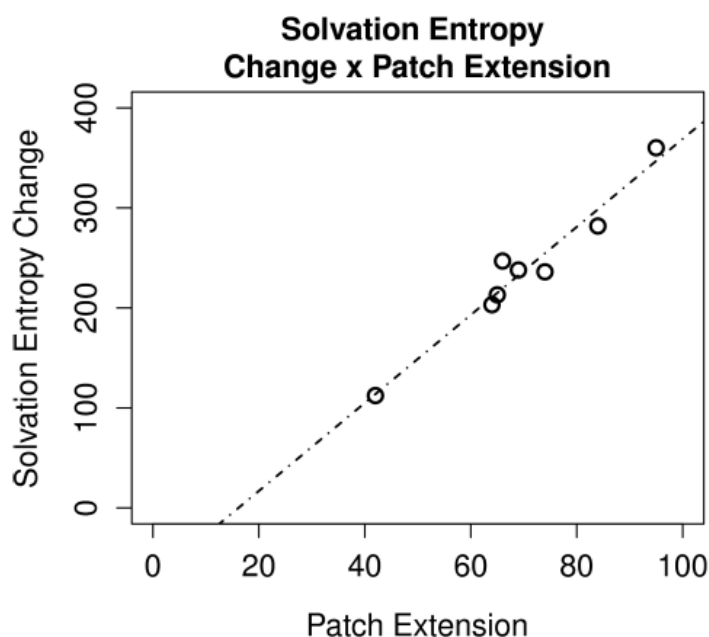
1. determinar a variação da capacidade calorífica (pressão constante) como:

$$\Delta C_p = a \cdot \Delta ASA_{apolar} + b \cdot \Delta ASA_{polar};$$
2. determinar a variação da entropia de solvatação: $\Delta S = \Delta C_p \cdot \ln\left(\frac{T}{T_s}\right)$.

Onde, a e b são parâmetros de ajuste, conforme estimados em [Murphy and Freire (1992)] como $1.88 \text{ JK}^{-1}\text{mol}^{-1}\text{A}^{-1}$ e $-1.09 \text{ JK}^{-1}\text{mol}^{-1}\text{A}^{-1}$, respectivamente. T é a temperatura experimental ou ambiente e T_s é a temperatura de referência onde a variação de entropia é considerada como zero (em torno de 385K) [Baldwin (1986)].

Na Figura 1.3, é possível ver que existe uma forte correlação linear (coeficiente de Pearson de 0.98) entre a variação da entropia de solvatação (inferido como descrito acima) e a “extensão” dos *patches*, medida pelo número de átomos hidrofóbicos que os compõe. O intercepto parece se distanciar de zero porque a variação da capacidade calorífica, como sugerido na metodologia de Murphy e co-autores, tem um termo polar.

Figura 1.3: Gráfico mostrando a correlação entre mudanças na entropia de solvatação e a extensão do *patch* para as peptidases: 1ACB 1CSE 1SBN 1TEC 1R0R 1PPF 1CHO 3SGB (ID's PDB).



Fonte: [Gonçalves-Almeida et al. (2012)].

Concluindo, acredita-se que os *patches* hidrofóbicos conservados têm um papel bem definido na inibição. Eles podem representar a rede de interações hidrofóbicas que organiza (e talvez até codifique) as interfaces das peptidases, indo além das descrições tradicionais em termos de tríades/díades catalíticas e sítios de estabilização do oxianion. Suas interfaces têm um fator hidrofóbico que é claramente distinto do resto da superfície, como mostrado na Figura 1.2. O estudo energético da elastase PPE calculada por [Baker and Murphy (1997)] indica que a interação entre proteases e inibidores pode ser orquestrada fundamentalmente por forças hidrofóbicas/entrópicas, chegando até a 96% da variação da energia livre na formação do complexo. Corroborando com essas afirmativas, os *patches* hidrofóbicos também têm uma correlação evidente com a variação da entropia de solvatação, como mostra a Figura 1.3.

Cabe salientar, no entanto, que a existência de padrões hidrofóbicos pode ser uma condição necessária mas talvez não suficiente para o reconhecimento. As interações polares, as complementaridades estereoquímicas, além da estrutura das águas nas primeiras camadas de solvatação também podem ter sua importância. Na verdade, tudo isso se correlaciona e se retroalimenta, como indicam vários estudos dos quais falaremos mais adiante.

Mesmo considerando apenas o fator apolar, parece também que ainda não há consenso na literatura sobre o exato papel das interações hidrofóbicas em pro-

teínas. Enquanto [Chothia and Janin (1975)] argumenta que *as hydrophobic contribution is entirely unpecific, it would lead to all kinds of incorrect interactions in a cell*, [Dill (1999)] defende que *hydrophobic interactions are not nonspecific glue, but a crucial structure-determining driving force*“.

Independente de quem tem razão, o primeiro ataque do nosso grupo de pesquisa a essa questão [Gonçalves-Almeida et al. (2012)] teve o mérito de colocar em evidência a existência de padrões hidrofóbicos robustos no lado enzima (se olhados no nível atômico), mesmo quando considerou-se peptidases tão díspares quanto Tripsinas e Subtilases. Também ficou evidenciado que esses padrões podem ajudar a explicar o fenômeno de inibição cruzada. O que esta tese almejou foi estender ainda mais essa análise, agora com um olhar especial para o lado inibidor, dando ênfase nas correspondências hidrofóbicas com o lado peptidase.

Capítulo 2

Objetivos

2.1 Objetivo Geral

Identificar a correspondência hidrofóbica entre complexos formados por serino peptidases e inibidores protéicos através de técnicas de varredura de agrupamento espectral, tendo a área de contato como parâmetro para inferir indiretamente as proximidades e/ou interações atômicas.

2.1.1 Objetivos Específicos

- Construir uma base de dados com complexos formados por enzimas tipo tripsina e tipo subtilisina e inibidores protéicos, com representantes em várias famílias de inibidores e complexos com baixa redundância;
- Desenvolver algoritmos em Perl e R para identificação de regiões hidrofóbicas na enzima e no inibidor e identificação automática da correspondência entre essas regiões;
- Montar os grafos tendo como peso entre as arestas as áreas de contato entre pares atômicos da enzima e do inibidor como medida do grau de proximidade e/ou interação entre ambos;
- Classificar os átomos conforme o perfil apolar/polar;
- Aplicar técnicas de agrupamento espectral com o objetivo de identificar subregiões de correspondências hidrofóbicas.

Capítulo 3

Serino peptidases e seus inibidores proteicos

3.1 Peptidases

As peptidases são enzimas que catalisam reações de hidrólise de ligações peptídicas de proteínas. Também são conhecidas alternativamente como proteases ou proteinases. Porém, como desde 1984, a União Internacional de Bioquímica e Biologia Molecular (UIBBM)[Webb et al. (1992)] recomenda o uso de *peptidase*, este termo foi adotado nesta tese. Enzima proteolítica e peptidase também eram tratadas como sinônimos e todas as enzimas proteolíticas eram consideradas como hidrolases. Porém, em 2004 foi descoberto novo grupo de enzimas que clivam ligações peptídicas sem hidrólise. São proteínas que realizam autoclivagem em resíduos de asparagina, como a proteína precursora *Tsh* na *Escherichia coli*. Essas enzimas estão no grupo das Liases [Rawlings et al. (2011)]. Portanto, nem toda enzima proteolítica é uma peptidase.

Nas células humanas, as peptidases correspondem a mais de 500 proteínas diferentes, conforme observado com o sequenciamento do genoma humano [Puate et al. (2005)]. Essas enzimas desempenham papéis importantes em diversos processos fisiológicos tais como: a digestão de proteínas dos alimentos, a degradação de proteínas pelos lisossomos, as cascatas de coagulação, complemento e as de sinalização intracelular, o desenvolvimento embrionário. Também estão envolvidas em várias patologias, como doenças inflamatórias, doenças hepáticas, doenças reumáticas, distúrbios neurológicos, hipertensão, doenças degenerativas, autoimunes, apoptose, cânceres, entre outras.

Em doenças causadas por vírus, como a AIDS, a hepatite C e a herpes, peptidases estão envolvidas nos processos de replicação desses microorganismos ou são usadas como fator de virulência no ciclo de infecção [Mitsuya et al. (1990)]. Esses fatos têm contribuído para que as peptidases sejam consideradas um alvo terapêutico valioso para o desenvolvimento de novos compostos farmacêuticos.

Em vegetais, as peptidases estão envolvidas nos processos de amadurecimento,

germinação, de diferenciação e morfogênese, de morte celular, de resposta de defesa de plantas a processos de estresse oxidativo ou às variações das condições ambientais [Lopez-Otin and Bond (2008)].

Além da sua importância fisiológica e farmacêutica, as peptidases têm importante aplicabilidade biotecnológica. As peptidases de origem microbianas, como as secretadas por fungos e leveduras para o meio externo, participam da degradação de proteínas cujos produtos da hidrólise servem como fonte de carbono e de nitrogênio no metabolismo celular. Na indústria, as peptidases têm sido empregadas na produção de pães e biscoitos mediante o enriquecimento proteico de farinhas de trigo, no preparo de detergentes, no processamento de couro, no amaciamento de carnes, dentre outros [Vermelho et al. (2008), Karbalaei-Heidari et al. (2009)]. Enzimas que podem ser extraídas em grandes quantidades têm significativa importância econômica, como aquelas envolvidas no amadurecimento de frutos: a ficina (figo), a papaína (mamão) e a bromelina (abacaxi).

A ação das peptidases na quebra das ligações peptídicas pode ocorrer em apenas uma ou em um número limitado de ligações de uma determinada proteína (proteólise limitada), levando à ativação ou a formação de uma proteína inativa, ou pode ocorrer em todas as ligações peptídicas, de modo que a proteína clivada é degradada em seus aminoácidos constituintes (proteólise ilimitada) [Neurath (1989)].

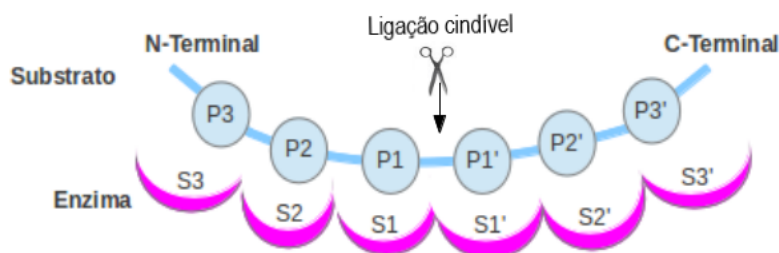
Em um complexo formado por uma peptidase e um substrato polipeptídico, os resíduos desse substrato se ligam em subsítios catalíticos da peptidase, os quais estão comumente localizados em bolsões ou sulcos na superfície da enzima.

A **especificidade de** uma peptidase para clivar uma ligação peptídica de resíduos particulares, cuja cadeia lateral está acomodada em cada subsítio específico é descrita numa terminologia cuja base é o trabalho de [Schechter et al. (1967)] que descreveram a especificidade da papaína, usando como substratos polialaninas de tamanhos variados (ala2 até ala6).

Na peptidase, esses subsítios são chamados S (**S**ubsítios) e os resíduos de aminoácidos do substrato são denominados P (**P**eptídeos), conforme pode ser visto na Figura 3.1. Os subsítios catalíticos da enzima são numerados como S_n, \dots, S_2, S_1 para o lado N-terminal e S_1', S_2', \dots, S_n' para o C-terminal. Os resíduos de aminoácidos do lado N-terminal da ligação peptídica do substrato que é hidrolisada pela peptidase são numerados P_n, \dots, P_2, P_1 e os resíduos do lado C-terminal são numerados P_1', P_2', \dots, P_n' . Os resíduos $P_1 - P_1'$ formam uma ligação chamada *scissile bond* ou ligação peptídica do substrato que é clivada pela peptidase na hidrólise.

Importante notar que na nomenclatura de Schechter & Berger, cada resíduo ficava associado a um único subsítio, como se houvesse um sub-bolsão na enzima para cada um deles. Percebam também que os substratos utilizados foram polialaninas de 2 a 6 resíduos de comprimento, ligantes tipicamente hidrofóbicos. Ressaltaremos esses detalhes em nossa discussão dos resultados mais adiante.

Figura 3.1: Nomenclatura dos subsítios de uma peptidase e os resíduos complementares de seu substrato - nomenclatura de Schechter e Berger.



Fonte: Adaptada de <https://swift.cmbi.umcn.nl/teach/B2/LINK/NOOT_32.html>. [Schechter et al. (1967)].

O sítio ativo da enzima é composto de $S3 - S1$ e $S1' - S3'$ localizados em ambos os lados (N e C terminal) do sítio catalítico da enzima. As posições P no substrato são enumeradas a partir do ponto de clivagem, portanto têm a mesma numeração dos subsítios que ocupam na enzima.

3.1.1 Classificação das Peptidases

De acordo com o sistema de classificação de enzimas, chamado EC (*number Enzyme Classification*), proposto pelo Comitê de Nomenclatura Enzimática da IUBMB (*The International Union of Biochemistry and Molecular Biology*), as enzimas são classificadas em seis classes: (1) Oxidorredutases, (2) Transferases, (3) Hidrolases, (4) Liases, (5) Isomerases e (6) Ligases. As peptidases pertencem à classe 3 (hidrolases) e subclasse 3.4 (peptídeo hidrolases ou peptidases) [Voet and Voet (2013)].

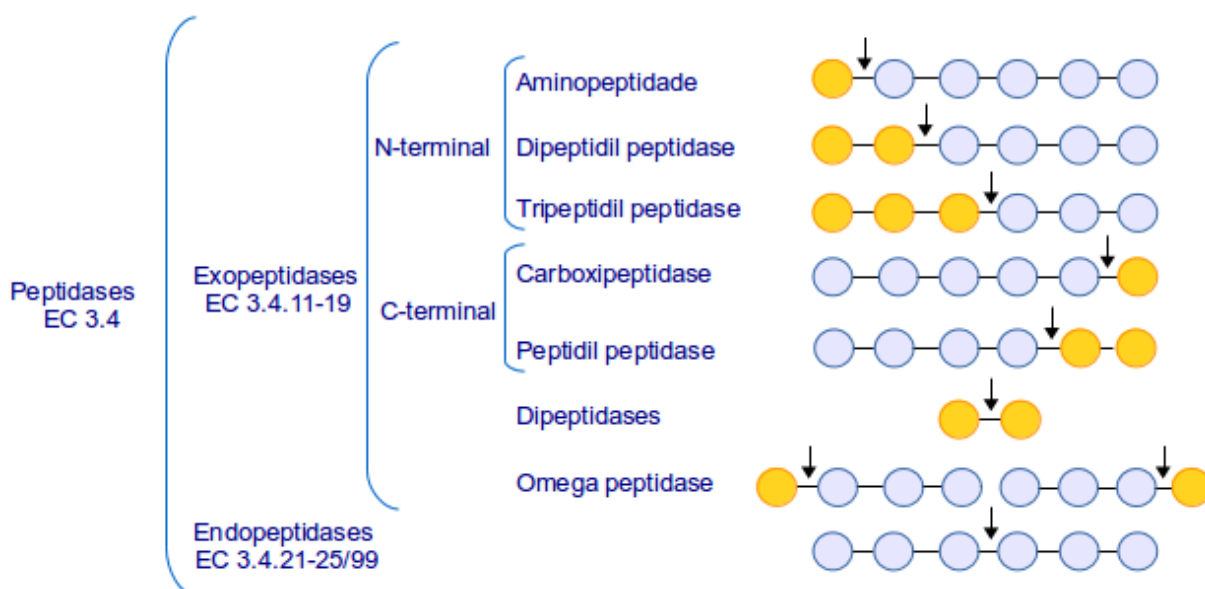
De acordo com [Merops (2015)], há três métodos de classificação das peptidases definidos, como segue:

- pelo tipo de reação química da catálise,
- pelo mecanismo de ação da catálise, e
- pela estrutura molecular e homologia.

3.1.1.1 Peptidases agrupadas pelo tipo de reação química

As peptidases catalisam uma mesma reação: pela hidrólise de uma ligação peptídica, mas apresentam seletividade para ligações peptídicas em posições particulares na cadeia polipeptídica do substrato, ou seja, se a ligação peptídica está na extremidade ou no interior da cadeia [Salleh et al. (2006)]. Com base nesse critério, as peptidases podem ser classificadas em *exopeptidases* (EC 3.4.11-19) e *endopeptidases* (EC 3.4.21-25/99) como é apresentado na Figura 3.2.

Figura 3.2: Tipos de peptidases e representação do modo de ação. Os círculos azuis representam os resíduos de aminoácidos na cadeia polipeptídica. Os círculos amarelos representam os resíduos terminais e as setas indicam o local da clivagem.



Fonte: Adaptado de [Rao et al. (1998)].

As exopeptidases são enzimas que atuam somente nos finais das cadeias polipeptídicas na região *N* ou *C* terminal, enquanto as endopeptidases, também conhecidas como proteinases, agem preferencialmente nas regiões internas da cadeia polipeptídica, entre as regiões *N* e *C* terminal.

As exopeptidases que atuam na região *N*-terminal com liberação de um único resíduo de aminoácido são chamadas de aminopeptidases, quando liberam dois resíduos (dipeptídeo) ou três resíduos (tripeptídeo) são chamadas, respectivamente, de dipeptidil-peptidases e tripeptidil-peptidases. As exopeptidases que atuam na região *C*-terminal livre liberando um único resíduo de aminoácido são denominadas carboxipeptidases, quando liberam dois resíduos de aminoácidos são denominadas peptidil-dipeptidase. Algumas exopeptidases hidrolisam dipeptídeos (dipeptidase). Existem ainda exopeptidases

que removem resíduos de aminoácidos substituídos, cíclicos ou em ligação isopeptídica, neste último caso são denominadas omega peptidases [Merops (2015)].

3.1.1.2 Peptidases agrupadas pelo mecanismo de ação da catálise

As peptidases também podem ser descritas de acordo com seus mecanismos e tipos catalíticos em: aspártico, cisteíno, treonino, glutâmico, metalo ou serino peptidases. Além destas classes mecanísticas, existe ainda uma classe destinada às enzimas cujo mecanismo catalítico ainda não é conhecido [Rawlings and Salvesen (2012)]. Falaremos em mais detalhes sobre alguns desses mecanismos mais adiante.

Um tipo catalítico está relacionado aos grupos químicos responsáveis pela catálise da ligação peptídica na hidrólise. Na serino peptidase, o nucleófilo catalítico é a hidroxila da cadeia lateral de uma serina no sítio ativo e na cisteíno peptidase, o nucleófilo catalítico é o o grupo tiol da cadeia lateral do sítio ativo da cisteína. Nas aspártico peptidases, o nucleófilo são moléculas de água em interação com as cadeias laterais de aspárticos. Nas metalopeptidases, um ou dois íons metálicos mantêm a água também como nucleófilo. O metal pode ser o zinco, o cobalto, o manganês, o níquel ou ferro. Nas treonino peptidase, o nucleófilo é a hidroxila da treonina. Ainda se tem pouco conhecimento sobre as glutâmico peptidases, cujos mecanismos catalíticos parecem ser realizados por uma díade catalítica *Glu/Gln* [Rawlings and Salvesen (2012)]. Na seção 3.2 deste capítulo, será apresentada uma breve descrição sobre as serino peptidases pois os complexos enzima/inibidor selecionados neste trabalho são desse tipo.

Essa classificação baseada no tipo catalítico é útil porque, de modo geral, enzimas do mesmo tipo tendem a ser inibidas pelos mesmos inibidores. Entretanto, podem ter pouca ou nenhuma relação evolutiva. Por exemplo, tripsina e subtilisina são serino peptidases que não têm mesma origem evolutiva. Por outro lado, peptidases de tipos catalíticos diferentes podem estar relacionadas evolucionariamente, como a peptidase cisteínica *picornain 3C poliovirus* e a tripsina [Rawlings et al. (2011)].

3.1.1.3 Peptidases agrupadas pela estrutura e homologia

A classificação de peptidases tendo como critérios as características estruturais da molécula e evolucionárias (homologia) foi proposta por [Rawlings and Barrett (1993)] em 1993 e utilizada no banco de dados MEROPS. Este é um banco de dados curado manualmente, composto de peptidases e seus inibidores proteico organizados em três níveis: peptidases, famílias e clãs [Rawlings et al. (2004), Rawlings et al. (2014a), Rawlings et al. (2014b), Rawlings and Salvesen (2012)].

As peptidases são diferenciadas entre si pelas atividades que desempenham. É uma “classificação simples” com critério similar ao utilizado na lista do IUBMB EC. As peptidases quando têm alguma similaridade sequencial significativa são agrupadas em famílias, as quais por similaridade estrutural são agrupadas em clãs.

Uma família é um conjunto de enzimas proteolíticas homólogas. Cada família tem uma peptidase “exemplo” com a qual um membro dessa família deve ter significativa similaridade estatística quando as respectivas sequências de aminoácidos são comparadas, ou quando essa comparação é feita com qualquer membro da família. Esse relacionamento deve existir em pelo menos uma parte (unidade) da peptidase responsável por sua atividade e se deve garantir que haja algum relacionamento evolucionário entre os membros. Esse exemplar, ou representante característico da família é o que dá nome a ela. Cada família é identificada por uma letra seguida por um número arbitrário, por exemplo, S1. A letra representa o tipo catalítico da enzima: *Aspartic(A)*, *Cysteine(C)*, *Glutamic(G)*, *Metallo(M)*, *Asparagine(N)*, *Mixed(P)*, *Serine(S)*, *Threonine(T)*, *Unknown(U)*. Algumas famílias são divididas em subfamílias porque há evidência de uma divergência muito antiga dentro da família, por exemplo, S1A, S1B.

Um clã contém todas as peptidases que surgiram a partir de uma única origem evolutiva de peptidases. Ele representa uma ou mais famílias cuja evidência de relação evolutiva se dá por meio da similaridade entre as suas estruturas tridimensionais. Quando essas estruturas não estão disponíveis, a relação evolutiva é dada pela ordem dos resíduos do sítio catalítico na cadeia polipeptídica e frequentemente por motivos comuns na sequência em torno dos resíduos catalíticos. Cada clã é identificado com duas letras. Do mesmo modo que nas famílias, a primeira letra representa o tipo catalítico das famílias incluídas no clã: *A*, *C*, *G*, *M*, *N*, *P*, *S*, *T*, *U*. A letra *P* é utilizada para um clã que contém famílias com peptidases de mais de um tipo catalítico (*S*, *T*, *C*). A segunda letra é atribuída arbitrariamente. Há algumas famílias que ainda não pode ser atribuídas a clãs. Neste caso, quando uma atribuição formal é necessária, a família é descrita como pertencente ao clã *A-*, *C-*, *M-*, *S-*, *T-* ou *U-*, de acordo com o tipo catalítico. Alguns clãs são divididos em subclãs porque há evidência de uma divergência muito antiga dentro do clã. No geral envolvem mudanças do tipo catalítico, mantendo o mesmo motivo estrutural,

como acontece nos subclãs PA(S) e PA(C), o primeiro com nucleófilo serina, o segundo com cisteína.

Os inibidores proteicos de peptidases são classificados no MEROPS da mesma forma que as peptidases. No entanto, cada família tem a primeira letra do nome identificada pela letra *I*. O identificar do clã, formado por duas letras, inicia-se com as letras *I* ou *J*.

3.2 Serino Peptidases

As serino peptidases formam uma ampla classe de enzimas proteolíticas que foram extensivamente estudadas, sendo encontradas em procariotos e eucariotos. Essas enzimas receberam esse nome por terem um mecanismo catalítico comum caracterizado pela existência de um resíduo de serina muito reativo (nucleófilo), essencial para a atividade enzimática [Voet and Voet (2013)].

No MEROPS, as serino peptidases estão agrupadas em cerca de 50 famílias distinguidas em função da similaridade sequencial dos aminoácidos e distribuídas em 14 clãs definidos com base na estrutura terciária e ordem dos resíduos catalíticos nas sequências [Barrett et al. (2012)]. Desses clãs, aqueles que foram objeto de pesquisa deste trabalho são: o clã PA (subclã PA(S)) com a família S1 e o clã SB com as famílias S8 e S53.

O clã PA agrupa famílias que possuem origem comum com a quimotripsina e contém serino e cisteíno peptidases, distribuídas nos subclãs PA(S) e PA(C), respectivamente. A similaridade entre as famílias do clã PA é baseada nos *folds* da proteína ou no arranjo dos resíduos catalíticos.

No subclã PA(S), todas as enzimas são endopeptidases, com exceção da dipeptidyl-peptidase 7 do *Porphyromonas gingivalis*. A tríade catalítica típica é formada por HIS, ASP, SER, onde o ASP ajuda a orientar o anel imidazol da histidina (que irá agir tanto como um ácido quanto uma base conjugada) e as estruturas terciárias consistem principalmente de folhas β organizadas em barris.

A família S1 é a maior família das peptidases tanto em termos de número de proteínas sequenciadas quanto em termos do número de peptidases com atividades reconhecidamente diferentes e é representada pela *quimotripsina A* do *Bos taurus*. Todos os membros dessa família têm a tríade disposta na seguinte ordem: HIS, ASP e SER.

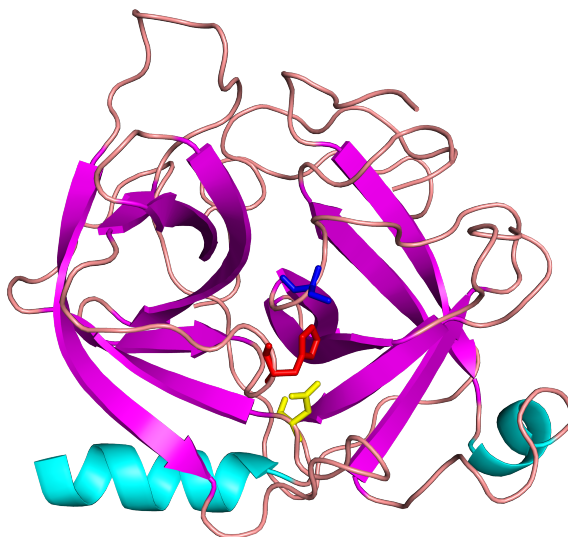
O clã SB contém a família S8 (família subtilisina) e a família S53 (família sedolisina). Em ambas as famílias, a serina catalítica está em um motivo GLY-THR-SER-XAA-XAA-XBB-PRO, onde XAA é um aminoácido alifático e XBB é um pequeno aminoácido (SER/THR/ALA/GLY) [Rawlings and Salvesen (2012)]. Somente a SER em ambas as

famílias, a HIS na família S8 e a GLU na família S53 têm posições equivalentes, ou seja, em S53 a GLU substitui a HIS [Page and Di Cera (2008)]. Na família S53, os resíduos catalíticos GLU e ASP ocorrem no motivo GLU-XAA-XAA-LEU-ASP. Uma asparagina tem papel importante como oxiânio na família S8 e o aspartato na família S53.

3.2.1 Enzimas Tipo Tripsinas

Segundo o Merops, as enzimas Tipo Tripsinas estão agrupadas no subclã PA(S), família S1, cuja enzima representativa é a quimotripsina *A*. Essa enzima é caracterizada por ter uma estrutura terciária constituída de dois domínios, onde cada domínio contém um β barril, e entre os quais situa-se o sítio ativo (Figura 3.3). Embora a quimotripsina seja descrita como uma proteína “*all – beta*” (SCOP-família b.47.1.2, [Fox et al. (2014)]), há uma α -hélice que interage com ambos os domínios e presumidamente estabiliza a sua interação. No barril N-terminal estão os resíduos HIS e ASP do sítio ativo e a SER (nucleofílica) está no barril C-terminal. Essa família inclui enzimas como a quimotripsina, tripsina, elastase, trombina, enzimas envolvidas na coagulação sanguínea, *streptogrisin A*, dentre outras.

Figura 3.3: Estrutura tridimensional da Quimotripsina (ID PDB: 1ACB). Tríade catalítica: ASP102 (amarelo), HIS57 (vermelho) e SER195 (azul).



Em relação à especificidade para o substrato, as enzimas da família S1 têm diferenças marcantes. A tripsina é específica para clivar ligações peptídicas contendo resíduos com cadeia lateral carregada positivamente, em condições fisiológicas, tais como a argi-

nina (ARG) e a lisina (LYS), do lado C-terminal. A tripsina é formada da clivagem do tripsinogênio na porção N-terminal que permite o enovelamento para a configuração ativa [Huber and Bode (1978)]. A quimotripsina é específica para ligações peptídicas contendo resíduos de aminoácidos com cadeias laterais hidrofóbicas também no lado C-terminal, como a fenilalanina (PHE), tirosina (TYR) e triptofano (TRP). A maioria das peptidases do clã PA, do tipo tripsina, tem seletividade por cadeias laterais da Arg. No entanto, muitos membros de origem viral ou bacteriana na família são específicos para a GLN. A elastase tem especificidade para a ALA, GLU ou VAL.

As peptidases do clã PA, família S1, são encontradas em plantas, protozoários, fungos, bactérias, animais e vírus. Os inibidores naturais da família S1 incluem inibidores das famílias *ovomucoid* (I1), *aprotinin* (I2) e *Bowman-Birk* (I12). As serpinas também inibem algumas peptidases na família S1.

3.2.2 Enzimas Tipo Subtilisinas

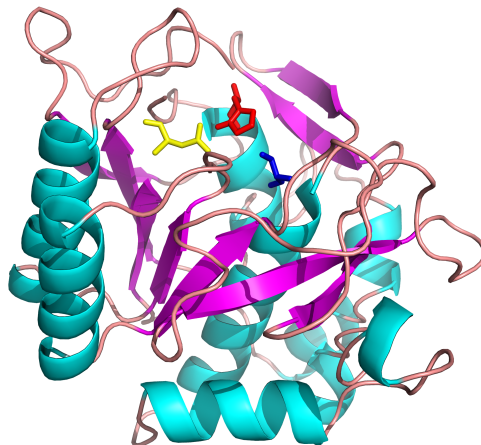
As enzimas tipo subtilisinas estão agrupadas no clã SB do Merops e são representadas pela subtilisina Carlsberg (*Bacillus licheniformis*).

A subtilisina difere bastante da quimotripsina em termos de sequência e estrutura. Enquanto a quimotripsina tem dois domínios, conforme já mencionado, contendo folhas β , a subtilisina tem somente um domínio contendo elementos α e β , sendo classificada como α/β no SCOP (família c.41.1.1, [Fox et al. (2014)]). As tipo subtilisinas, na perspectiva tridimensional, têm três camadas com sete folhas β sobrepostas entre duas camadas de hélices, como pode ser observado na Figura 3.4.

As peptidases do clã SB estão presentes em plantas, protozoários, fungos, bactérias, vírus e com poucos representantes no genoma animal.

Os inibidores proteicos da família S8 incluem o Ovomucoide (OMTK3) da família I1, *Streptomyces subtilisin inhibitor* da família I16, membros da família I13 tais como a eglina C e o inibidor da cevada CI-1A (I13.005), muitos dos quais também inibem a quimotripsina. Na subtilisina propetidase (POA) há autoinibição.

Figura 3.4: Estrutura tridimensional da subtilisina (ID PDB: 1GCI). Tríade catalítica: ASP32 (amarelo), HIS64 (vermelho) e SER221 (azul).



3.2.3 Mecanismo catalítico clássico

A homologia entre os sítios das várias serino peptidases indica que possuem o mesmo mecanismo catalítico. Por exemplo, a despeito das diferenças de ordem na sequência polipeptídica dos resíduos da tríade catalítica nos clãs PA(S) e SB, sendo respectivamente, His/Asp/Ser e Asp/His/Ser e de diferenças na estrutura primária e terciária (veja Figuras 3.5 e 3.6), o mecanismo de ação para as enzimas desses clãs é o mesmo, indicando a convergência evolutiva.

Com base em dados estruturais e químicos obtidos em vários laboratórios foi formulado o mecanismo catalítico mostrado em termos da quimotripsina na Figura 3.7.

A hidrólise de uma ligação peptídica pela quimotripsina tem duas fases. Na fase da acilação, a formação de um intermediário covalente acil-enzima é acoplada à fase da ligação peptídica: a enzima liga-se ao substrato formando um complexo e em seguida ocorre o ataque da Ser reativa à carbonila da ligação peptídica do substrato para formar o intermediário acil-enzima. A formação desse intermediário é facilitada pela presença da His e do Asp da tríade catalítica. A histidina serve para posicionar a cadeia lateral da serina e para polarizar a sua hidroxila. Ao fazer isto, age como catalisador básico geral, como acceptor de íon de hidrogênio, porque a hidroxila polarizada da serina fica pronta para desprotonação. O aspartato ajuda a orientar a histidina e torna-la um melhor acceptor de prótons, por efeitos eletrostáticos e/ou ponte de hidrogênio. Outro aspecto importante, na formação desse intermediário tetraédrico, é a existência de um bolsão da enzima denominado sítio do oxianion (*oxyanion-binding site*) constituído de dois grupos

Figura 3.5: Alinhamento das sequências do tipo Tripsina (complexo 1PPF contendo enzima *Human Leukocyte Elastase de Homo sapiens*) e do tipo Subtilisina (complexo 1SBN contendo enzima *Subtilisin NOVO BPN*). Score de identidade de 19,61% - ClustalO.

```

CLUSTAL O(1.2.1) multiple sequence alignment

1PPF:E  -----IVGRRARPHAWPFMV5
1SBN:E  AQSVPYGV5QIKAPALHSQYTG5NVKVAVIDSGID55HPDLKVAGGAS5MVP5ETNPFQD
          :. **      *      :.

1PPF:E  LQLRGGHFCGATLIAPNFVMSAAHCVANVNVRAVRVVLGAHNLSRREPTRQVFAVQRIFE
1SBN:E  NNSHGTHVAGTVAA-----LNN5IG---VLGVAPSASLYAVKVLGA
          : * * . * :.          :. : * : : * : : : * : :

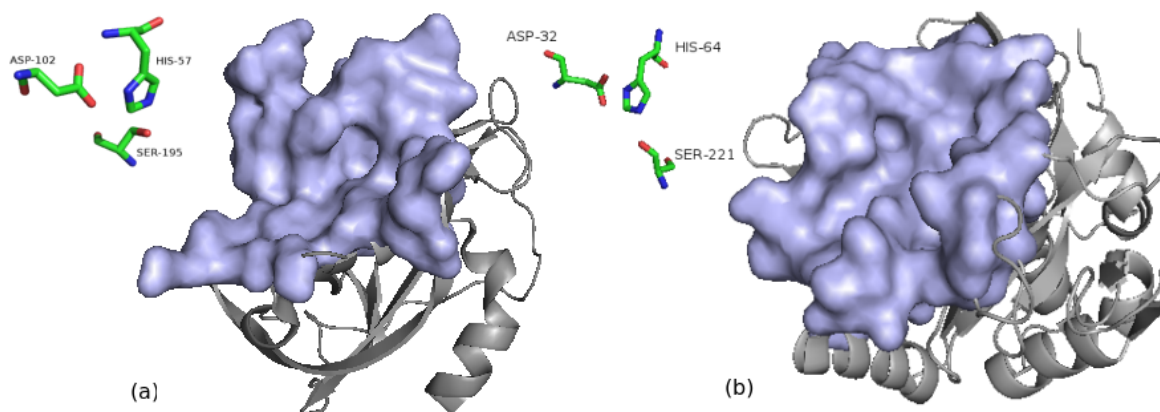
1PPF:E  NGYDPVNLNDIVILQLNGSATINANVQVAQLPAQGRRLGNGVQCLAMGWLLGRNRGIA
1SBN:E  DG5G-----Q5SWIINGI5EWAIA
          : *                          : * : * : *

1PPF:E  SVLQELNVTVVTSLC-----RR5NVCTLV-RGR-----
1SBN:E  NNMDVINMSLGGPSGSAALKA5VDKAVASGVVVVAAAGNEGTSG5S5TVGYPGKYPSVIA
          : : : * : : :          * * . . . *

1PPF:E  -----QAGVCFGDSG5SPLVCNGLIHGIA5FVRGGCASGLYP-----DAFAPV
1SBN:E  VGAVD55NQRA5F5SVGPELDVMAPGV5IQ5TLPG-NKYGAYNGT5MASPHVAGAAALIL
          : . . * . . * * . . * * : * * : * * : * * :

1PPF:E  AQFVNWIDSIIQ-----
1SBN:E  SKHPNWTNTQVR5SLENTTTLKLD5SFY5GKGLINVA5AAQ
          : : . * * : : :
  
```

Figura 3.6: Semelhança topológica dos resíduos da tríade catalítica de enzimas tipo Tripsina e tipo Subtilisina. Em (a), uma enzima do tipo Tripsina e sua tríade (complexo 1PPF) e em (b) do tipo Subtilisina e sua tríade (complexo 1SBN).



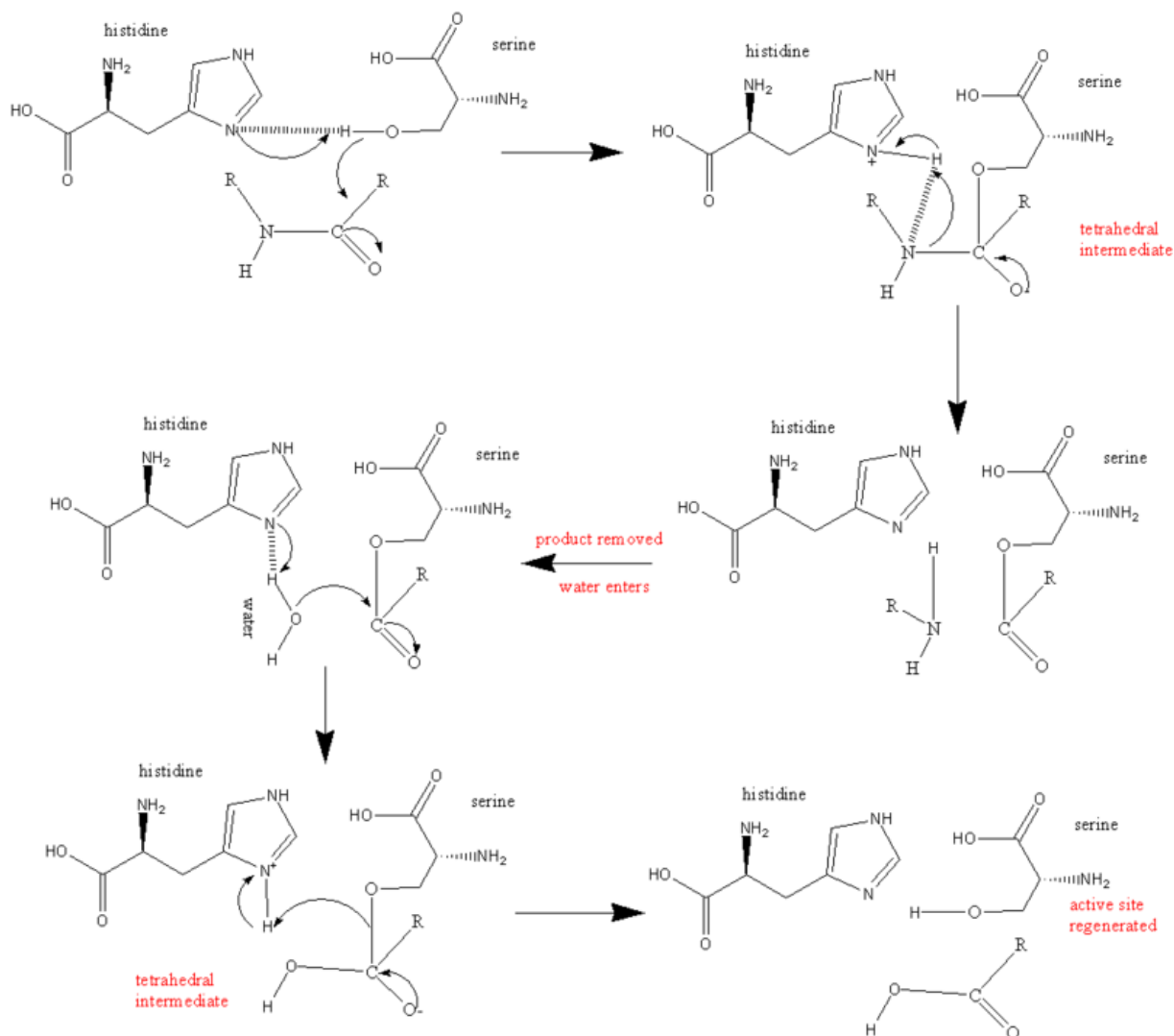
Fonte: Adaptado de [Gonçalves-Almeida (2011)].

NH do backbone da cadeia polipeptídica associados com a Gly193 e Ser195. O oxigênio da Ser195, ao perder seu próton para a His, ataca o carbono da carbonila, transformando-o num intermediário tetraédrico ligado a um oxigênio carregado negativamente (o oxiânio).

Na fase de desacilação, regenera-se a enzima livre. Isso é essencialmente o reverso da fase de acilação, com uma molécula de água como nucleófilo no lugar da Ser [Lehninger et al. (2006)].

Na subtilisina do exemplo acima (Figura 3.6, complexo 1SBN), a tríade catalítica é constituída por Asp32, His64, Ser221, nessa ordem e o sítio do oxiânion é constituído

Figura 3.7: Mecanismo catalítico das serino peptidases.



Fonte: Snellios at the English-language Wikipedia, 13/12/2006. Licenciado sob CC BY-SA 3.0, via *Wikimedia Commons*. Disponível em: <https://commons.wikimedia.org/w/index.php?curid=5327342>. Acesso em Outubro 2015 .

pelo grupo NH da Ser221 do backbone e pela cadeia lateral do grupo amida de Asn155.

3.3 Inibidores de Serino Peptidases

As peptidases mesmo realizando funções essenciais à manutenção da vida, também podem ser perigosas e por isso devem ser controladas. No caso das enzimas digestivas, por exemplo, se essas não fossem sintetizadas na forma de precursores inativos, elas poderiam digerir os tecidos onde são sintetizadas. A pancreatite aguda pode ser precipitada por

trauma no pâncreas e é caracterizada pela ativação prematura das enzimas digestivas sintetizadas nessa glândula [Voet and Voet (2013)]. Entre os mecanismos de controle das peptidases, dois prevalecem: a biossíntese de precursores inativos chamados zimogênios e a ação de inibidores proteicos na forma ativa das enzimas [Laskowski Jr and Kato (1980)].

Os zimogênios são armazenados e ativados sob demanda, como por exemplo, o tripsinogênio precursor da tripsina. A ativação envolve proteólise de uma ou mais ligações peptídicas da porção N-terminal em direção ao sítio catalítico da enzima. Essa forma localizada da ativação garante que a enzima ativa não seja formada antes de passar pelo mesmo processo de ativação proteolítica comum à biossíntese de proteínas, que ocorre a partir do N-terminal para o C-terminal.

Uma vez que as enzimas estão ativadas, outro mecanismo é necessário para controle das mesmas. Os inibidores proteicos de peptidases cumprem esse papel formando complexos completamente inativos (inibição total) ou parcialmente ativos com suas enzimas cognatas (inibição parcial). Esses inibidores apresentam estruturas distintas e podem inibir suas enzimas alvo tanto por inibição reversível, que é o mecanismo mais comum observado, quanto por inibição irreversível. Segundo [Krowarsch et al. (2003)], os inibidores proteicos podem ser classificados em canônicos (mecanismo padrão), não-canônicos e serpinas.

3.3.1 Inibidores canônicos

Nessa classe de inibidores de serino peptidases, divididos em várias famílias de pequenas proteínas (14 a 200 resíduos de aminoácidos), nem todos os inibidores são homólogos, mas a grande maioria interage com suas enzimas cognatas seguindo um mecanismo padrão. As enzimas do tipo subtilisina e do tipo tripsina, por exemplo, as quais não compartilham uma estrutura tridimensional comum, hidrolisam seus substratos e são inibidas pelos seus inibidores pelo mesmo mecanismo [Laskowski Jr and Kato (1980)].

Os inibidores canônicos são assim chamados porque se ligam às suas enzimas cognatas do mesmo modo do substrato, mas são clivados de forma muito mais lenta. Apresentam como característica cinética marcante uma ligação forte, não covalente, com o sítio ativo da enzima à semelhança do complexo na interação enzima-substrato. O segmento do inibidor responsável pela inibição, chamado de alça de ligação à peptidase (*RCL-reactive centre loop*), tem uma conformação similar, canônica em todos os inibidores de estruturas conhecidas [Bode and Huber (1992)].

O exato mecanismo por trás da inibição canônica em serino peptidases ainda não é todo conhecido [Radisky and Koshland (2002)]. Vários estudos mostram que eles se

ligam de maneira semelhante aos substratos [Laskowski Jr and Kato (1980), Radisky and Koshland (2002)]. Na verdade, eles tendem a se complexar de forma muito mais intensa que substratos normais (constante de associação da ordem de $10^{14}M^{-1}$ contra $10^{-1}M^{-1}$ para alguns substratos, como β -Lactoglobulina por tripsina [Olsen et al. (2000)], embora o passo seguinte (a taxa de hidrólise) seja retardada por um fator entre 10^6 a 10^{10} .

Postula-se que entre as razões desse “estranho” comportamento estariam: (i) a extrema rigidez dos complexos, de forma a prevenir por imobilidade ataques nucleófilos produtivos; (ii) a orientação inadequada dos grupos reagentes, como os nucleófilos, as bases gerais no lado enzima, o carbono da carbonila no lado inibidor, bem como os átomos que participam na estabilização no estado intermediário; (iii) o posicionamento do grupo abandonador ($H_2N - R_2$) no complexo acil-enzima favorecendo a reação reversa.

Na literatura, encontra-se evidências cada vez maiores de que o postulado (ii) não procede para a maioria dos casos [Radisky and Koshland (2002)]. Já o postulado (iii) vem ganhando cada vez mais força. Radisky & Koshland [Radisky and Koshland (2002)], com base em dados experimentais e simulacionais envolvendo a inibição da subtilisina BPN com quimotripsina (CI2), encontraram uma vasta rede de pontes de hidrogênio no complexo (algo não evidente em substratos normais) que poderiam estabilizar o grupo abandonador, literalmente “segurando-o” e evitando o seu afastamento do complexo, o que favoreceria a reação reversa de desacilação. Logo, sobre o postulado (i), percebe-se que não estaria de todo incorreto: há um nível de interação maior entre inibidores e enzimas que substrato e enzimas, tornando seus complexos mais “próximos” e justapostos, mas isso não implica necessariamente em maior “rigidez”. Há espaço para uma certa dinâmica, em que o grupo abandonador é até formado, mas ele tem dificuldade de escapar do complexo em função das interações. E como não escapa, favorece a reação reversa, mantendo o inibidor em complexo com a enzima.

As famílias de inibidores canônicos de serino peptidases com estruturas resolvidas têm representantes de origem animal, vegetal e em micro-organismos, como pode ser visto na Tabela 3.1.

3.3.2 Inibidores não canônicos

Os inibidores não canônicos por meio do segmento N-terminal se ligam ao sítio ativo da peptidase formando uma pequena folha β paralela, exibindo estrita especificidade para uma determinada peptidase. Essa ligação do inibidor com a enzima assemelha-se ao complexo enzima-substrato, como é observado para os inibidores canônicos. Inibidores não canônicos também mantêm interações secundárias com a peptidase fora do sítio

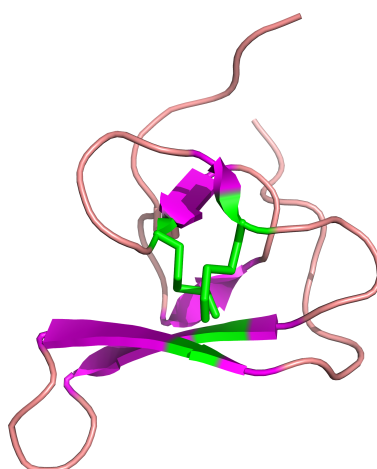
Tabela 3.1: Famílias de inibidores de serino peptidases com estrutura cristalina resolvida. Inibidores de origem ^aanimal, ^vvegetal e de ^mmicroorganismos.

Família	PDB ID representativo (Å)	
	Livre	Complexo
BPTI Kunitz ^a	1BPTI	2PTC
STI Kunitz ^v	1AVU	1AVW
BBI ^v	1PI2	1TAW
Potato I ^v	1MIT	1ACB
Potato II ^v	1TIH	4SGB
Squash ^v	2CTI	1PPE
Kazal ^v	2OVO	1PPF
SSI ^m	2SSI	1SIC
Antistasina ^a	1SKZ	1HIA
Chelonianina ^a	2REL	1FLE
Ascaris ^a	1ATA	1EAI
Ecotina ^m	1ECY	1EYS

ativo. Com isso uma maior área fica enterrada e contribui significativamente para a força, velocidade e especificidade do reconhecimento molecular, aumentando a seletividade da inibição [Farady et al. (2008)].

O exemplo clássico desse tipo de inibidor é a hirudina (antitrombina) que inibe a trombina. A hirudina é purificada da saliva da sanguessuga *Hirudo medicinalis* e pode ser usada na fabricação de fármacos capazes de dissolver coágulos sanguíneos ou como template para a criação de inibidores sintéticos usados na profilaxia e tratamento de trombose e distúrbios correlatos. A família das hirudinas é composta por inibidores homólogos de cerca de 7kDa com uma cadeia polipeptídica estabilizada por três pontes dissulfeto altamente conservadas (Figura 3.8).

Figura 3.8: Estrutura tridimensional da hirudina extraída da saliva do *Hirudo medicinalis* (PDB 2HIR). Pontes dissulfeto representadas em verde.



Inibidores do tipo Kazal e inibidores tipo canônicos também apresentam interações secundárias como os não canônicos, mas com uma conformação distorcida do loop de li-

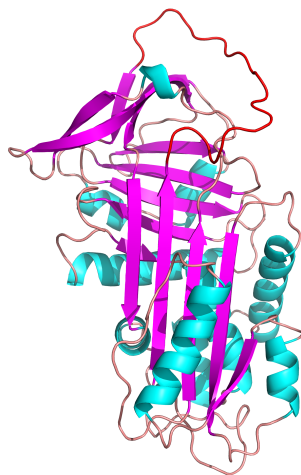
gação. Os inibidores não canônicos são menos abundantes que os canônicos e as serpinas, ocorrendo predominantemente em organismos sugadores de sangue, como o sanguessuga. Inibem peptidases envolvidas na cascata de coagulação como a trombina ou fator Xa, além de estarem envolvidos na regulação de mecanismos como apoptose, sinalização intracelular, embriogênese, angiogênese, neurogênese e em eventos relacionados à resposta imune [Woods et al. (2008)]. Até o momento, poucos inibidores não canônicos foram caracterizados em termos de estrutura e cinética da interação com as proteínas alvo.

3.3.3 Serpinas

As Serpinas (*SER*rine *PRO*tease *INH*ibitors) são glicoproteínas que constituem uma superfamília de inibidores de serino peptidases com estruturas muito semelhantes. Têm relevância associada a várias doenças humanas. Deficiências na capacidade inibitória da α -1-antitripsina, originárias de alterações na estrutura protéica causadas por mutações em seu gene, podem comprometer a sua ação sobre a elastase neutrofílica que tem a capacidade de hidrolisar as fibras de elastina no pulmão, refletindo-se no pulmão sob a forma de enfisema [Abboud et al. (2005)].

São proteínas cuja estrutura terciária é caracterizada por um número variado de folhas β , oito ou nove α -hélices, e um loop central reativo (RCL) composto de aproximadamente 20 resíduos próximos à região C-terminal (Figura 3.9). Na região RCL está o centro de clivagem entre P1 e P1'.

Figura 3.9: Estrutura tridimensional da Serpina α ₁-antitripsina humana (PDB 1PSI). Os inibidores do tipo Serpina são formado por folhas β (em magenta), oito ou nove α -hélices (cyan) e um *loop* na porção superior que é a RCL (vermelho).



Além de inibir irreversivelmente serino peptidases como quimotripsina, tripsina,

trombina, elastase, fatores da coagulação sanguínea, as Serpinas inibem também algumas cisteíno-peptidases [Rawlings et al. (2004)].

Capítulo 4

Hidrofobicidade e formação de complexos proteína-proteína

Este capítulo pretende aprofundar o conceito de hidrofobicidade e a sua relação com a formação de complexos proteína-proteína.

4.1 Hidrofobicidade

O efeito hidrofóbico - tendência da porção apolar do soluto se segregar da parte polar em soluções polares como a água, formando aglomerados - é um fenômeno importante nas interações proteína-proteína. Para mapear o padrão de hidrofobicidade/hidrofilicidade das superfícies de proteínas, escalas da hidropatia dos aminoácidos são frequentemente usadas, por exemplo, escalas de Miyazawa, Eisenberg e Kyte-Doolittle [Campbell (2007)]. Essas escalas classificam os aminoácidos com base na energia livre de transferência e em muitos métodos essas escalas são melhoradas incluindo parâmetros de solvatação da área exposta ao solvente [Eisenberg and McLachlan (1986)] ou a partir do número de contatos por resíduo [Colonna-Cesari and Sander (1990)]. No entanto, porque essas escalas diferem entre si, para uma mesma proteína ou complexo proteico valores diferentes costumam ser obtidos na quantificação do efeito hidrofóbico. Além disso, para se determinar eficientemente a hidrofobicidade é importante considerar a dependência do contexto, ou seja, do ambiente local onde a molécula está inserida [Jamadagni et al. (2011)].

Estudos mais recentes têm mostrado que o efeito hidrofóbico depende do tamanho da partícula e do comportamento do solvente. Essa caracterização da hidrofobicidade com base em flutuações de densidade naturalmente considera como a água responde à química local e ao contexto topográfico das moléculas protéicas [Chandler (2005)].

Numa rápida revisão da termodinâmica envolvida nesses fenômenos, pode-se dizer que ela essencialmente nos permite dizer quando um processo poderá ocorrer espontaneamente num sistema em busca de equilíbrio. Um dos parâmetros que é convenientemente

usado para quantificar tal “espontaneidade” é a variação de energia livre (ΔG): fenômenos espontâneos oferecem uma variação de energia livre negativa. No contexto de uma molécula em solvatação, ΔG estará relacionado ao trabalho (teoricamente reversível) necessário para a água se reorganizar para assimilar o soluto (no nosso caso, hidrofóbico) no sistema.

A variação de energia livre, tal qual definida por Willard Gibbs (1839-1903) em um clássico trabalho de 1873 [Gibbs (1873)], tem dois principais componentes: $\Delta G = \Delta H - T\Delta S$, onde ΔH e ΔS são a variação de entalpia e variação de entropia que ocorreram durante a solvatação. A parte entálpica pode ser vista como uma medida do potencial médio de interação entre moléculas; a parte entrópica, como uma inferência da “ordem” aferida pelo grau de correlações intermoleculares. Do ponto de vista de um sistema solvente - soluto hidrofóbico, a entalpia vai estar relacionada com as interações, principalmente a quebra ou formação de pontes de hidrogênio, enquanto que a entropia com os desvios da aleatoriedade, em termos de organização espacial e padrões dessas pontes de hidrogênio induzidos pela presença do soluto hidrofóbico [Chandler (2005)].

4.1.1 Influência do tamanho das partículas hidrofóbicas

É comum para quem faz macarrão, adicionar um pouco de óleo vegetal na água em aquecimento. Como o óleo vegetal é formado por ácidos graxos com longas cadeias carbonadas hidrofóbicas, há indução de fases macroscopicamente bem distintas entre óleo e água. No início, o óleo tende a formar numerosas gotas de diferentes tamanhos, mas com o tempo, vão se coalescendo em poucas gotas maiores. Por que isso acontece?

[Chandler (2005)] analisaram a influência do tamanho de partículas esféricas (simulando um soluto hidrofóbico) em diluição num solvente aquoso. Perceberam que o comportamento termodinâmico é dependente do tamanho do soluto. Para solutos pequenos, como o metano (0.35 nm de raio), embora as águas de solvatação sejam entropicamente afetadas por interferências em sua liberdade de compor pontes de hidrogênio, o número médio dessas pontes não é muito afetado em relação ao solvente distante (em inglês, *bulk*). Isso cria correlações, fazendo com que as águas de solvatação tenham um comportamento diferente das águas do *bulk*, e permaneçam mais próximas da partícula hidrofóbica, numa densidade maior.

Se aumentarmos o raio dessa partícula hidrofóbica, algo curioso acontece: perto de 1.0 nm, as correlações desaparecem, e acima disso, invertem-se, conforme mostrado em [Chandler (2005)]. Uma correlação menor que um, indica que a densidade das águas de solvatação em torno do grande soluto hidrofóbico ficou menor que do *bulk*. Diz-se então,

quando a partícula hidrofóbica tem raio menor que 1.0 nm, que ela é “molhada” (do inglês *wet*), ao passo que se é maior que um, a partícula torna-se “seca” (*dry*).

Dados experimentais e simulacionais indicam que o fenômeno por trás dessa transição de molhado para seco estaria na incapacidade das águas de solvatação manterem um número de pontes de hidrogênio similares à do *bulk*. Quando o soluto é maior, essas águas tenderão a formar um número menor de pontes, podendo até não formar nenhuma. Nessa condição, elas teriam um comportamento próximo do estado de vapor. Isso levaria à criação de uma interface de duas fases (microscópicas) análogas ao vapor-líquido. E mais: até que se alcance o equilíbrio, parte dessas águas “desconexas” tenderiam a ser capturadas pela consistente rede de pontes de hidrogênio das águas na fase líquida, fazendo com que as primeiras camadas de solvatação em torno do grande soluto tornem-se mais “secas”. Como houve quebra e formação de pontes de hidrogênio, a partir de agora a variação de entalpia passa a ser um componente importante no ΔG [Chandler (2005)].

Chandler e coautores também demonstram que, quando a esfera hidrofóbica é pequena, o ΔG de solvatação escala-se em função de R^3 , já que a dinâmica rede de pontes de hidrogênio está cercado um **volume** excluído pelo soluto. Mas, tão logo essa rede se desfça, na medida que o raio da partícula hidrofóbica cresce além de 1.0 nm, a criação de uma interface análoga ao vapor-líquido faz com o ΔG de solvatação escale-se melhor com R^2 , já que a interface é definida pela área. para formar aglomerados maiores de partículas, induzem mudanças na maneira como a energia livre irá crescer. Se elas ficassem separadas, G cresceria com R^3 ; mas, se juntam e formam agregados cujo raio é maior que 1.0nm, passam a crescer com R^2 . O resultado disso é um $\Delta G < 0$. Se é menor que zero, é espontâneo, o que implica que partículas hidrofóbicas pequenas buscarão naturalmente o equilíbrio em partículas maiores. Pode-se agora entender melhor o comportamento do óleo em nosso macarrão.

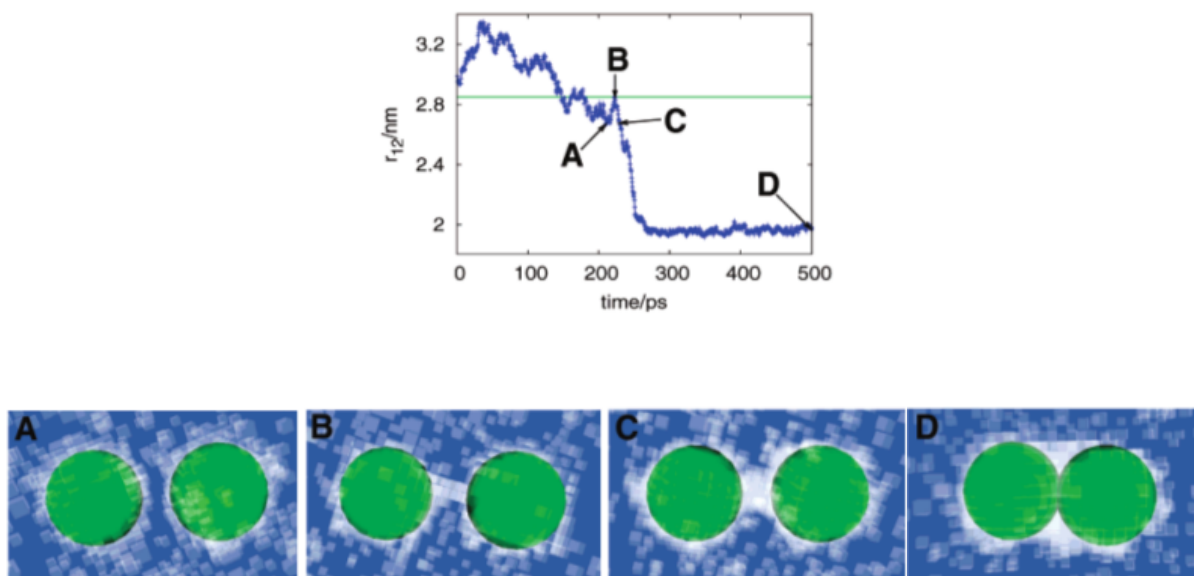
4.1.2 Flutuações da densidade do solvente

O mesmo Chandler em outro trabalho [Willard and Chandler (2008)] mostrou que flutuações na densidade das primeiras camadas de solvatação também são importantes para o processo do colapso hidrofóbico de duas partículas, especialmente aquelas em escalas nanométricas. Para isso, foi usado um modelo de simulação das partículas de 1nm de raio, em água, usando uma série de aproximações (*coarse-graining model*), num reticulado cúbico (*cubic lattice*).

Nessa simulação, como foi percorrido, percebe-se a formação de uma interface vapor-líquido para as primeiras águas de solvatação, conforme visto na figura 4.1, pelas

cores mais claras. Após um certo período difusivo, nota-se inesperadamente a formação de um túnel ou canal de vapor, representando moléculas de água com baixa (ou nenhuma) interação com outras moléculas de água. Esse túnel leva a um desbalançamento de forças sobre as partículas, com os lados opostos ao túnel com mais moléculas de água que os lados que participam do canal. Cria-se, assim, um diferencial de pressão em favor da agregação das duas partículas.

Figura 4.1: Trajetória relativa em vários pontos ao longo do tempo. Cada figura (A, B, C, D) está rotulada na trajetória plotada no espaço unidimensional de r_{12} , indicando a distância entre as partículas.



Fonte: Figura utilizada com permissão de [Willard and Chandler (2008)]. Copyright {2008} American Chemical Society.

A Figura 4.1 mostra uma série de figuras de uma trajetória relativa dos dois solutos do modelo. Essa trajetória é vista como uma função relacionada à separação dos solutos, ou seja, à distância entre eles é dada por $r_{12} = |\vec{r}_1 - \vec{r}_2|$, onde \vec{r}_1 é a posição do soluto 1 e \vec{r}_2 é a posição do soluto 2. Para esse tipo de trajetória são observados três comportamentos. Para os maiores valores de r_{12} (acima de 2.8), as partículas apresentam um movimento difuso independente. As partículas estão agregadas formando um dímero estável quando os valores de r_{12} são menores. Para os valores intermediários de r_{12} , as nanopartículas sofrem agregação que se manifestam nas trajetórias com uma rápida diminuição de r_{12} ao longo do tempo, como pode ser visto no gráfico apresentado no topo da Figura 4.1. Após 200ps, há uma transição de afastamento e aproximação dos solutos (Figura 4.1-A, B, C) culminando na junção dos mesmos (Figura 4.1-D). A partir de D é observável a formação do túnel, ampliado em C. As variações na densidade da água são observadas pelas cores: quanto mais azul, maior concentração de moléculas de água, Em C e D, é observado a maior formação da interface líquido-vapor em volta dos solutos. Na agregação, é obser-

vável que os solutos estão mais “secos”, conforme foi também observado por [Chandler (2005)] para um único soluto com superfície maior e esboçado acima.

Até agora, viu-se estudos teóricos e simulacionais envolvendo sistemas soluto-solvente simples, com modelos de partículas hidrofóbicas esféricas. Muito mais difícil tem sido compreender sistemas mais complexos, como os inerentes à agregação proteína-proteína. Nas interfaces de interação entre duas proteínas, não se tem superfícies de hidrofobicidade e curvaturas homogêneas. Elas são intercaladas por regiões hidrofílicas em meio à superfícies rugosas. Essa intercalação hidrofóbica-hidrofílica numa área irregular pode ter severos efeitos sobre o comportamento das águas de solvatação, tornando difícil previsões sobre a condição seca ou molhada das interfaces.

Mesmo assim, alguns progressos estão acontecendo. Por exemplo, [Patel et al. (2012)] usaram simulação de dinâmica molecular, com técnicas de amostragem especializadas, para mostrar que as flutuações na densidade da água comportam-se similarmente aos modelos idealizados quando próximas a superfícies de biomoléculas complexas, de modo que estas ficam na fronteira de transição de um estado de desumidificação e é sensível a pequenas perturbações. Essa sensibilidade fornece às biomoléculas a capacidade de sintonizar suas interações e funções pela manipulação do contexto, por exemplo, ao confinar a água entre elas ou com mudanças na conformação da superfície, da topologia e da química.

Capítulo 5

Agrupamento espectral em grafos

5.1 Notação matemática

Para este capítulo, usaremos as seguintes notações especiais:

- Matrizes serão identificadas por letra maiúscula em negrito, tipo **A**.
- Vetores serão identificados por letra minúscula em negrito, tipo **v**.
- O restante segue sem negrito.

5.2 Apresentação

Um dos objetivos da Biologia de Sistemas é a análise em larga escala de redes de moléculas, incluindo expressões de genes e interações de proteína, para revelar as relações entre as estruturas da rede e suas funções biológicas. As redes de interação proteína-proteína (PPI - *protein-protein interaction*) podem ser transformadas em um grafo, onde geralmente um nó é uma molécula e uma aresta é a interação. No entanto, se a granularidade da análise é menor, esses nós (vértices) podem ser representados pelos átomos das proteínas e as arestas representarem algum tipo de relação entre eles. De acordo com [Verli et al. \(2014\)](#),

“o emprego da teoria de grafos e suas aplicações têm apresentado um crescimento explosivo devido a sua multidisciplinaridade e ao seu conceito de modelo que permite estudar um objeto específico sem negligenciar o meio em que este objeto se encontra”.

Um modo comum de analisar uma rede consiste em particioná-la em sub-redes responsáveis por funções biológicas específicas. Uma vez que estas podem ser realizadas

por determinados grupos de proteínas, dividir a rede em partes naturalmente agrupadas (grupos ou comunidades - *clusters or communities*) é um modo essencial para investigar algumas relações entre a função e a topologia das redes ou para revelar conhecimento implícito (escondido) por trás deles. Dessa forma, o uso de métodos de agrupamento são importantes nesse contexto, como métodos aglomerativos que empregam uma variedade de medidas de similaridade entre os nós para particionar uma rede PPI e métodos como o agrupamento espectral em grafos.

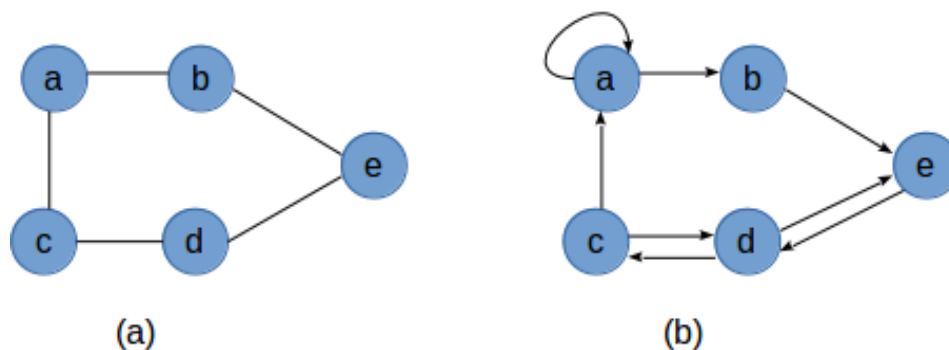
No nosso caso em especial, no objetivo de tentar encontrar correspondências entre átomos hidrofóbicos do inibidor e das peptidases, modelou-se o desafio através de um grafo simbolizando as redes de interações intercadeia em nível atômico. De posse do grafo, o objetivo seguinte foi minerar ou identificar os agrupamentos mais densos, e reforçar ou refutar a hipótese de que há casamentos hidrofóbicos consistentes em complexos inibidor - serino peptidases.

Como o agrupamento espectral foi o método utilizado para caracterização das regiões hidrofóbicas, neste capítulo são apresentados uma visão geral sobre o agrupamento espectral e conceitos correlatos, como grafos e agrupamento em grafos. Uma revisão de algoritmos de agrupamento em grafos pode ser encontrada em [Aggarwal and Wang (2010)] e [Schaeffer (2007)]. Uma visão mais aprofundada sobre teoria espectral em [Chung (1997), Von Luxburg (2007)].

5.3 Grafos: conceitos básicos

Um grafo G é definido pela função $G = (V, E)$, onde V é um conjunto finito e não vazio cujos elementos são denominados nós ou vértices e E é um conjunto de conectores ou arestas que ligam os elementos de V .

Figura 5.1: Exemplos de grafos. Em (a), o grafo pode ser representado por $V = \{a, b, c, d, e\}$ e $E = \{ab, ac, be, cd, de\}$ (ou $E = \{ba, ca, eb, dc, ed\}$, visto que o grafo é não dirigido). A cardinalidade deste grafo ou número de vértices é dada por $n = |V| = 5$ e o número de arestas por $m = |E| = 5$. Em (b), $V = \{a, b, c, d, e\}$ e $E = \{aa, ab, be, ca, cd, dc, de\}$, $n = 5$ e $m = 7$. Neste grafo dirigido, também é exemplificada a ocorrência de laço, representado pela aresta aa .



Na Figura 5.1 são mostrados dois grafos, dois quais um grafo é não dirigido (a) e o outro é dirigido (b). Em um grafo não dirigido, cada aresta é um par não ordenado $\{a, b\}$. Em um grafo dirigido (também chamado de dígrafo), as arestas são pares ordenados, ou seja, cada aresta é orientada para alguma direção ($a \rightarrow b$). Em (a) o grafo pode ser representado por $V = \{a, b, c, d, e\}$ e $E = \{ab, ac, be, cd, de\}$, onde uma aresta de E deve apresentar suas extremidades ligadas a vértices que pertençam ao conjunto V . Um grafo também pode ser definido como um grafo ponderado quando as arestas apresentam valores numéricos. Algumas definições relacionados a grafo não dirigido, sem laços e sem arestas múltiplas são apresentadas abaixo, pois, neste trabalho foi utilizado apenas este tipo de grafo e também são definições usadas no agrupamento espectral. A principal referência bibliográfica utilizada para essas definições é [Von Luxburg \(2007\)](#).

Definição 1 (Função de pesos) *Seja $G = (V, E)$, um grafo ponderado não dirigido, u e v vértices pertencentes a V e uv uma aresta de E . A função de pesos é definida por $w(u, v)$, tal que $w : E \rightarrow \mathbb{R}$, com as seguintes propriedades: $w(u, v) \geq 0$ e $w(u, v) = w(v, u)$.*

Definição 2 (Matriz de adjacência) *A matriz de adjacência \mathbf{W} de um grafo $G = (V, E)$ de ordem n é uma matriz simétrica, quadrada $n \times n$, onde:*

$$w_{ij} = \begin{cases} w(i, j), & \text{se } \{i, j\} \in E \\ 0, & \text{caso contrario.} \end{cases}$$

Se $w(i, j) = 0$, então os vértices v_i e v_j não estão conectados.

Definição 3 (Grau do vértice) *O somatório dos pesos das arestas incidentes em um determinado vértice $v_i \in E$ é o grau desse vértice e é definido por:*

$$d_i = \sum_{j=1}^n w_{ij} \quad (5.1)$$

Definição 4 (Matriz de graus) A matriz de graus \mathbf{D} de um grafo $G = (V, E)$ é definida como a matriz diagonal com os graus d_1, \dots, d_n :

$$d_{ij} = \begin{cases} d_i, & \text{se } i = j \\ 0, & \text{caso contrario.} \end{cases}$$

$$\begin{bmatrix} d_1, & \dots & \dots & 0 \\ 0, & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & d_n \end{bmatrix}$$

Definição 5 (Tamanho de um subconjunto de vértices) Seja A um subconjunto de V ($A \subset V$) do grafo $G = (V, E)$. O tamanho de A pode ser dado pelo número de vértices de A ($|A|$) ou pelo volume de A :

$$\text{vol}(A) = \sum_{i \in A} d_i \quad (5.2)$$

O $\text{vol}(A)$ mede o tamanho de A pela soma dos pesos de todas as arestas conectadas pelos vértices em A .

Definição 6 (Grafo bipartido) um grafo bipartido é um grafo $G = (V, E)$ cujo conjunto de vértices V pode ser separado em dois conjuntos disjuntos U e V tal que toda aresta de E liga somente vértices de U com vértices de V .

5.4 Grafos: particionamento

O particionamento de nós de grafos consiste em dividi-lo em k grupos. Dado um grafo não dirigido $G = (V, E)$, com $V = v_1, v_2, \dots, v_n$ e $E = e_1, e_2, \dots, e_n$, o k -particionamento consistem em dividir o conjunto de vértices em k subconjuntos disjuntos V_1, V_2, \dots, V_n , onde $V_1 \cup V_2 \cup \dots \cup V_n$ e $V_1 \cap V_2 \cap \dots \cap V_n = \emptyset$.

Em [Von Luxburg (2007)], o problema do particionamento de grafos é apresentado da seguinte forma: deseja-se encontrar uma partição do grafo tal que as arestas entre grupos diferentes tenham pesos muito baixos (o que significa que os pontos (vértices) em diferentes grupos são dissimilares entre si), e as arestas dentro de um mesmo grupo, tenham pesos muito altos (o que significa que os vértices dentro do mesmo grupo são semelhantes entre si). O problema do particionamento pode ser abordado por critérios de cortes (*cuts*), que têm como objetivo eliminar arestas de um grafo de forma a produzir ou descobrir subgrupos densamente conexos (k subcomponentes conexas).

5.4.1 Corte de grafos (Cuts)

Um modo simples de se construir uma partição de um grafo é solucionar o problema do corte mínimo (*minicut*, em inglês) [Flake et al. (2004)]. Para uma matriz de adjacências \mathbf{W} , e dois conjuntos $A, \bar{A} \subset V$, onde \bar{A} é o complemento de A , a soma dos elementos de W é dada por:

$$W(A, \bar{A}) = \sum_{i \in A, j \in \bar{A}} w_{ij} \quad (5.3)$$

Para um dado número k de subconjuntos, a abordagem *minicut* consiste em escolher a partição A_1, \dots, A_k que minimiza

$$cut(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k W(A_i, \bar{A}_i) \quad (5.4)$$

Em particular, quando $k = 2$, o *minicut* é um problema fácil de resolver, mas na prática muitas vezes ele não leva a partições satisfatórias. O problema é que, em muitos casos, a solução do *minicut* simplesmente separa um vértice individual dos demais vértices do grafo. Isso não é o desejável, pois bons *clusters* devem agrupar grandes grupos de vértices. Para obter grupos com essas características, duas funções são comumente utilizadas: corte normalizado pelos vértices - *RatioCut* ou *RCut* [Hagen and Kahng (1992)] e o corte normalizado pelas arestas - *NCut* [Shi and Malik (2000)]. No *RCut*, o $cut(A, \bar{A})$ é normalizado pelo número de seus vértices ($|A|$), enquanto no *NCut* o tamanho é normalizado pelo somatório dos pesos de suas arestas ($vol(A)$):

$$RCut(A_1, \dots, A_k) = \sum_{i=1}^k \frac{cut(A_i, \bar{A}_i)}{|A_i|}$$

$$NCut(A_1, \dots, A_k) = \sum_{i=1}^k \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}$$

Para *cuts* normalizados ou balanceados acima, infelizmente o problema de encontrar o corte mínimo é provado ser NP-difícil (*NP-Hard*) [Von Luxburg (2007)] em razão da sua natureza combinatória. Para encontrar a solução ótima, no caso de *clustering* em 2 grupos, teria que se gerar todas as 2^n possibilidades de partição (cada vértice v_i poderia pertencer a A ou \bar{A}), e então escolher qual teria o *Rcut* ou *Ncut* mínimo. É fácil perceber que isso pode ser generalizado para k^n no caso de *clustering* em k grupos. Em razão disso, algoritmos espectrais (seção 5.5.2) foram desenvolvidos como uma heurística viável para resolver esse problema, desde que se relaxem algumas restrições [Von Luxburg (2007)]. Como uma heurística, não se espera a solução ótima, mas uma solução boa ou próxima de ótima.

É possível demonstrar que minimizar o *RCut* passa por uma modelagem com agrupamento espectral não-normalizado e minimizar o *NCut* passa pela modelagem com

agrupamento espectral normalizado. Veremos a parte do *NCut* em mais detalhes daqui a pouco, porque foi a preferida em detrimento do *RCut*. Por quê? Porque se pode perceber que *RCut* normaliza com respeito ao número de vértices em cada *cluster*, independentemente se os clusters formam um agrupamento denso em arestas ou não. Já o *NCut* leva isso em consideração, ao normalizar por $Vol(A_i)$, que por definição, constitui o somatório dos pesos das arestas que tem A_i em uma de suas pontas. Logo, minimizar o *Ncut* tende a produzir agrupamentos semanticamente mais adequados a uma boa parte de problemas do que quando se usa *RCut*, inclusive o nosso problema de encontrar agrupamentos densos hidrofóbicos nas interfaces enzima-inibidor.

5.5 Agrupamento Espectral de Grafos

A teoria espectral de grafos é aplicada a dados que estão representados em forma de grafos. Por ser abrangente é usada em várias áreas. Serão apresentados nesta seção somente os conceitos e técnicas necessários à realização desse trabalho. As principais ferramentas para o agrupamento espectral são as matrizes Laplacianas do grafo: matriz Laplaciana não normalizada e matriz Laplaciana normalizada.

Por que uma matriz tipo Laplaciana e não uma de adjacência comum? Porque, como será visto, ela encerra propriedades intrigantes, muito úteis para o *spectral clustering*. Seu nome advem do fato de representar um caso discreto dos operadores de Laplace-Beltrami, usados em espaços contínuos na geometria diferencial [Skillicorn (2007)].

Para entender essas propriedades, vamos necessitar de um conceito fundamental da álgebra linear.

Definição 7 (Autovetores e Autovalores) *Dada uma matriz quadrada simétrica \mathbf{A} , de tamanho n , um par (escalar, vetor) dado por (\mathbf{v}, λ) , tem-se que:*

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \quad (5.5)$$

Os vetores \mathbf{v} e respectivos escalares λ que satisfazem a equação acima são chamados de eigenvectors e eigenvalues, ou autovetores e autovalores, respectivamente. Se agruparmos os autovetores numa matriz \mathbf{V} e os autovalores numa matriz diagonal $\mathbf{\Lambda}$, a relação abaixo também é válida:

$$\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Lambda} \quad (5.6)$$

A palavra “espectral” no contexto matemático parece ter origem histórica nos trabalhos de Hilbert [Harrell (2015)], como uma analogia (talvez fraca) das decomposições

espectrais da luz. No âmbito da álgebra linear, os “espectros” de uma matriz estão associados ao conjunto de seus autovalores.

Podemos falar também na decomposição espectral de uma matriz:

Definição 8 (Decomposição Espectral) *Dada uma matriz quadrada simétrica \mathbf{A} , de tamanho n , sua decomposição espectral pode ser dada por:*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} \quad (5.7)$$

onde \mathbf{V} é uma matriz de autovetores e $\mathbf{\Lambda}$ uma matriz diagonal de autovalores. Essa decomposição pode ser deduzida a partir da pós-multiplicação por \mathbf{V}^{-1} em ambos os lados de (5.6).

Para as definições que serão apresentadas abaixo, considere o grafo G não dirigido, a matriz de adjacências \mathbf{W} ponderada e simétrica, onde $w_{ij} = w_{ji} \geq 0$, a matriz de graus \mathbf{D} [Von Luxburg (2007)]:

Definição 9 (Matriz Laplaciana não normalizada do grafo)

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \quad (5.8)$$

Na literatura são duas matrizes, relacionadas uma com a outra, chamadas laplacianas normalizadas do grafo. A primeira é uma matriz simétrica (Equação (5.9)) e a segunda (Equação (5.10)) está diretamente relacionada com o *random walk*. Este, em um grafo, consiste de um processo estocástico onde se passa aleatoriamente de um vértice para outro.

Definição 10 (Matrizes Laplacianas normalizadas do grafo)

$$\mathbf{L}_{\text{sym}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} \quad (5.9)$$

$$\mathbf{L}_{\text{rw}} = \mathbf{D}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W} \quad (5.10)$$

Onde \mathbf{I} é a matriz identidade.

5.5.1 Propriedades das matrizes Laplacianas

Aqui são apresentadas as principais proposições que resumem os aspectos mais importantes para no agrupamento espectral.

Mas, antes, vamos precisar ainda desta definição:

Definição 11 (Matrizes Positivas Semidefinidas) *Uma matriz simétrica quadrada $n \times n$ real \mathbf{A} é considerada positiva semidefinida se existe um vetor \mathbf{v} também real tal que $\mathbf{v}^t\mathbf{A}\mathbf{v} \geq 0$.*

5.5.1.1 Matriz Laplaciana não Normalizada

Proposição 1 (Propriedades de \mathbf{L}) *A matriz \mathbf{L} satisfaz às seguintes propriedades intrínsecas:*

1. Para todo vetor $\mathbf{f} \in \mathbb{R}^n$ temos:

$$\mathbf{f}^t \mathbf{L} \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2$$

2. \mathbf{L} é simétrica e positiva semidefinida;

3. O menor autovalor de \mathbf{L} é 0, e o autovetor correspondente é um vetor constante indicador $\mathbb{1}_A = [a, a, \dots, a]^t$, onde a é uma constante real qualquer, e A é um subconjunto de G ;

4. \mathbf{L} possui n autovalores reais e não negativos $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$.

A demonstração da propriedade 1 pode ser encontrada em [Von Luxburg (2007)]. Para a propriedade 2, vemos que \mathbf{L} será simétrica porque \mathbf{D} e \mathbf{W} são simétricas por definição; e será positiva semidefinida (não negativa) dado que $\mathbf{f}^t \mathbf{L} \mathbf{f} \geq 0$, na condição de que w_{ij} são pesos não-negativos e $(f_i - f_j)^2$ será sempre zero ou positivo.

A propriedade 3 é interessantíssima. Se $\lambda = 0$, então pela definição de autovetor-autovalor $\mathbf{L} \mathbf{v} = 0$. Como as linhas de \mathbf{L} somam zero, temos que basta um vetor \mathbf{v} constante para zerar $L\mathbf{v}$.

A propriedade 4 vem do fato que, como \mathbf{L} é positiva semidefinida, ela só admite autovalores não negativos. Para perceber isso, pela definição de autovetores-autovalores, temos que $\mathbf{L} \mathbf{v} = \lambda \mathbf{v}$. Pré-multiplicando por \mathbf{v}^t ambos os lados, temos $\mathbf{v}^t \mathbf{L} \mathbf{v} = \mathbf{v}^t \lambda \mathbf{v} = \mathbf{v}^t \mathbf{v} \lambda = \|\mathbf{v}\|^2 \lambda$. Logo, se $\mathbf{v}^t \mathbf{L} \mathbf{v} \geq 0$, então $\|\mathbf{v}\|^2 \lambda \geq 0$, o que faz com que $\lambda \geq 0$.

Temos, então, como anunciar o importante resultado abaixo:

Proposição 2 (Número de componentes conexas e espectro de L) *Seja G um grafo não direcionado com pesos (w_{ij}) não negativos, a multiplicidade k do autovalor 0 é igual ao número de componentes conexas A_1, \dots, A_k do grafo G . E os autovetores associados a cada autovalor zero serão também vetores indicadores $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_k}$ de cada componente conexo.*

Em outras palavras, o número de autovalores iguais a zero conta a quantidade de elementos conexas descritos por \mathbf{L} . E mais, os valores constantes presentes nos respectivos autovetores são indicativos de quais nós estão associados a cada subgrafo. Isso é uma consequência direta da propriedade 3, dado que, se há subgrafos desconexos dentro do

grafo maior, eles estarão presentes na matriz de adjacência (e na matriz Laplaciana) como submatrizes também desconexas.

A prova dessa proposição passa por perceber que, para o caso mais simples $k = 1$, o grafo contém apenas um componente conexo (ou seja, o grafo todo é conexo), e terá um autovalor zero. Logo, $\mathbf{v}^t \mathbf{L} \mathbf{v} = \|\mathbf{v}\|^2 \lambda = 0$, dado que $\lambda = 0$. Se fazemos $\mathbf{f} = \mathbf{v}$, pela propriedade 1, temos que:

$$0 = \mathbf{f}' \mathbf{L} \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2 \quad (5.11)$$

Como os pesos (w_{ij}) são não-negativos, se dois vértices v_i e v_j estão conectados, então a única forma de zerar a equação acima (5.11) é ter $f_i = f_j$. Logo, f terá que ser constante para todos os vértices que formam o elemento conexo. A extensão para $k > 1$ segue facilmente. Basta perceber que se pode agrupar os vértices de cada componente conexo na matriz \mathbf{L} de modo a formar uma submatriz Laplaciana própria.

Notem a profundidade desse resultado: simplesmente contando os autovalores zerados e olhando para seus respectivos autovetores, já temos condições de ter uma primeira partição (ainda que trivial) em termos de seus subgrafos desconexos.

E as Laplacianas normalizadas são mais interessantes ainda. Uma delas, a chamada *Laplacian Random Walking* - \mathbf{L}_{rw} , tem uma propriedade fundamental para esta tese.

5.5.1.2 Matriz Laplaciana Normalizada

Proposição 3 (Propriedades de \mathbf{L}_{sym} e \mathbf{L}_{rw}) *As laplacianas normalizadas satisfazem às seguintes propriedades:*

1. Para todo vetor $\mathbf{f} \in \mathbb{R}^n$ temos:

$$\mathbf{f}' \mathbf{L}_{sym} \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^n w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2$$

2. λ é um autovalor de \mathbf{L}_{rw} com autovetor u , se e somente se, λ é um autovalor de \mathbf{L}_{sym} como autovetor $w = D^{1/2}u$

3. λ é um autovalor de \mathbf{L}_{rw} com autovetor \mathbf{u} , se e somente se, λ é \mathbf{u} solucionam o problema geral para autovalores e autovetores $\mathbf{L} \mathbf{u} = \lambda \mathbf{D} \mathbf{u}$;

4. 0 é um autovalor de \mathbf{L}_{rw} quando o autovetor correspondente é um constante indicador $\mathbf{1}_A = [a, a, \dots, a]^t$, onde a uma constante real qualquer, e A um subconjunto de G . 0 é um autovalor de \mathbf{L}_{sym} com autovetor $\mathbf{D}^{1/2} \mathbf{1}$;

5. \mathbf{L}_{sym} e \mathbf{L}_{rw} são matrizes semi-definidas positivas e têm n autovalores reais e não negativos $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$.

As demonstrações para \mathbf{L}_{sim} e \mathbf{L}_{rw} são semelhantes às de \mathbf{L} , e detalhes podem ser vistas em [Von Luxburg (2007)]. Mas, cabe destacar que os resultados relativos ao autovalores zeros continuam válidos para ambas, mas os respectivos autovetores indicativos como constantes aplicam-se somente para \mathbf{L}_{rw} .

O resultado maior, que muito nos interessa, vem agora. Para uma bipartição $k = 2$, é possível escolher o vetor \mathbf{f} com dois valores constantes algebricamente estratégicos (um associado a A outro a \bar{A}) tal que, a igualdade $\mathbf{f}^t \mathbf{L} \mathbf{f} = 2vol(V)Ncut(A, \bar{A})$ torna-se válida.

Logo, minimizar $Ncut$ agora nos leva a minimizar $\mathbf{f}^t \mathbf{L} \mathbf{f}$, sob certas condições de contorno. Mas, \mathbf{f} é um vetor discreto de dois valores, e isso ainda nos aprisiona num problema *NP-Hard*. Se relaxarmos essa condição binária em favor de um vetor contínuo $\mathbf{f} \in \mathbb{R}^n$, podemos fazer uso do curioso Teorema de Rayleigh-Ritz [Von Luxburg (2007)], que diz que a solução a essa minimização passa por fazer \mathbf{f} igual ao autovetor de Lw cujo autovalor seja o menor e não-zero. Para $k > 2$, uma generalização possível seria escolher os autovetores com os respectivos k menores autovalores não-zeros. Em essência, essa é a heurística proposta por [Shi and Malik (2000)] para o *spectral clustering* descrita a seguir.

5.5.2 Algoritmo para agrupamento espectral

A ideia principal do agrupamento espectral é a geração de autovetores e autovalores para definir os membros de um *cluster* a partir das matrizes Laplacianas e consequente uso de suas propriedades.

Pode-se perceber nas propriedades acima que os k autovetores associados aos menores autovalores em \mathbf{L}_{rw} carregam significados especiais: são tanto discriminadores dos vértices do grafo associados a cada componente conexo, como também a matriz formada por esses k autovetores pode formar um subespaço simplificado, mais estruturado e menos convoluto, onde algoritmos de agrupamentos mais simples, como k -means e k -medoides, podem produzir bons resultados. Nesse sentido, não deixa de ser valiosa a relação estabelecida pelo poderoso Teorema de Rayleigh-Ritz entre os menores espectros de \mathbf{L}_{rw} e a minimização do $Ncut$.

Eis o algoritmo sugerido em [Shi and Malik (2000)] e utilizado nesta tese para definir agrupamentos entre átomos hidrofóbicos nas interfaces dos complexos peptidases-inibidores. Cabe já adiantar que preferiu-se o k -medoides como algoritmo de agrupamento final.

Algoritmo 1: Agrupamento espectral normalizado (L_{rw})**Entrada:** Matriz de similaridades $S \in \mathbb{R}^{n \times n}$, número k de clusteres

- 1: Construa um grafo de similaridades e faça W ser a matriz de adjacências ponderada
- 2: Calcule a Laplaciana L_{rw} normalizada
- 3: **Calcule os k autovetores** u_1, \dots, u_k de L_{rw} conforme os k menores autovalores
- 4: Faça $U \in \mathbb{R}^{n \times k}$ ser a matriz contendo os vetores u_1, \dots, u_k como colunas
- 5: Para $i = 1, \dots, n$, faça $y_i \in \mathbb{R}^k$ ser o vetor correspondente a i – ésima linha de U
- 6: Agrupe os pontos $(y_i)_{i, \dots, n} \in \mathbb{R}^k$ com o algoritmo k -means ou k -medoides em C_1, \dots, C_k clusteres

Saída: Clusteres A_1, \dots, A_k , com $A_i = \{j | y_j \in C_i\}$

Finalizamos este capítulo com uma breve discussão sobre o algoritmo de agrupamento k -medoides.

5.5.2.1 Algoritmo K -Medoides

O k -medoides é uma variação do algoritmo k -means [MacQueen (1965)]. Ambos os métodos são particionais e têm como objetivo a divisão de um conjunto de dados em k clusteres disjuntos. No k -means, cada grupo é representado pelo seu centro de gravidade, também chamado centroide, dado pela média das medidas usadas no grupo (normalmente a distância). No algoritmo k -medoides, o centro de cada grupo é um objeto representativo, chamado medoide, localizado próximo ao centro do grupo. Por essa razão, esse método é menos sensível a *outliers* do que o k -means, que pode calcular centroides que não estão dentro do conjunto de dados. Em vista disso, o k -medoides é útil para o agrupamento de dados categóricos onde uma média é difícil de definir ou interpretar. Entretanto utiliza um maior tempo de processamento quando comparado ao k -means.

O k -medoides, realiza a cada passo uma busca exaustiva pela troca de um dos k -medoides, previamente selecionados por um dos demais $n - k$ objetos de forma que minimize as dissimilaridades entre os k -medoides e os membros dos k grupos. O algoritmo PAM (*Partitioning Around Medoids*), abaixo descrito, é um algoritmo clássico da família dos métodos k -medoides [Kaufman and Rousseeuw (2009)].

5.5.2.2 Avaliação dos grupos formados

A qualidade de agrupamento de um grafo é aferida por meio de métricas que possibilitam avaliar as melhores formas, dentre as várias possibilidades de agrupar n vértices

Algoritmo 2: Algoritmo K -Medoides (PAM)**Entrada:** Número k de clusteres e a base de dados com N elementos

- 1: Escolher, arbitrariamente, k elementos da base de dados como os medoides iniciais dos grupos
- 2: Repetir:
 - 3: atribua cada elemento remanescente ao grupo com o medoide mais próximo
 - 4: aleatoriamente, selecione um elemento o que não esteja como medoide
 - 5: calcule o custo total, C , de trocar o medoide m_i pelo elemento o
 - 6: se $C < 0$, então troque m_i por o para formar o novo conjunto de k -medoides
 - 7: Até que não haja mudança de objetos de um grupo para outro

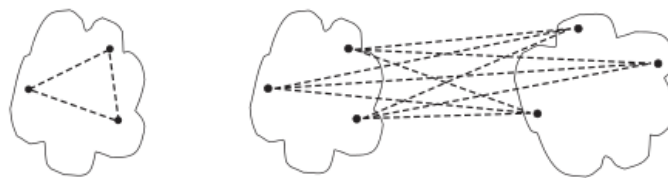
Saída: Conjunto de k clusteres disjuntos

em k conjuntos. Em outras palavras, os grupos determinados por uma métrica de qualidade devem apresentar alta homogeneidade interna (proximidade) e alta heterogeneidade externa, ou seja, os vértices devem ser dissimilares em relação a vértices de outros grupos.

Há várias métricas propostas na literatura para a avaliação da qualidade do agrupamento de grafos, mas não há um consenso sobre qual é a melhor, em razão do modo como são comparadas e dos diferentes tipos de grafos. Entre essas métricas, cita-se: condutância, modularidade, cobertura, *performance* e coeficiente de silhueta [Kannan et al. (2004), Newman and Girvan (2004), Brandes et al. (2007), Van Dongen (2001), Rousseeuw (1987)].

O **coeficiente de silhueta**, métrica empregada para avaliação dos clusteres obtidos neste trabalho (capítulo 8: Resultados e discussões), é um método popular que combina os conceitos de coesão e separação. Para clusteres baseados em grafo, a coesão é definida como a soma dos pesos das arestas que ligam vértices próximos dentro do cluster (Figura 5.2, à esquerda). Do mesmo modo, a separação entre dois grupos pode ser medida pela soma dos pesos das arestas que ligam um vértice de um grupo a um vértice de outro grupo (Figura 5.2, à direita).

Figura 5.2: Visão, baseada em grafo, de coesão (à esquerda) e separação (à direita) de clusteres.



Fonte: [Tan et al. (2005)].

Matematicamente, a coesão e separação são expressas, respectivamente, pelas equações 5.12 e 5.13, onde C_i e C_j são clusteres e u e v são vértices.

$$coesao(C_i) = \sum_{u \in C_i, v \in C_i} proximidade(u, v) \quad (5.12)$$

$$separacao(C_i, C_j) = \sum_{u \in C_i, v \in C_j} proximidade(u, v) \quad (5.13)$$

Um modo de calcular o coeficiente de *silhouette* baseia-se nas distâncias entre os vértices dos grafos, embora seja possível utilizar outra abordagem para definir as similaridades entre os vértices. O *coeficiente de silhueta* para um vértice i é dado pela equação 5.14:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (5.14)$$

Que pode ser escrita como:

$$s_i = \begin{cases} 1 - a_i/b_i, & \text{se } a_i < b_i \\ 0, & \text{se } a_i = b_i \\ b_i/a_i - 1, & \text{se } a_i > b_i \end{cases}$$

Onde, a_i é a dissimilaridade média entre o vértice i e todos os outros vértices do mesmo cluster de i , e b_i é a dissimilaridade média entre o vértice e todos os nós do cluster mais próximo. O valor de s_i é adimensional e varia no intervalo $-1 \leq s_i \leq 1$. Um valor negativo não é desejado, pois corresponde ao caso no qual $a_i > b_i$, ou seja, o vértice i , em média, está mais distante dos vértices de seu próprio grupo do que dos vértices de outro grupo. Se o coeficiente de *silhueta* está próximo a zero, quando $a_i = b_i$, então o vértice i está num ponto intermediário entre dois clusters. Logo, um agrupamento é de boa qualidade quando o coeficiente de *silhueta* é próximo ou igual a 1.

Na Tabela 5.1 é apresentada uma proposta de interpretação de agrupamentos avaliados com o coeficiente de *silhueta* elaborada por [Kaufman and Rousseeuw (2009)].

Tabela 5.1: *Coeficiente de silhueta e interpretação de agrupamentos*

s_i	Interpretação
0.71 – 1.00	Grupos descobertos possuem uma estrutura muito robusta
0.51 – 0.70	Grupos possuem uma estrutura razoável
0.26 – 0.50	Os grupos encontrados possuem uma estrutura fraca e pode ser artificial. É aconselhável tentar outros métodos sobre o conjunto de dados
≤ 0.25	Nenhuma estrutura foi descoberta

O coeficiente de *silhueta* apresenta algumas limitações. Em primeiro lugar, o seu cálculo é muito caro, pois é necessário encontrar todos os pares de menor caminho. O outro limite diz respeito ao seu comportamento quando há clusters com um único vértice. Neste caso, como não há arestas, a distância interna é nula e estes *clusters* são erroneamente considerados de boa qualidade pois $s_i \simeq 1$. Desta forma, agrupamentos constituídos de um único vértice sempre têm altos valores para o coeficiente de *silhueta*, não importando a qualidade dos outros *clusters*.

Capítulo 6

Trabalhos correlacionados

Neste capítulo, são apresentados alguns trabalhos correlacionados que se referem à identificação de regiões hidrofóbicas em proteínas e complexos protéicos. Primeiramente são apresentados os conceitos de área ou superfície acessível ao solvente (ASA) e área de contato entre proteínas. Em seguida, são apresentados alguns métodos para identificação de regiões hidrofóbicas e algumas aplicações do uso de grafos e métodos de agrupamento em redes biológicas, com ênfase para a detecção de regiões hidrofóbicas em proteínas e agrupamento espectral, temas centrais no presente trabalho.

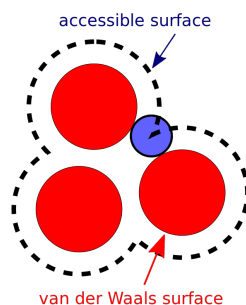
6.1 Cálculo da superfície de proteínas

Como já mencionado anteriormente, as interações entre proteínas são fortemente dependes de suas superfícies. A superfície acessível ao solvente (ASA), em especial, permite inferir sobre essas interações, auxiliando, por exemplo, na localização de sítios de ligação, e no exame de diferenças entre a interface e o resto da superfície, como os tipos de aminoácidos que constituem cada uma dessas superfícies.

Lee and Richards (1971) foram os primeiros a propor o conceito de ASA. Esta superfície, traçada em torno de uma molécula protéica, é descrita pelo centro de uma esfera denominada probe ou esfera teste (em geral uma molécula de água) rolando em torno da proteína. Cada átomo tem o seu raio de van der Waals aumentado com o valor do raio da probe. O raio da probe mais usado quando o solvente é a água é 1.4Å (Figura 6.1)

O software Naccess [Hubbard and Thornton (1993)] usa esse método numérico proposto por Lee and Richards (1971).

Figura 6.1: Área acessível ao solvente.



Fonte: Keith Callenberg, *Illustration of solvent accessible surface in relation to the van der Waals surface of an atom*, 06/05/2010. Licenciado sob CC BY-SA 3.0, via *Wikimedia Commons*. Disponível em:

<https://commons.wikimedia.org/wiki/File:Accessible_surface.svg>. Acesso em Outubro 2015 .

Shrake and Rupley (1973) adaptaram o método de **Lee and Richards (1971)** e propuseram um algoritmo onde a superfície expandida da molécula formada pelos átomos expandidos (raio de Van der Waals do átomo + raio da esfera probe) é discretizada como uma *grid* (*lattice*) que contém um conjunto de pontos. Para cada par de átomos se distribui um número fixo de pontos. O conjunto de pontos que não estão contidos na esfera expandida de qualquer outro átomo é considerado acessível ao solvente.

Eisenhaber et al. (1995) propuseram o algoritmo DCLM (*Double Cubic Lattice Method*) que é uma variante de **Shrake and Rupley (1973)**. Procurou-se com esse método reduzir o tempo de execução do algoritmo original. A ideia é dividir a molécula em cubos de modo a comparar os átomos apenas em suas vizinhanças, não no todo, diminuindo assim o número de comparações entre os pontos gerados para cada esfera. O primeiro cubo determina a vizinhança entre os átomos e o segundo determina os quadrantes de um átomo que interessam efetivamente às comparações, ou seja, os intervalos de pontos em um átomo que são candidatos à oclusão pela projeção de um segundo átomo sobre ele. Esse método numérico é utilizado no Gromacs, um software *open source* de dinâmica molecular projetado principalmente para moléculas bioquímicas como proteínas, lipídios e ácidos nucleicos.

O valor da ASA, além da técnica em si, é afetado pelos valores do tamanho da probe, dos raios de van der Waals, e quando calculada para os aminoácidos individuais, pelo contexto estrutural onde se encontram, isto é, se o resíduo está enterrado na proteína ou na superfície.

6.2 Cálculo da superfície de contato entre proteínas

A ASA, além de ser utilizada para identificar resíduos envolvidos em alguma interação, também pode ser utilizada para anotar resíduos que já foram identificados como participantes de uma interação com métodos baseados em algum delimitador (*cut-off*) de distância e/ou angular, conforme é feito no Piccolo, um banco dados de interações proteínas-proteínas caracterizadas estruturalmente. [Bickerton et al. (2011)]. A ASA também pode ser usada para calcular a superfície de contato/interação, ou seja, a área de superfície que fica enterrada entre duas moléculas de proteínas, também conhecida como interface proteína-proteína.

Por [Chothia and Janin (1975)], o tamanho da interface proteína-proteína (ΔASA) é dado pela equação (6.1), abaixo. Separadamente, três cálculos são realizados. Primeiro, calcula-se a ASA da cadeia A e a ASA da cadeia B em separado. Em seguida, calcula-se a ASA do complexo A U B.

$$\Delta ASA = ASA_A + ASA_B - ASA_{AUB} \quad (6.1)$$

Para Jones and Thornton (1996) a média da interface (ΔASA) na formação do complexo (indo de um estado monomérico para um estado dimérico) é dada pela metade da soma do total de ΔASA para ambas as moléculas ($1/2[\Delta ASA_A + \Delta ASA_B]$). A área de contato em Ponstingl et al. (2000) foi medida de modo semelhante a [Chothia and Janin (1975)], mas a equação (6.1) é dividida pela metade:

$$\Delta ASA = \frac{1}{2}(ASA_A + ASA_B - ASA_{AUB}) \quad (6.2)$$

6.3 Identificação de regiões hidrofóbicas em proteínas

Vários processos biológicos, como a catálise realizada por enzimas e as interações proteínas-proteínas, envolvem reconhecimento molecular, normalmente direcionado pela interação entre partes das superfícies das moléculas envolvidas. Essas partes das superfícies ou *patches* são complementares em termos de aspectos estéricos, eletrostáticos, hidrofóbicos, dentre outros. É comum que esses *patches* sejam analisados quanto ao potencial de solvatação, propensão do resíduo de estar nessa região, planaridade, protusão, área da superfície acessível ao solvente e hidrofobicidade [Jones and Thornton (1997), Janin et al. (1990), Laskowski et al. (1996), Argos (1988)].

Em termos de hidrofobicidade, a interação ocorre por meio do contato entre *patches* hidrofóbicos os quais costumam estar localizados nas interfaces intramoleculares ou subunidades das proteínas e também nas interfaces intermoleculares, quando se considera a associação proteína-proteína [Jones and Thornton (1997), Janin et al. (1990)]. As interações hidrofóbicas estão envolvidas em fenômenos como o enovelamento de proteínas, a interação entre cadeias de proteínas multiméricas, a formação de complexos ligante-proteína, complexos inibidor-enzima, estabilização destes complexos, dentre outros.

Numerosos casos são conhecidos onde a ligação de substratos e cofatores às enzimas ocorre em *patches* hidrofóbicos, frequentemente presentes nos sítios de ligação. Em virtude disso, diversas pesquisas têm sido realizadas para identificar esse tipo de sítio a partir da identificação de regiões onde as interações são mais densas, também denominadas neste caso como *hot spots* [Jain et al. (1994), Keskin et al. (2005), Pettit et al. (2007), Gao and Skolnick (2012)]. No processo de oligomerização de proteínas, os *patches* hidrofóbicos presentes na região de interface entre as subunidades são essenciais, tal que qualquer alteração pode afetar a associação entre elas [Jones and Thornton (1996)].

Os *patches* hidrofóbicos, em cooperação com interações hidrofílicas, são claramente pertinentes para a função da proteína. Estudos empíricos e experimentais têm indicado que frequentemente as proteínas se associam por meio dos *patches* hidrofóbicos em suas superfícies, mas interações polares na superfície também ocorrem e influenciam a formação de complexos [Korn and Burnett (1991)].

Ao longo dos anos, métodos computacionais (*in silico*) têm sido propostos para detectar os *patches* hidrofóbicos. Geralmente esses métodos se baseiam no cálculo da área acessível ao solvente (ASA) e seleção dos resíduos ou átomos como hidrofóbicos.

Korn and Burnett (1991) quantificaram, analiticamente, a hidrofobicidade calculando as complementaridades hidropáticas para as interfaces (superfície de contato entre subunidades) e para o resto da superfície em proteínas multiméricas. É definida uma função de hidropatia que considera as hidropatias atômicas. Em uma das aplicações dessa abordagem, foi mostrado que as superfícies de contatos são mais hidrofóbicas que o resto da superfície. Young et al. (1994) descreveram um método, onde são determinados clusters hidrofóbicos de resíduos na superfície da proteína. Os *clusters* dos resíduos da superfície são ordenados em função da hidrofobicidade total obtida a partir da hidrofobicidade dos aminoácidos constituintes da superfície. Uma *grid* cúbica é usada para representar as posições do C_α e a partir desses pontos estimar uma aproximação da ASA.

Para detectar *patches* hidrofóbicos, Lijnzaad et al. (1996) usaram o DCLM, para identificar a ASA, em conjunto com uma busca em profundidade para selecionar os átomos apolares. Neste método, denominado QUILT, *patches* hidrofóbicos são pedaços contíguos da superfície da proteína, de tamanhos e formas arbitrários e constituídos de carbono e enxofre. Embora por definição, a área de um *patch* é contígua, neste método, ela pode conter *gaps* formados por átomos hidrofílicos. Para resolver isso, os raios dos átomos

polares são expandidos, mas em muitos casos ocluem *patches* hidrofóbicos pequenos.

Gonçalves-Almeida et al. (2012), propôs a metodologia *Hydropace* para identificar *patches* hidrofóbicos em enzimas do tipo tripsina e do tipo subtilisina complexadas com inibidores envolvidos no fenômeno denominado inibição cruzada. Nessa metodologia, cada molécula de enzima foi modelado como um grafo, onde os átomos hidrofóbicos, correspondem aos vértices e os contatos entre eles são representados por arestas. Um *patch* hidrofóbico foi identificado como um componente conectado. Também foi suposto que a propriedade mais importante de um *patch* hidrofóbico é onde ele está posicionado para interagir com o ligante. Então, abstraiu-se sua composição, forma, volume e densidade e um *patch* foi representado como um centroide geométrico, chamado de Hp-centroide (*Hydrophobic Patch Centroid*).

6.4 Agrupamento espectral

O agrupamento espectral tem sido utilizado para várias finalidades na Biologia de Sistemas: identificação de famílias de proteínas, estabelecimento de correlação entre mutações numa determinada proteína, alinhamento múltiplo global de proteínas, detecção de complexos em PPI, detecção de cadeias laterais e agrupamentos hidrofóbicos, melhoria dos particionamentos das redes biológicas.

Em 2007, Brewer (2007) publicou o uso de um algoritmo de agrupamento espectral para analisar dados moleculares. A motivação foi a seleção de *scaffolds* moleculares de um conjunto de dados químicos. Um *scaffold* molecular na química é uma estrutura que pode ser usada para dar suporte à construção de outro material, como um fármaco ou uma proteína. Para demonstrar a aplicabilidade do agrupamento espectral, este método foi aplicado a um conjunto de 125 inibidores de COX-2 (*Cyclooxygenase-2*), uma enzima associada com processos inflamatórios. Posteriormente, o algoritmo de Brewer foi utilizado para agrupar 1800 inibidores de trans-sialidase do *Trypanosoma cruzi* [Neres et al. (2009)] em 690 grupos e para agrupar 2700 compostos em 126 grupos durante o desenvolvimento de um modelo para antagonistas de MCH-1R (*melanin-concentrating hormone-1 receptor*) usados no tratamento da obesidade [Heifetz et al. (2013)].

Inoue et al. (2010) propôs o algoritmo ADMSC (*Adjustable diffusion matrix-based spectral clustering*) que analiticamente particiona as redes de PPI em grupos biologicamente significativos. Esse algoritmo comparado a outros já bem estabelecidos, facilita a decomposição das redes PPI de modo claro e rápido, de acordo com os resultados apresentados no estudo. Para lidar com a heterogeneidade das redes, um fator de potência é introduzido para ajustar a matriz de difusão atribuindo pesos à matriz de adjacências de

acordo com os graus dos nós da matriz. Segundo os referidos propositores desse algoritmo, a análise espectral aplicada ao particionamento de redes PPI tem como resultados, grafos mais significativos biologicamente do que algoritmos baseados em modelos físicos como o MCL (*Markov Clustering*). O MCL [Enright et al. (2002)] é um algoritmo para grafos não supervisionado, rápido e escalável que simula caminhos aleatórios num grafo utilizando matrizes de Markov para determinar probabilidades de transição entre os vértices do grafo. Com esse algoritmo pode ocorrer a perda de proteínas periféricas, mas representativas de interações significativas verificadas experimentalmente, que se conectam aos agrupamentos centrais com poucas arestas.

Qin and Gao (2010) estudaram o agrupamento espectral para a detecção de complexos de proteínas em redes PPI com foco em dois aspectos: construção dos grafos de similaridade e determinação do número de grupos. Segundo esses autores, os resultados obtidos para identificar as proteínas em complexos são comparáveis aos obtidos com aplicação de algoritmos tradicionais para esse fim, como o MCL. Isso prova que o agrupamento espectral é bom para a decomposição de rede PPI.

Paccanaro et al. (2006) empregaram o agrupamento espectral para determinar grupos de proteínas homólogas a partir das sequências. O método é descrito como global, diferindo-se, portanto daqueles que são locais e que têm seus resultados limitados por medidas de distância entre as sequências que devem obedecer a um limiar. Com os métodos espectrais, argumentam esses autores, que ao atribuir uma proteína a um grupo tendo-se em conta todas as distâncias entre cada par de proteínas do conjunto, ou seja, globalmente, pode-se agrupar proteínas cujas sequências protéicas têm baixa identidade.

Nos experimentos realizados, Paccanaro et al. (2006) mostraram que a qualidade dos grupos e a quantidade, apresentaram em média melhorias de cerca de 84% sobre o agrupamento hierárquico, 34% sobre a análise de componentes conectados (*Connected Component Analysis (CCA)*) e 72% sobre o método global TribeMCL (algoritmo baseado no MCL para detecção de famílias de proteínas [Enright et al. (2002)]). O conjunto de sequências foi obtido com no banco de dados SCOP (*Structural Classification of Proteins*) e as famílias obtidas como resultado do agrupamento espectral comparadas (bem) com a própria classificação do SCOP [Andreeva et al. (2008)].

Esse método proposto por Paccanaro et al. (2006) teve a implementação melhorada e foi incorporado a uma plataforma Web denominada SCPS (*Spectral Clustering of Protein Sequences*) [Nepusz et al. (2010)]. Phuc and Phung (2010) para visualizar proteínas com estruturas similares, usou a teoria espectral combinada com redes neurais.

Liao et al. (2009) também aplicaram métodos espectrais para o alinhamento múltiplo global de redes de proteínas. Foi desenvolvido o programa IsoRankN (*IsoRank-mordidela*) onde a teoria espectral é aplicada a grafos que contêm os *scores* dos alinhamento par-a-par. O método não requer treinamento ou dados filogenéticos. Para demonstrar sua efetividade, esse algoritmo foi testado em cinco redes eucarióticas (*Homo sapiens*

(*human*), *Mus musculus* (*mouse*), *Drosophila melanogaster* (*fly*), *Caenorhabditis elegans* (*worm*) and *Saccharomyces cerevisiae* (*Yeast*)). Pelos resultados do estudo, o IsoRankN apresentou boa cobertura e consistência quando comparado a outras abordagens e pode ser utilizado para a predição de ortólogos. O agrupamento espectral também tem sido utilizado para agrupar conformações de proteínas a partir de simulações dinâmicas [Mittag and Forman-Kay (2007)].

Na genômica, Perkins and Langston (2009) aplicaram técnicas espectrais em grafos para desenvolver um método sistemático para a seleção de um limiar (*threshold*) em redes de expressão gênica. Autovalores e autovetores foram obtidos pela transformação da matriz de adjacências construída com valores de (*threshold*). Foi utilizado um método básico de agrupamento espectral para examinar o conjunto de relações gene-gene e selecionar o (*threshold*) dependendo das estruturas dos dados. A abordagem foi aplicada a dois *microarrays* com dados de *Homo sapiens* e *Saccharomyces Cerevisi*. [Yu et al. (2012)] utilizaram agrupamento espectral para a descoberta de classes de câncer a partir de perfis de expressão gênica e Loss et al. (2010) para a predição de genes regulados epigeneticamente em linhagens de células de câncer de mama.

Kannan and Vishveshwara (1999) apresentou um método baseado na teoria espectral em grafos para encontrar grupos de cadeias laterais em proteínas e identificar o resíduo que tem o maior número de interações entre os resíduos de cada grupo. Detectar tais clusteres e centros do cluster são importantes no enovelamento das proteínas. Além dos clusteres formados pelas cadeias laterais o método também é aplicado para pesquisar por domínios nas proteínas e por clusteres hidrofóbicos.

Capítulo 7

Materiais e Métodos

Neste capítulo são apresentados os procedimentos realizados na montagem e seleção da base de dados, os métodos e técnicas empregados para descoberta de regiões hidrofóbicas correspondentes na interface de complexos formados por enzimas do tipo tripsina e subtilisina (serino peptidases) e inibidores protéicos.

Os complexos protéicos utilizados neste trabalho foram obtidos do banco de dados PDB (*Protein Data Bank*) [Berman et al. (2000b)]. Também foram utilizados dados do PFAM [Finn et al. (2014)] e MEROPS [Rawlings et al. (2014a)]. Para descoberta de regiões hidrofóbicas (enzima-inibidor) empregou-se o agrupamento espectral. Os *scripts* foram desenvolvidos em R, PERL e PYMOL.

7.1 Bases de dados utilizadas

7.1.1 PDB - *Protein Data Bank*

O PDB (<http://www.rcsb.org/>) é um banco de dados que contém informações sobre estruturas tridimensionais de macromoléculas biológicas, incluindo proteínas, ácidos nucleicos e outros complexos [Berman et al. (2000a)]. Esses dados, são obtidos por meio de experimentos de difração de raios-X, de Ressonância Magnética Nuclear (RMN) e microscopia eletrônica. Criado em 1971 com apenas 7 (sete) estruturas de proteínas, no momento que esta tese é escrita contém cerca de 110619 moléculas com estruturas resolvidas.

No PDB, à cada estrutura depositada, determinada independentemente, é atribuído um identificador de quatro caracteres únicos (PDBid), no qual o primeiro caractere é um dígito (1-9) e os demais são letras que podem indistintamente serem maiúsculas ou minúsculas. Por exemplo, 1GCI é o PDBid da estrutura da enzima subtilisina (Figura

3.4, capítulo 3).

A informação primária armazenada no PDB consiste de arquivos de coordenadas para as moléculas biológicas. Esses arquivos listam os átomos em cada proteína e sua localização no espaço 3D. Estes arquivos estão disponíveis em vários formatos (PDB, mmCIF, XML, PDBx/mmCIF). Até 2014 o formato PDB era considerado o arquivo padrão do banco PDB, e foi substituído pelo formato PDBx/mmCIF. O formato de arquivo PDB será retirado em 2016.

Um típico arquivo no formato PDB inclui uma seção “header” de texto que resume a proteína, informação de citação e os detalhes de como a estrutura foi determinada, seguidos pela sequência e uma longa lista de átomos e suas coordenadas. Esse arquivo contém também observações experimentais que são usadas para determinar as coordenadas atômicas.

O formato PDB tem restrições quanto ao número de átomos e cadeias poliméricas em virtude do seu formato fixado em 80 colunas e campos com tamanho fixo. Consequentemente, grandes estruturas não podiam ser acomodadas nesse formato. O arquivo mmCIF (*macromolecular Crystallographic Information*) foi introduzido, em 1996, para solucionar esse problema. Seu dicionário continha mais de 3000 definições de conceitos que abrangem os resultados de experimentos cristalográficos, bem como os experimentos em si mesmos. O mmCIF prevê tipificação e relacionamentos entre os dados e em razão de ser auto-definido é ideal para aplicações computacionais. Quando foram incluídas definições nesse dicionário para RMN e microscopia eletrônica (3DEM), ele foi renomeado para PDBx [Berman et al. (2014)].

Os complexos enzima-inibidor constituintes da base de dados deste trabalho são arquivos PDB obtidos do banco PDB.

7.1.2 MEROPS

O MEROPS, versão 9.13, foi utilizado como referência para a classificação das enzimas do tipo tripsinas e do tipo subtilisinas e consequente caracterização da relação evolutiva, proporcionada pela própria estruturação dessa base de dados.

Na construção da base de dados deste trabalho, foram selecionados, em razão dos critérios adotados (descritos na seção 7.3), complexos com enzima do tipo tripsinas da família S1 e subfamílias S1A e S1E. A família S1 inclui enzimas como a quimotripsina, tripsina, elastase, streptogrisin, dentre outras.

As enzimas do tipo subtilisinas selecionadas pertencem à família S8. A família S53 não teve exemplares selecionados para constituir a base de dados. Isso se deve ao fato

de que a maioria das estruturas resolvidas e com representantes dessa família no banco de dados PDB correspondem somente à enzima (com ou sem mutações). Quando existe o complexo enzima-inibidor, o inibidor não atendia a um dos critérios de seleção que é ter comprimento igual ou superior a dez resíduos. Esse critério foi determinado, tendo-se como base o estudo de [Li et al. (2007)], no qual foi sintetizado, a partir do *loop canônico* de um pequeno peptídeo ORB (AALKGCWTKSIPPKPCFGKR), um potente inibidor de tripsina contendo onze resíduos, denominado *Trypsin Inhibitory Loop*(TIL). Uma vez que esse estudo mostrou a existência desse inibidor com esse tamanho e seu potencial de inibição, definimos o número de dez resíduos como o limite inferior para o tamanho do inibidor.

7.1.3 PFAM

O PFAM é uma base de dados de famílias de proteínas, representadas por alinhamentos múltiplos de sequência e por modelos ocultos de Markov (*hidden Markov models-HMMs*). O PFAM, assim como o MEROPS, agrupa em clãs famílias relacionadas entre si pela similaridade de sequência, estrutura ou por modelos HMM. É muito usado para confirmar a identidade de uma proteína por meio dos domínios da proteína ou partes estruturais, não somente pela sequência dos aminoácidos.

A cada família é atribuído um código de acesso, como por exemplo, o código *PF00089* para a família das tripsinas. Esse código foi utilizado no presente trabalho como parte do protocolo de pesquisa submetido ao PDB (*Protein Data Bank*) para seleção dos complexos enzima-inibidor, visto que os identificadores do MEROPS não são parâmetros compatíveis com esse protocolo. Foram obtidas no próprio MEROPS as correspondências entre os códigos PFAM e a classificação MEROPS.

Ressalta-se que o MEROPS é um banco de dados específico para a classificação das peptidases, enquanto o PFAM é uma coleção abrangente de alinhamentos e modelos de Markov ocultos que representam famílias de proteínas e domínios. Dada a especificidade do MEROPS para as peptidases, optamos por utilizar a classificação desse banco de dados e fazer a correspondência com o PFAM que oferece o código que pode ser usado como entrada para pesquisa no PDB e conseqüente obtenção dos arquivos com dados dos complexos.

7.2 Scripts utilizados

Os algoritmos desenvolvidos foram implementados em Perl, Pymol e R. Os *scripts* em Perl foram escritos para geração das matrizes de adjacências de cada complexo, tendo como valores as distâncias euclidianas entre os átomos da enzima e do inibidor. Essas matrizes foram usadas como entrada para o cálculo da área de contato entre os átomos da enzima e os átomos do inibidor, implementado em R.

No Pymol, os *scripts* foram gerados como método alternativo para o cálculo da área de contato. No entanto, devido ao alto tempo de execução por complexo, esse método foi descartado. Também foram implementados no Pymol, *scripts* para realizar o alinhamento estrutural de um complexo escolhido como *template* (1PPF) com os demais complexos. Esse alinhamento foi feito tendo como referência as tríades catalíticas das enzimas de ambos os complexos.

Como já foi citado, o R foi utilizado para a implementação do cálculo da área de contato. Também foram desenvolvidos em R, os algoritmos para o agrupamento espectral visando a identificação das regiões hidrofóbicas nas enzimas e nos inibidores e os gráficos que serão apresentados no capítulo 8 (Resultados e discussões).

7.3 Construção da base de dados

O ponto de partida para a seleção de complexos serino peptidases e inibidores protéicos foi a base de dados constituída por 9 complexos com inibição cruzada utilizada no Hydropace - uma metodologia desenvolvida pelo nosso grupo de pesquisa em trabalho anterior para análise de inibição cruzada em serino peptidases por meio de centroides de regiões hidrofóbicas [Gonçalves-Almeida et al. (2012)]. Essa base foi utilizada para os estudos iniciais, pois utilizando o paradigma dividir para conquistar [Shaffer (2013)], o problema foi considerado em menor escala para depois estender as proposições, análises e resultados para uma base de dados mais ampla. Essa base é constituída por enzimas do tipo tripsina e do tipo subtilisina e há 5 (cinco) complexos com inibição cruzada com o inibidor Eglina C (Tabela 7.1) e 4 (quatro) com inibição cruzada com o inibidor Ovomucoide-OMTKY3 (Tabela 7.2).

Dentro do universo dessa base de dados inicial e após algumas análises visuais sobre a interface entre enzima e inibidor, tanto de regiões polares como apolares, foram escolhidos 4 complexos para análise de grupos hidrofóbicos e treinamento dos algoritmos. Nas

regiões de interface, foram identificadas semelhanças estruturais que poderiam supervalorizar uma subclasse em detrimento a outra. Então, restringiu-se o estudo aos complexos 1PPF, 1ROR (tipo tripsina e tipo subtilisina, respectivamente, com Ovomucoide) e 1ACB e 1TEC (tipo tripsina e tipo subtilisina, respectivamente, com Eglina C).

Tabela 7.1: Tabela de inibição cruzada com Eglina C

PDB ID	Molécula-enzima	Família
1ACB	<i>Alpha-Chymotrypsin</i>	Tipo tripsina
1TEC	<i>Thermitase</i>	Tipo subtilisina
1CSE	<i>Subtilisin Carlsberg</i>	Tipo subtilisina
1MEE	<i>Mesentericopeptidase</i>	Tipo subtilisina
1SBN	<i>Subtilisin Novo BPN'</i>	Tipo subtilisina

Tabela 7.2: Tabela de inibição cruzada com Ovomucoide (OMTKY3)

PDB ID	Molécula-enzima	Família
1ROR	<i>Subtilisin Carlsberg</i>	Tipo subtilisina
1PPF	<i>Human Leukocyte Elastase</i>	Tipo tripsina
1CHO	<i>Alpha-Chymotrypsin A</i>	Tipo tripsina
3SGB	<i>Protense B (SGPB)</i>	Tipo tripsina

Para ampliar a base de dados foram selecionadas serino peptidases do tipo tripsina e subtilisina. A escolha desses dois tipos se justifica pela realização dos estudos preliminares realizados, cuja intenção inicial era aplicar o Hydropace no lado do inibidor e porque essas serino peptidases são bem conhecidas e têm grande importância biológica.

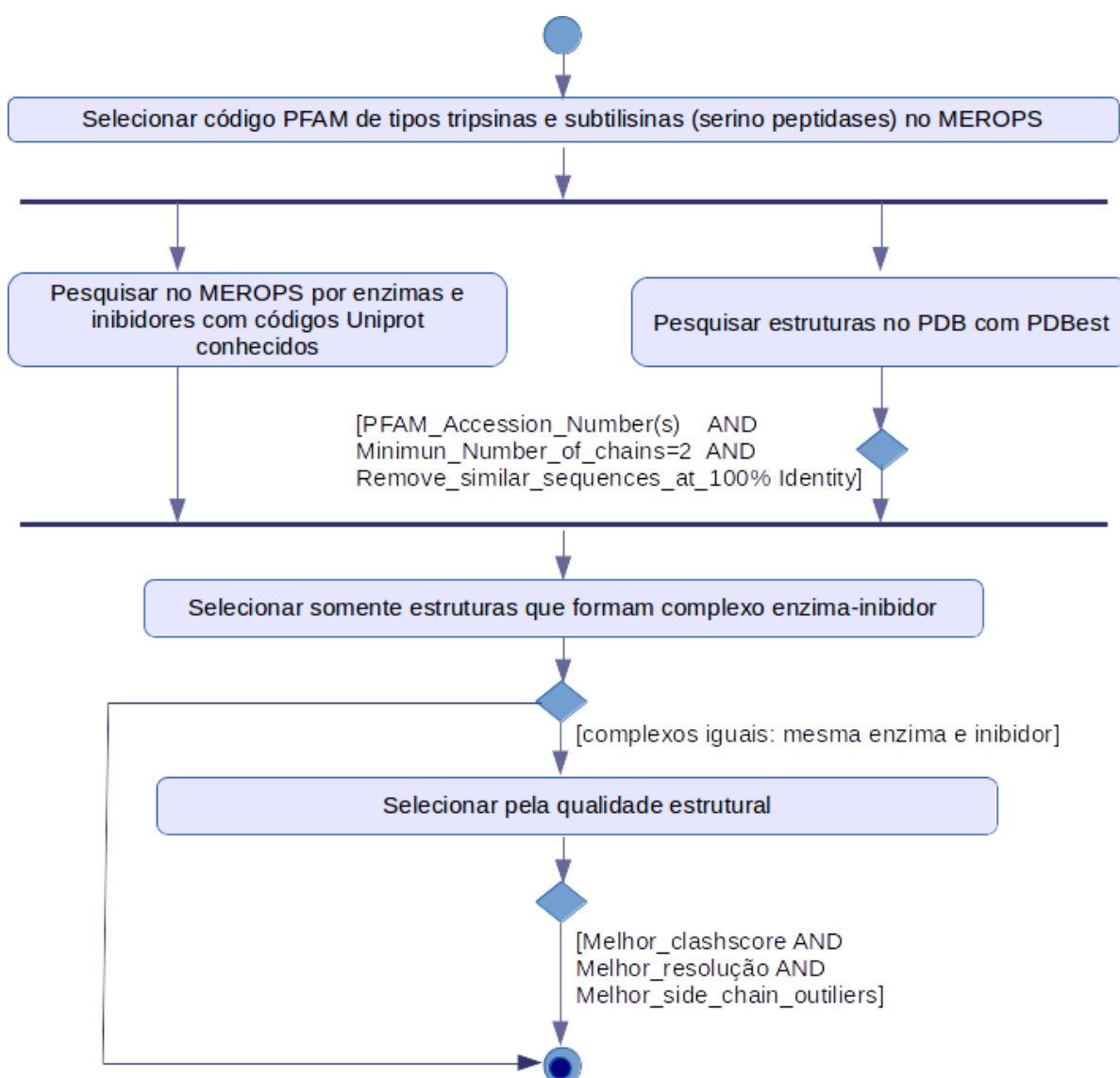
Foi feita uma pesquisa no MEROPS, visto que esse banco de dados de peptidases é atualizado constantemente, para identificar os códigos PFAM das tipo tripsinas e subtilisinas. No PDBEST (PDB *Enhanced Structures Toolkit*) [Goncalves et al. (2015), Pires et al. (2007)], software utilizado para pesquisar o banco de dados PDB (*Protein Data Bank*) e realizar automaticamente o download dos arquivos, não é possível fornecer como parâmetro o identificar do MEROPS para as enzimas, mas é possível informar o código PFAM. Por isso, foram identificados os códigos PFAM correspondentes às classificações no Merops. O protocolo de busca executado no PDBEST tem os seguintes critérios:

- *Id(s) and Keywords: Pfam Accession Number(s) Search;*
- *Structure Features: Minimum Number of chains = 2;*
- *Remove Similar Sequences at 100% Identity.*

Com esses critérios, foram selecionadas 132 estruturas. Por meio de uma análise manual das estruturas, removemos aquelas que correspondiam a uma estrutura que era uma enzima ou um inibidor com mais de uma cadeia. Também removemos os inibidores cuja sequência de aminoácidos apresentava tamanho menor ou igual a dez. Resultaram,

então 30 estruturas. Para descobrir novos complexos, a partir das enzimas e dos inibidores identificados nas 30 selecionadas e também considerando inibidores de serino-peptidases identificados na revisão bibliográfica fizemos uma pesquisa no MEROPS pelo código uniprot dessas enzimas e inibidores. Para os casos nos quais foram identificadas mais de uma estrutura para a mesma enzima e mesmo inibidor, selecionamos apenas um representante com os seguintes critérios relacionados à qualidade estrutural: melhor *clashscore*, melhor resolução e melhor *sidechain outliers* [Read et al. (2011)]. A base final resultante passou a ser composta por 36 complexos enzima-inibidor. A Figura 7.1 apresenta o fluxo dessas atividades de construção da base de dados.

Figura 7.1: Fluxograma para construção da base de dados



Em razão da existência de várias famílias de inibidores de serino peptidases, procurou-se construir a base de dados de modo abrangente, então foram escolhidas as famílias dos inibidores identificados na inibição cruzada no trabalho de [Gonçalves-Almeida et al. (2012)] e aquelas famílias mais estudadas que contemplassem o reino animal e vegetal,

que são: *ecotin*, *Kasal*, *BPTI (Kunitz)*, *Bowman-Birk*, *Potato I e II*, *Squash*, *Serpin*, *WAP* (Vide na Tabela 7.5).

7.3.1 Classificação das enzimas e inibidores

As enzimas e inibidores dos complexos selecionados foram categorizados de acordo com a classificação utilizada no MEROPS. O princípio organizacional desse banco de dados é uma classificação hierárquica na qual conjuntos de peptidases e sequências de inibidores protéicos são agrupados em espécies de peptidases e de inibidores, os quais são agrupados em famílias que são agrupadas em clãs. Uma família contém sequências relacionadas e um clã contém estruturas relacionadas [Rawlings et al. (2014a)]. A análise sequencial está restrita à porção da proteína diretamente responsável pela atividade da peptidase ou do inibidor.

Por meio desses conceitos de famílias e clãs, as enzimas e os inibidores foram organizados nas tabelas 7.3 e 7.4, respectivamente. As enzimas estão organizadas nas famílias S1 e S8. As enzimas tipo tripsinas estão na família S1. A família S8 inclui as subtilisinas. Pode-se observar também que em cada grupo (tipo tripsinas e tipo subtilisinas) as enzimas estão organizadas em subfamílias, o que indica que há diversidade nas sequências do conjunto selecionado. Os complexos com mesma identificação de enzima, por exemplo, 1PPE, 1TAW e 1F2S na tabela 7.4 contêm enzimas que têm maior similaridade sequencial, mas não de 100%, dado que a remoção de 100% de identidade foi critério utilizado na seleção dos complexos. Adite-se que essas enzimas mais próximas sequencialmente estão complexadas com inibidores diferentes. Essas mesmas observações se aplicam ao caso dos inibidores (Tabela 7.3).

Do ponto de vista de similaridade estrutural, a evidência de relacionamento evolucionário entre uma ou mais famílias é dada pelo agrupamento em clãs. As enzimas do tipo tripsinas estão no clã PA, que é um clã misto contendo enzimas com tipo catalítico para Serina(S), Cisteína(C), Treonina(T) e as tipo subtilisinas estão no clã SB. Os inibidores estão distribuídos em vários clãs mostrando diversidade para os dados selecionados.

A tabela 7.5 mostra a relação entre os inibidores e as enzimas selecionados e o identificador PDB dos complexos nos quais os inibidores e enzimas estão combinados. Nesta tabela também pode ser observado, que das estruturas escolhidas onde a enzima é a mesma, os inibidores são diferentes. Ressalta-se que pelos Clãs, é possível verificar a diversidade do ponto de vista estrutural, maior para os inibidores.

Tabela 7.3: Classificação dos inibidores segundo o Merops. O inibidor WCI da família I3 pertence a subfamília A. Os demais inibidores selecionados não possuem subfamílias.

Inibidores			
Família	Clan	Sigla	PDB
I1	IA	OVO GRG	1Z7K 1HJA 1PPF 3SGB 1R0R 4GI3
I2	IB	BPT APP BPT	2FI5 4DG4 1T8O 1TAW 1YC0
I52	IB	TAP	1KIG
I3	IC	WCI	3VEQ
I4	ID	SER	1OPH 1K9O
I7	IE	SQF	1MCT 1F2S 1PPE
I9	JC	POA	1V5I
I11	IE	ECO	1XX9 1Ezs 1FI8
I12	IF	BBI	3MYW
I13	IG	EGL CL2	4B2B 1ACB 4H4F 1SBN 2SEC 1TEC 1TM1 1LW6
I17	IP	SLP SKI	4DOQ 2Z7F 1FLE
I20	JO	POT WIP	4SGB 1OYV

7.3.2 Normalização dos dados

Os complexos enzima-inibidor foram normalizados por meio do PDBest. O objetivo desse procedimento foi corrigir possíveis inconsistências e anomalias nos arquivos PDB, como por exemplo, a duplicação de átomos. A remoção de átomos de hidrogênio e separação das cadeias referentes à enzima e ao inibidor também foram procedimentos realizados com o PDBest.

A decomposição do arquivo PDB em duas partes (enzima e inibidor) teve a finalidade de analisar separadamente as características sequenciais e estruturais das enzimas e dos inibidores. Nos casos, onde o PDB contém mais de uma cadeia correspondente à enzima e estas são iguais, escolhemos apenas uma cadeia. O PDB do complexo 1FI8, por exemplo, tem as cadeias enzimáticas A e B iguais. Então selecionamos somente a cadeia A. Quando a enzima é formada por cadeias que não são iguais, consideramos todas essas cadeias. No PDB 1HJA, por exemplo, as cadeias ABC foram consideradas como componentes da enzima. Os mesmos critérios foram usados para os inibidores.

Tabela 7.4: Classificação das enzimas segundo o Merops. As enzimas tipo tripsina selecionadas pertencem ao subclã PA(S), enquanto as tipo subtilisinas não são agrupadas em subclãs. Todas as enzimas estão classificadas na subfamília A, exceto a tipo tripsina SGY ou SGPB dos complexos 3SGB e 4SGB, cuja subfamília é E.

Enzimas						
PDB	Clan	Família	Sigla	Tipo	Organismo	
1ACB, 1HJA, 1T8O	PA	S1	CHY	Tipo tripsina	<i>Bos Taurus</i>	
4H4F	PA	S1	CHY	Tipo tripsina	<i>Homo sapiens</i>	
1PPE, 1TAW, 1F2S,	PA	S1	TRY	Tipo tripsina	<i>Bos Taurus</i>	
1OPH, 2FI5, 3VEC, 4B2B						
1Z7K, 1MCT, 3MYW, 4DOQ	PA	S1	TRY	Tipo tripsina	<i>Sus scrofa</i>	
1K9O, 1E2S	PA	S1	TRY	Tipo tripsina	<i>Rattus norvegicus</i>	
4DG4	PA	S1	TRY	Tipo tripsina	<i>Homo sapiens</i>	
1KIG	PA	S1	XAY	Tipo tripsina	<i>Bos Taurus</i>	
1XX9	PA	S1	XAY	Tipo tripsina	<i>Homo sapiens</i>	
1F18	PA	S1	GRY	Tipo tripsina	<i>Rattus norvegicus</i>	
1YC0	PA	S1	HGY	Tipo tripsina	<i>Homo sapiens</i>	
1PPF, 2Z7F	PA	S1	ELY	Tipo tripsina	<i>Homo sapiens</i>	
1FLE	PA	S1	ELY	Tipo tripsina	<i>Sus scrofa</i>	
3SGB, 4SGB	PA	S1	SGY	Tipo tripsina	<i>Streptomyces griseus</i>	
1LW6, 1SBN, 1TM1, 1V5I	SB	S8	BNP	Tipo subtilisina	<i>Bacillus amyloliquefaciens</i>	
1OYV, 1R0R, 2SEC, 4GI3	SB	S8	CAN	Tipo subtilisina	<i>Bacillus licheniformis</i>	
1TEC	SB	S8	THN	Tipo subtilisina	<i>Thermoactinomyces vulgaris</i>	

Tabela 7.5: Relação de inibidores e enzimas em complexo

PDB	Inibidores			Enzimas				
	Clã	Família	Sigla	Clã	Família	Sigla		
1Z7K	IA	I1	OVO	PA	S1	TRY		
1HJA						CHY		
1PPF						ELY		
3SGB						SGY		
1R0R				SB	S8	CAN		
4GI3							GRG	
2FI5	IB	I2	BPT	PA	S1	TRY		
4DG4						CHY		
1T8O								
1TAW			APP	PA	S1	TRY		
1YC0			PBP			HGY		
3VEQ	IC	I3	WCI	PA	S1	TRY		
1OPH	ID	I4	SER	PA	S1	TRY		
1K9O								
1MCT	IE	I7	SQF	PA	S1	TRY		
1F2S								
1PPE								
1V5I	JC	I9	POA	SB	S8	BPN		
1XX9	IN	I11	ECO	PA	S1	XAY		
1EZS						TRY		
1FI8						GRY		
3MYW	IF	I12	BBI	PA	S1	TRY		
4B2B	IG	I13	EGL	PA	S1	TRY		
1ACB						CHY		
4H4F								
1SBN				SB	S8	BPN		
2SEC						CAN		
1TEC			THN					
1TM1			IG		CL2	SB	S8	BPN
1LW6								
4DOQ			IP	I17	SLP	PA	S1	TRY
2Z7F	ELY							
1FLE	SKI							
4SGB	JO	I20	POT	PA	S1	SGY		
1OYV			WIP	SB	S8	CAN		
1KIG	IB	I52	TAP	PA	S1	XAY		

7.4 Cálculo da área de contato entre átomos da enzima e inibidor

As interações entre moléculas são fortemente dependentes de suas superfícies, de modo particular, a interface entre elas permite inferir sobre alguns tipos de interações e a influência que uma exerce sobre a outra, auxiliando, por exemplo, na formação de complexos enzima-inibidor. Há muitas formas de definir uma interface entre duas cadeias. Uma delas é pelo conjunto de átomos ou resíduos na interface e que perderam acessibilidade ao solvente (ASA) [Janin et al. (1990)]. Outra forma, a usada nesta tese, envolve o cálculo da área de contato (Ac) entre átomos intercadeia. No nosso caso, entre átomos da enzima com átomos do inibidor. São da interface todos os átomos que apresentam $Ac > 0$.

A Ac tem um conceito que é oposto ao ASA. Enquanto o último envolve encontrar a área exposta ao solvente, o primeiro calcula a área que o solvente não teve acesso. Ou seja, ao se rolar uma *probe* esférica (representando a água) ao redor de dois átomos (também considerados esféricos) em contato, a superfície dos átomos que a *probe* tocou constitui a ASA, e a que ela não tocou a Ac .

A Ac pode ser calculada com base na ASA, e tem sido o método padrão na literatura. Essa abordagem, essencialmente heurística e aproximativa, foi utilizada no início de nossos trabalhos, mas demandava um tempo computacional relativamente alto para cada complexo, da ordem de dezenas ou centenas de minutos, numa estação Quadcore padrão. Como alternativa, foi desenvolvido uma metodologia puramente analítica, descrita por uma única equação, em que a escala de tempo de processamento foi reduzida a poucos milissegundos por complexo. Isso também simplificou deveras os *scripts* envolvidos nos cálculos. Tal método, chamado Silveira-Romanelli (SR), foi o adotado nesta tese, e será descrito em detalhes a seguir.

7.4.1 Cálculo da Ac utilizando a ASA

A Ac calculada a partir da ASA e exemplificada na Figura 7.2 pode ser obtida, semelhantemente ao que foi apresentado na seção 6.2 do capítulo 6, da seguinte forma para um par de átomos:

$$Ac_{ei} = (ASA_{atm_i} + ASA_{atm_e}) + ASA_{ei} \quad (7.1)$$

Onde:

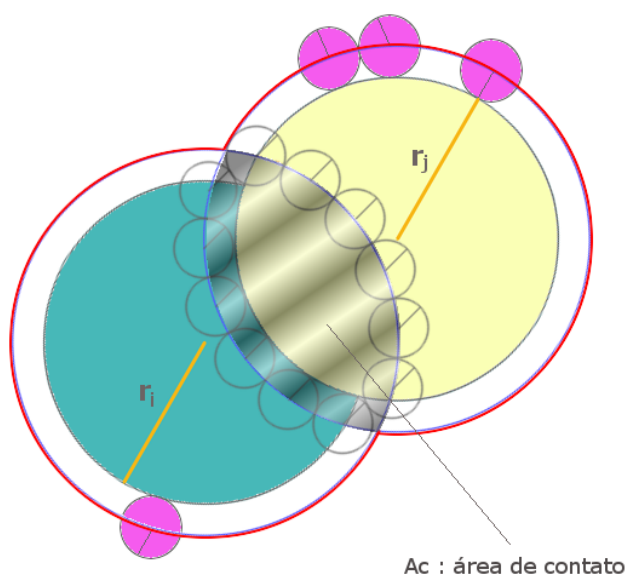
Ac_{ei} é a área de contato para dois átomos e e i , pertencentes, respectivamente à enzima e ao inibidor;

ASA_{atm_i} é a ASA de um átomo do inibidor considerado sozinho;

ASA_{atm_e} é a ASA de um átomo da enzima considerado sozinho;

ASA_{ei} é a ASA calculada para dois átomos juntos, em contato. Sendo um do inibidor e o outro da enzima.

Figura 7.2: Ac para dois átomos vizinhos de raios r_i e r_j .



O Pymol foi utilizado para calcular a Ac a partir da ASA. A função `get_area` desse software calcula a ASA por átomo ou para um conjunto de átomos, isto é, o cálculo é feito dependendo do objeto que é criado. Esse pode ser constituído de um átomo ou conjunto de átomos ou mesmo de resíduos. O resultado é uma aproximação discreta da ASA e depende também da densidade dos pontos (*dot_density*) escolhida entre 1 e 4. Nos experimentos realizados foi escolhido o valor 4 para *dot_density*. Abaixo é apresentado um fragmento de um *script* desenvolvido para o cálculo da Ac , obedecendo à equação (7.1).

```
set dot_solvent, 1
```

```

set dot_density, 4
create at154, 1PPF and (id 154)
asa_at154 = cmd.get_area("at154")
create at1724, 1PPF and (id 1724)
asa_at1724 = cmd.get_area("at1724")
create ats154_1724, at154 at1724
asa154_1724 = cmd.get_area("ats154_1724")
sumAse = asa_at154 + asa_at1724
delta = sumAse - asa154_1724

```

7.4.2 Cálculo da A_c utilizando SR

A metodologia SR foi desenvolvida conjuntamente pelo orientador desta tese, Prof. Carlos Silveira, e seu colega matemático da Unifei - Campus de Itabira, Prof. João Romanelli. O primeiro encontrou a equação que descreve a A_c entre dois átomos esféricos iguais (de mesmo raio de van der Waals), e o segundo a generalizou para átomos de raios diferentes. Um resumo da dedução da equação SR encontra-se no apêndice A.

A equação SR é usada com um duplo propósito na metodologia homônima: definir a interface como o conjunto de átomos envolvidos em contatos intercadeia cujo $A_c > 0$; utilizar esse A_c como peso para as arestas do grafo de contatos entre átomos de cada cadeia.

O A_c tem como parâmetros: os raios de van der Waals, o raio da probe e a distância entre os centros de dois átomos a_i e a_j , em cadeias diferentes (a_i no lado enzima e a_j no lado inibidor, no nosso caso). As mesmas considerações empregadas no cálculo da ASA por [Connolly (1983)] são observadas:

1. as cadeias nos complexos são tidas como um conjunto de átomos $[a_1, a_2, \dots, a_n]$;
2. cada átomo (a_i) é visto como uma aproximação esférica, dada pelo seu centro (x,y,z) e seu raio (r_i);
3. existe uma esfera imaginária, denominada *probe* (esfera teste), com raio próximo ao de uma molécula de solvente (p). Se esta for a água, na Literatura o raio mais usado é de 1.4Å.

A equação SR é apresentada a seguir:

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi(R_1 + R_2)d \left[1 + \left(\frac{R_1 - R_2}{d} \right)^2 \right] \quad (7.2)$$

Sendo:

$$R_1 = r_1 + p$$

$$R_2 = r_2 + p$$

r_1 = raio de van der Waals do 1º átomo

r_2 = raio de van der Waals do 2º átomo

p = raio da probe

d = distância entre o 1º e o 2º átomo

Assume-se $r_1 \geq r_2$.

Percebe-se que Ac e d são grandezas inversamente proporcionais. Seus limites em função dos raios R_1, R_2 podem ser dados pela tabela abaixo:

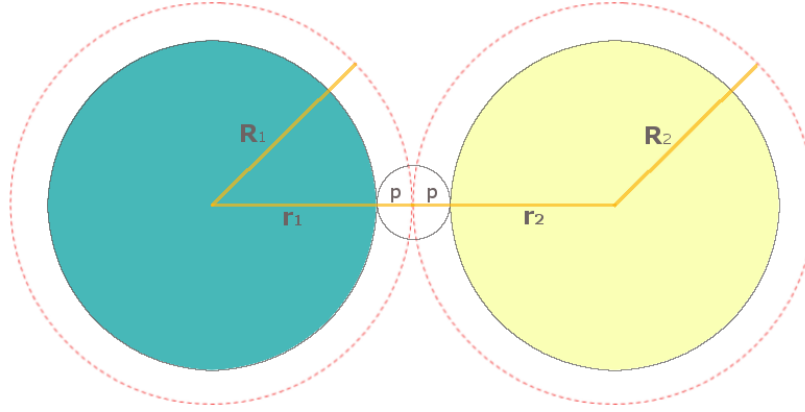
$0 \leq Ac \leq 4\pi R_2^2$ $R_1 + R_2 \geq d \geq R_1 - R_2$ para $R_1 > R_2$
--

É importante ter uma noção mais profunda do significado e das implicações desses limites. Servirá também para nos dar maior confiabilidade na veracidade termodinâmica e na consistência matemática da equação SR.

Para as análises a seguir iremos assumir que o raio da *probe* é constante envolvendo uma molécula de água ($p = 1.4\text{Å}$).

7.4.2.1 Limite inferior para Ac

Pela equação SR 7.2, nota-se que $Ac = 0$ quando $d = R_1 + R_2$. Na verdade, podemos considerar, sem perda de generalidade, que $Ac = 0$ para qualquer $d > R_1 + R_2$. Nessas distâncias, haveria espaço para que uma ou mais moléculas de água estivessem intervenientes entre dois átomos quaisquer. Como estamos interessados nesta tese nos contatos hidrofóbicos, o parâmetro Ac ganha um certo aporte termodinâmico. Se esse parâmetro vai a zero, podemos inferir que os átomos envolvidos suportam a condição de moléculas de águas ao seu redor. Mas, se forem átomos hidrofóbicos, torna-se improvável que isso ocorra. A tendência termodinâmica é que esses átomos encontrem-se a uma distância tal que não permita uma água interveniente. Essas distâncias serão exatamente aquelas em que $Ac > 0$. Logo, para o caso de átomos hidrofóbicos, o parâmetro Ac guarda relação com a intensidade do contato hidrofóbico.

Figura 7.3: Condição limite mínimo para A_c 

Isso pode ser demonstrado matematicamente. Dada que a distância $d = R_1 + R_2$, ao substituir d em (7.2), obtém-se o valor nulo para A_c , ou seja, $A_c = 0 \text{ \AA}^2$:

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi(R_1 + R_2)d \left[1 + \left(\frac{R_1 - R_2}{d} \right)^2 \right]$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi(R_1 + R_2)(R_1 + R_2) \left[1 + \frac{(R_1 - R_2)^2}{(R_1 + R_2)^2} \right]$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi(R_1 + R_2)^2 \left[\frac{(R_1 + R_2)^2 + (R_1 - R_2)^2}{(R_1 + R_2)^2} \right]$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi \left[(R_1 + R_2)^2 + (R_1 - R_2)^2 \right]$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi \left[R_1^2 + 2R_1R_2 + R_2^2 + R_1^2 - 2R_1R_2 + R_2^2 \right]$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - 2\pi(R_1^2 + R_2^2)$$

$$A_c(R_1, R_2, d) = 0$$

Como não poderia deixar de ser, o cálculo do A_c a partir da ASA 6.1, apresentada no capítulo 6, oferece o mesmo resultado:

$$A_c(R_1, R_2) = \Delta ASA = ASA_A + ASA_B - ASA_{AUB}$$

$$A_c(R_1, R_2) = 4\pi R_1^2 + 4\pi R_2^2 - (4\pi R_1^2 + 4\pi R_2^2) = 0$$

Logo, o limite inferior de A_c na equação SR é consistente, tanto matematicamente quanto termodinamicamente.

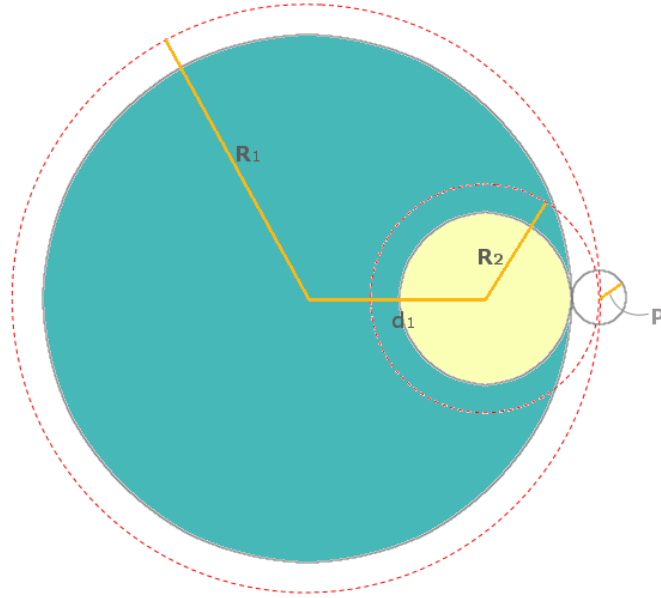
Na condição $A_c = 0$, qual o valor mínimo da distância inter-atômica d ? Pela figura 7.3, percebe-se que $d = r_1 + 2p + r_2$. Se considerarmos $r_1 = r_2 = r$ como sendo os raios de van der Waals de 2 carbonos tetraédricos (SP³) hidrofóbicos, temos $d = 2(r + p)$. Os

valores de r para esse tipo de carbono varia na literatura. [Bondi (1964)] considera $r = 1.7\text{\AA}$ e [Richards (1974)] $r = 2.0\text{\AA}$. Isso nos dá um $d = 6.2\text{\AA}$ e $d = 6.8\text{\AA}$, respectivamente. Trata-se de um valor consistente com outro trabalho desenvolvido por nosso grupo de pesquisa, publicado na *Proteins* em 2009 [da Silveira et al. (2009)]. Nesse artigo, foi encontrado o valor de distância ótima de $d = 7\text{\AA}$ para contatos interresíduos, independente de fatores como o tipo de cadeia (se todas betas, todas alfas ou alfa-betas); e do tipo de representação dos resíduos (se pelo centroide no carbono alfa CA ou no centro geométrico da cadeia lateral CG). Algo que intriga é a convergência de valores para análises feitas, não somente por diferentes métodos, mas também diferentes granulosidades: desta tese, ocorre no nível atômico; a do trabalho na *Proteins*, no nível de resíduo. O porquê dessa convergência ainda nos é obscuro, e foge ao escopo desta tese. Mas, já está anotado como algo a ser investigado em trabalhos futuros.

7.4.2.2 Limite superior para Ac

Viu-se na seção anterior que um $Ac > 0$ implica em distâncias interatômicas cada vez menores. Em $Ac = 0$ há possibilidade de uma molécula de água interveniente, e isso nos dá (para carbonos SP3) um d entre 6.2\AA e 6.8\AA . Na medida que essa distância é encurtada, a água é expulsa, podendo chegar a uma distância em que as nuvens de van der Waals dos carbonos se tocam, algo entre $d = 3.4\text{\AA}$ e $d = 4.0\text{\AA}$, o que implica em Ac variando de 54.5\AA^2 e 59.8\AA^2 . Distâncias ainda menores são permitidas, mas não sem interposição das nuvens e possível formação de ligações covalentes (não considerando os artefatos das técnicas de resolução estrutural). Por exemplo, para o caso de uma ligação covalente carbono-carbono como no etano, a distância esperada é de $d = 1.54\text{\AA}$, o que nos fornece um Ac entre 90.8\AA^2 e 112.4\AA^2 .

Se a distância pudesse ser encurtada ainda mais, chegaríamos na condição hipotética do átomo maior “engolfar” completamente o menor, algo obviamente irreal do ponto de vista químico. Mas, não matematicamente. Se fosse possível um átomo “assimilar” outro, é fácil perceber que a área de contato Ac deveria ser dada pela área da esfera do menor átomo e que a distância inter-atômica seria dada por $d = R_1 - R_2$. Podemos usar essa estranha hipótese para checar a consistência matemática da equação SR nesse limite surreal.

Figura 7.4: Condição limite máximo para A_c 

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi(R_1 + R_2)d \left[1 + \left(\frac{R_1 - R_2}{d} \right)^2 \right]$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - \pi(R_1 + R_2)(R_1 - R_2) \left[1 + \frac{(R_1 - R_2)^2}{(R_1 - R_2)^2} \right]$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - 2\pi(R_1 + R_2)(R_1 - R_2)$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2) - 2\pi(R_1^2 - R_2^2)$$

$$A_c(R_1, R_2, d) = 2\pi(R_1^2 + R_2^2 - R_1^2 + R_2^2)$$

$$A_c(R_1, R_2, d) = 4\pi R_2^2$$

Como esperado, converge para a área esférica do átomo menor. E o mesmo acontece com o A_c calculado a partir do ASA:

$$A_c(R_1, R_2) = \Delta ASA = ASA_A + ASA_B - ASA_{A \cup B}$$

$$A_c(R_1, R_2) = 4\pi R_1^2 + 4\pi R_2^2 - 4\pi R_1^2 = 4\pi R_2^2$$

7.5 Cálculo das distâncias e das matrizes de similaridades

Na equação SR, apresentada anteriormente, faz-se necessário determinar a distância entre dois átomos para cálculo da Ac . Computou-se a distância Euclidiana entre os átomos da enzima e do inibidor e selecionou-se somente aqueles átomos com distância inferior ou igual a 7\AA ($cutoff \leq 7\text{\AA}$). *Scripts* em Perl foram desenvolvidos para calcular a distância Euclidiana e gerar uma matriz de adjacências \mathbf{A} , para cada complexo, formada pelos resíduos da enzima e do inibidor que têm pelo menos um átomo com a distância inferior ao *cutoff* pré determinado.

Nessa matriz de adjacências \mathbf{A} , as colunas são formadas pelos átomos constituintes destes resíduos da enzima, seguidos com os do inibidor e as linhas obedecem ao mesmo critério. Assume-se que os valores da matriz são formados pela distância d_{ij} e que $d_{ij} \geq 0$ se o átomo a_i da enzima está em contato com o átomo a_j do inibidor, caso contrário $d_{ij} = 0$. Portanto, a matriz representa um grafo valorado ou ponderado, onde o peso corresponde a d_{ij} e os vértices são os átomos da enzima e do inibidor. Percebam que a matriz \mathbf{A} será quadrada e simétrica (vide figura 7.5), condição necessária às nossas decomposições espectrais.

Formalmente: dado o grafo $G = (V, E)$, onde V é o conjunto de vértices e E o conjunto de arestas, a matriz de adjacências será uma matriz \mathbf{A} de ordem $n \times n$, quadrada e simétrica, sendo que:

$n = E + I$, onde E é o número de átomos da enzima e I o número de átomos do inibidor (em contato com $d \leq 7\text{\AA}$);

$\mathbf{A}[ij] = d_{ij}$, onde $d_{ij} \geq 0$ e $d_{ij} \leq 7\text{\AA}$ quando há aresta de i a j e $d_{ij} = 0$ quando não há aresta de i a j .

Figura 7.5: Matriz de adjacências feita quadrada e simétrica para grafos bipartidos de contatos entre átomos (nós) enzima (ENZ) e inibidor (INB).

	nós ENZ	nós INB
nós ENZ	0	
nós INB		0

Notem pela forma como a matriz \mathbf{A} foi arranjada (figura 7.5), que ela representa um grafo bipartido. Os vértices em ENZ têm arestas zero com outros vértices em ENZ, assim como vértices em INB têm arestas zero com outros vértices INB. Apenas vértices entre ENZ e INB têm arestas, com pesos dados por Ac , diferentes de zero. Isso visa cumprir o objetivo de tentar mapear as correspondências hidrofóbicas entre enzimas e seus inibidores.

Para cada complexo, a partir dessa matriz de adjacências \mathbf{A} ponderada pela distância d_{ij} , foi construída, por meio de *scripts* em R, uma matriz de similaridades ou contato \mathbf{AC} , onde o peso corresponde à área de contato Ac calculada com base na equação de SR.

Deve-se mencionar que foi considerada somente $Ac > 2\text{\AA}^2$ como área de contato válida. Verificou-se empiricamente que esse era o erro médio do cálculo do Ac a partir do ASA (com *scripts* Pymol). Esse corte teve o objetivo também de ser mais conservativo no traço de uma aresta representando um contato hidrofóbico, evitando poluir o grafo com arestas pouco significativas.

Os valores dos raios de van der Waals (vdw) adotados nesta tese são os mesmos valores *default* do Pymol, que correspondem aos propostos por [Bondi (1964)], visto que há variações na literatura sobre esses valores (Tabela 7.7).

Tabela 7.7: Raios de van der Waals (R_{vdw})

Elemento	R_{vdw} (Å)	
	[Bondi (1964)]	[Richards (1974)]
C	1.70	2.00
O	1.52	1.40
N	1.55	2.00
S	1.80	1.80

7.6 Classificação dos átomos

Neste trabalho, o interesse é somente por interações atômicas não covalentes. Dado que as interações entre os átomos da enzima e do inibidor são consideradas quando o *cutoff* de distância é dado por $d \leq 7\text{\AA}$ e cada interação (aresta) é mensurada pela Ac , para tipificar os *clusters* em relação à hidrofobicidade, os átomos foram categorizados.

A classificação atômica, à semelhança de estudos realizados por [Sobolev et al. (1999)], permite categorizar as interações atômicas em apolares (hidrofóbicas) e polares com base no tipo de átomo. O método de [Sobolev et al. (1999)], desenvolvido para proteína-ligante, primeiramente categoriza os átomos em classes em função do comportamento eletrostático semelhante. Um tipo de contato entre dois átomos (a_i e a_j) é determinado a partir da classe de cada átomo e de restrições de distância experimentalmente definidas. Dessa forma, se ambos os átomos a_i e a_j são hidrofóbicos, o contato é potencialmente do tipo hidrofóbico, se ambos são hidrofílicos, o contato é considerado hidrofílico.

Neste trabalho, as análises foram concentradas somente nas interações entre os átomos que estão em condições hidrofóbicas (apolares), embora possam ser estendidas para condições polares, pois nossos algoritmos já dão suporte a isto. A motivação principal, para considerar as interações atômicas e não os resíduos em si, deve-se ao fato de que os resíduos têm porções (átomos) polares e apolares.

É comum em alguns trabalhos se considerar de forma generalizada todos os carbonos como átomos apolares [Sobolev et al. (1999)]. Sabe-se que isso não é de todo verdade. Carbonos ligados ou próximos a átomos polares, tendem-se a tornar mais polarizados também, por indução. Átomos mais eletronegativos, por exemplo, vão tender a atrair mais a nuvem eletrônica desses carbonos.

A tabela 7.8 mostra as 6 (seis) subclasses definidas para categorizar os átomos. Conforme experimentos descritos com mais detalhes na seção 8.2 e no intuito de compor uma classificação mais fidedigna com a realidade, os átomos foram agrupados nessas subclasses e posteriormente definidos no perfil predominante como polar ou apolar. Separamos os carbonos em 4 tipos: a , para aqueles que foram considerados tipicamente apolares; c , o carbono alfa do *backbone*; i o carbono planar (SP2) da carbonila; m , carbonos considerados "mesos", geralmente ligados diretamente a um grupo polar, e que podem, portanto, apresentar grande chance de estarem polarizados. A tabela 7.8 mostra quais carbonos foram considerados mesos, com base na nomenclatura PDB (exemplo: Q.CD indica o carbono delta de uma glutamina). Isso nos deu flexibilidade para investigar os efeitos de diferentes composições entre esses tipos de carbono.

Continuando, os nitrogênios e oxigênios foram categorizados em dois grupos: se eram do *backbone*, receberam o rótulo b ; senão, o rótulo p de polar. A intenção aqui

era poder fazer composições entre átomos do *backbone* e não-*backbone*, se necessário. Embora, tecnicamente, somente os enxofres nos agrupamentos tiois oxidados (em ponte dissulfeto) devessem ser classificados como apolares, todos foram incluídos no grupo *a*. Nas cadeias que formam nossas bases de dados, nenhuma ponte dissulfeto reduzida foi encontrada. Também o enxofre tioeter da metionina ficou em *a*. Todos os demais átomos não mencionados até o momento (como fósforo, selênio, metais, etc) ficaram categorizados em *o* de outros.

Na seção 8.2, o processo de definição dos tipos de átomos é apresentado detalhadamente.

Tabela 7.8: Classificação dos átomos - subclasses

Subclasse	Descrição atômica	Classe
a	Carbonos não meso (inclui os C aromáticos) + todos os enxofres (S)	Apolar
c	Carbono alpha (C_{α})	Polar
i	Carbono da carbonila	
m	C meso (anfipáticos): Q.CD, S.CB, N.CG, E.CD, D.CG, K.CE, T.CB, R.CZ, Y.CZ	
b	N e O do backbone	
p	N, O (não pertencentes ao backbone)	
o	Outros	

7.7 Agrupamento espectral

Conforme visto no capítulo 5, o agrupamento espectral fica melhor definido se usamos a Laplaciana normalizada L_{rw} , em função da sua associação com a minimização do Ncut. O Teorema de Rayleigh-Ritz nos autoriza teoricamente a usar a decomposição espectral de L_{rw} , usando a matriz de autovetores construída com os respectivos k menores autovalores para agrupar em k *cluster*. Como essa matriz espectral comporta um subespaço mais estruturado e menos convoluto, utilizamos k -medoides para rotular os k *cluster*, conforme algoritmo definido por [Shi and Malik (2000)].

Podemos então, definir a seguinte sequência de filtros e operações matriciais:

1. Geração da matriz de adjacências \mathbf{A} com base na distâncias Euclidianas interatômicas com corte (*cutoff*) em 7.0\AA . A matriz \mathbf{A} representa um grafo bipartido, onde são computadas apenas as distâncias entre átomos da enzima com átomos do inibidor.
2. Aplicação da equação SR para gerar a matriz de áreas contatos \mathbf{AC} .

3. Criação da matriz Laplaciana Normalizada \mathbf{ACL}_{rw} .
4. Decomposição espectral de $\mathbf{ACL}_{rw} = \mathbf{PSP}^{-1}$.
5. Seleção de uma submatriz \mathbf{P}_k contendo apenas os k últimos autovetores correspondentes aos k menores autovalores em \mathbf{S} .
6. Aplicação de k-medoides para rotular k grupos em \mathbf{P}_k .

7.7.1 Exemplo

Apresentamos nessa seção, um exemplo para ajudar a elucidar a sequência de passos e operações matriciais descritas anteriormente sobre o agrupamento espectral.

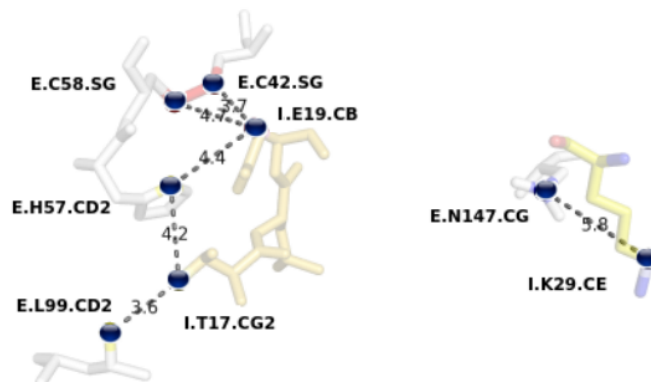
A partir do complexo 1PPF - formado entre uma tripsina (*Human Leukocyte Elastase*) e o inibidor Ovomucoide (OMTK3) - selecionamos um subconjunto de átomos da enzima e átomos da alça do inibidor. Na figura 7.6, é mostrado o modelo desse subconjunto de átomos, onde os átomos da enzima e do inibidor são representados pela notação “*Letra.Aminoacido.Atomo*”, a seguir descrita:

Letra: pode ser E ou I, abreviaturas para Enzima e Inibidor, respectivamente;

Aminoacido: corresponde ao código de uma letra do aminoácido associado ao número do aminoácido;

Atomo: código do átomo selecionado.

Figura 7.6: Representação simplificada do complexo 1PPF



7.7.1.1 Passo 1: Geração da matriz de distâncias

A figura 7.7 mostra o grafo correspondente ao modelo da figura 7.6. Observe que a distância euclidiana (d) entre os átomos da enzima e os átomos do inibidor é inferior ou igual a 7\AA ($d \leq 7\text{\AA}$). O grafo é representado pela matriz de distâncias entre os átomos, conforme abaixo (figura 7.8).

Figura 7.7: Grafo da representação simplificada do complexo 1PPF

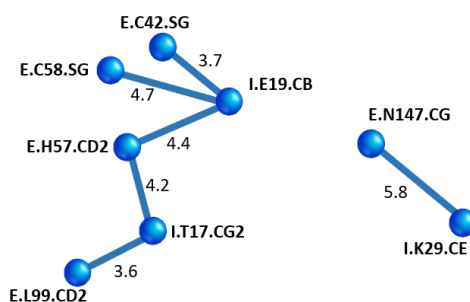


Figura 7.8: Matriz de distâncias do grafo da representação simplificada do complexo 1PPF

1PPF-ELY-OVO	E.CYS42.SG	E.HIS57.CD2	E.CYS58.SG	E.LEU99.CD2	E.ASN147.CG	I.THR17.CG2	I.GLU19.CB	I.LYS29.CE
E.CYS42.SG	0	0	0	0	0	0	0	3.7
E.HIS57.CD2	0	0	0	0	0	0	4.2	4.4
E.CYS58.SG	0	0	0	0	0	0	0	4.7
E.LEU99.CD2	0	0	0	0	0	0	3.6	0
E.ASN147.CG	0	0	0	0	0	0	0	0
I.THR17.CG2	0	4.2	0	3.6	0	0	0	0
I.GLU19.CB	3.7	4.4	4.7	0	0	0	0	0
I.LYS29.CE	0	0	0	0	5.8	0	0	0

A matriz de distâncias é quadrada e simétrica ($S \in \mathbb{R}^{n \times n}$), onde n é dado pelo número de átomos da enzima mais o número de átomos do inibidor. No exemplo, temos 5 átomos da enzima (E.CYS42.SG, E.HIS57.CD2, E.CYS58.SG, E.LEU99.CD2, E.ASN147.CG) e 3 átomos do inibidor (I.THR17.CG2, I.GLU19.CB, I.LYS29.CE). Logo $n = 8$, e a matriz de entrada é $M_{8 \times 8}$, cujos valores correspondem às distâncias euclidianas entre os átomos da enzima e do inibidor.

7.7.1.2 Passo 2: Geração da matriz de áreas de contatos AC

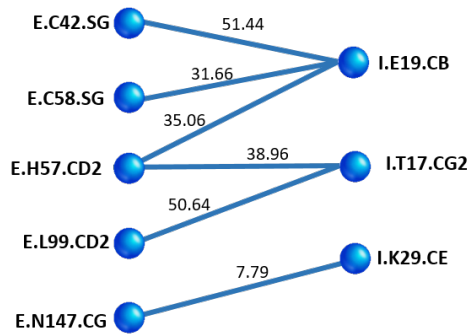
A partir da matriz de distâncias é gerada a matriz com as áreas de contato entre os átomos da enzima e do inibidor (matriz AC). Recordar-se que o cálculo da área de contato

foi realizado empregando-se o método SR (veja seção 7.4.2). A figura 7.9 é a matriz de contatos AC representativa do grafo bipartido da figura 7.10, cujos valores das arestas são os valores das áreas de contato entre os átomos da enzima e do inibidor.

Figura 7.9: Matriz de contatos AC representativa das áreas de contatos do modelo simplificado 1PPF

		"E.H57.CD2"	"E.L99.CD2"	"E.N147.CG"		"I.E19.CB"		
	"E.C42.SG"	"E.C58.SG"	"E.L99.CD2"	"E.N147.CG"	"I.T17.CG2"	"I.E19.CB"	"I.K29.CE"	
	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]	[, 7]	[, 8]
"E.C42.SG" [1,]	0.00	0.00	0.00	0.00	0.00	0.00	51.44	0.00
"E.H57.CD2" [2,]	0.00	0.00	0.00	0.00	0.00	38.96	35.06	0.00
"E.C58.SG" [3,]	0.00	0.00	0.00	0.00	0.00	0.00	31.66	0.00
"E.L99.CD2" [4,]	0.00	0.00	0.00	0.00	0.00	50.64	0.00	0.00
"E.N147.CG" [5,]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.79
"I.T17.CG2" [6,]	0.00	38.96	0.00	50.64	0.00	0.00	0.00	0.00
"I.E19.CB" [7,]	51.44	35.06	31.66	0.00	0.00	0.00	0.00	0.00
"I.K29.CE" [8,]	0.00	0.00	0.00	0.00	7.79	0.00	0.00	0.00

Figura 7.10: Grafo representativo da matriz de contatos AC



7.7.1.3 Passo 3: Criação da matriz Laplaciana Normalizada ACL_{rw}

A matriz Laplaciana Normalizada ACL_{rw} é obtida a partir da matriz Laplaciana e matriz diagonal D . Por sua vez, a matriz Laplaciana é dada pela fórmula $ACL = D - AC$, onde AC é a matriz de áreas de contatos calculada no passo 2.

Na matriz D , cada valor da diagonal corresponde ao grau do vértice que, por sua vez, é obtido pela soma dos valores da linha correspondente na matriz AC . Logo, o valor da segunda linha e segunda coluna d_{22} é igual à soma dos valores ac_{21} , ac_{22} , ..., ac_{28} da segunda linha da matriz AC , ou seja, $d_{22} = ac_{21} + ac_{22} + \dots + ac_{28}$. Por

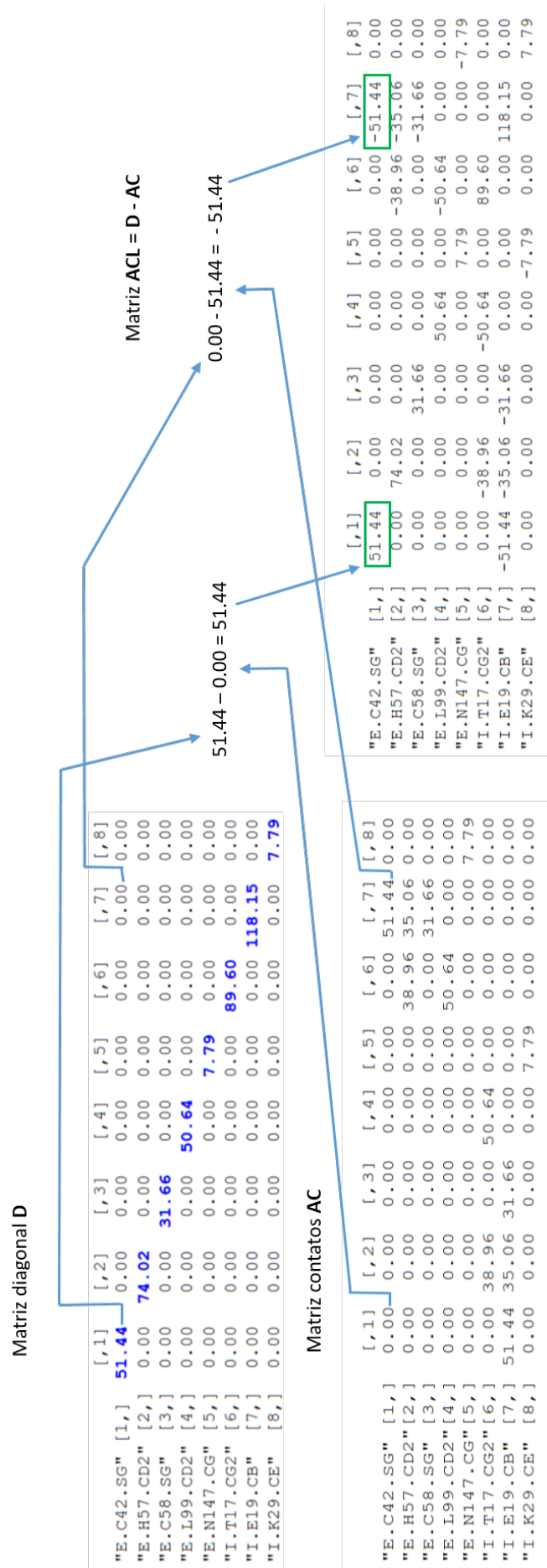
exemplo, a partir da segunda linha da matriz da AC (figura 7.9), ao somarmos os valores $0 + 0 + 0 + 0 + 0 + 38.96 + 35.06 + 0$, obtemos o total de 74.02 para o grau do vértice que ficará na segunda linha e segunda coluna da matriz diagonal D (veja a figura 7.11).

Figura 7.11: Matriz diagonal D calculada para o modelo simplificado 1PPF

"E.H57.CD2" [2,]	0.00	0.00	0.00	0.00	0.00	38.96	35.06	0.00	
						┌ + ┐			
						74.02			
		↙							
		[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]	[, 7]	[, 8]
"E.C42.SG" [1,]	51.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
"E.H57.CD2" [2,]	0.00	74.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
"E.C58.SG" [3,]	0.00	0.00	31.66	0.00	0.00	0.00	0.00	0.00	0.00
"E.L99.CD2" [4,]	0.00	0.00	0.00	50.64	0.00	0.00	0.00	0.00	0.00
"E.N147.CG" [5,]	0.00	0.00	0.00	0.00	7.79	0.00	0.00	0.00	0.00
"I.T17.CG2" [6,]	0.00	0.00	0.00	0.00	0.00	89.60	0.00	0.00	0.00
"I.E19.CB" [7,]	0.00	0.00	0.00	0.00	0.00	0.00	118.15	0.00	0.00
"I.K29.CE" [8,]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.79

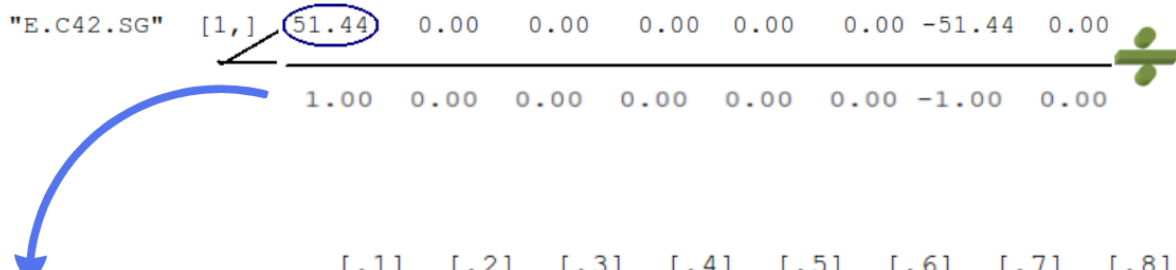
Uma vez calculadas a matriz AC e a matriz diagonal D , calcula-se a matriz Laplaciana ACL subtraindo-se de D a matriz AC : $ACL = D - AC$. Na figura 7.12, vemos o resultado dessa subtração. Note, por exemplo, que o elemento acl_{11} , ou seja, o elemento da primeira linha e primeira coluna da matriz ACL é obtido subtraindo-se 0 do valor 51.44, onde cada um corresponde, respectivamente, ao elemento ac_{11} da matriz AC e ao elemento d_{11} da matriz D .

Figura 7.12: Cálculo da matriz Laplaciana *ACL* do modelo simplificado 1PPF



Obtida a matriz laplaciana ACL , podemos gerar a matriz laplaciana normalizada $ACLrw$. Os valores dessa matriz são calculados dividindo-se cada linha da matriz laplaciana ACL pelo respectivo grau da linha que corresponde ao valor do elemento da diagonal ACL . Como podemos ver no exemplo abaixo, figura 7.13, o grau da primeira linha de ACL corresponde ao valor 51.44 do elemento acl_{11} da diagonal de ACL . Dividindo-se a primeira linha de ACL por 51.44, obtemos a primeira linha da matriz laplaciana normalizada $ACLrw$. Para obter a segunda linha de $ACLrw$, dividimos a segunda linha de ACL pelo grau 74.02. Logo temos, $aclr_{26} = -38.96/74.02 = -0.53$ e $aclr_{27} = -35.06/74.02 = -0.47$. O mesmo procedimento é realizado sucessivamente para as demais linhas da laplaciana normalizada.

Figura 7.13: Cálculo da matriz Laplaciana Normalizada $ACLrw$ do modelo simplificado 1PPF



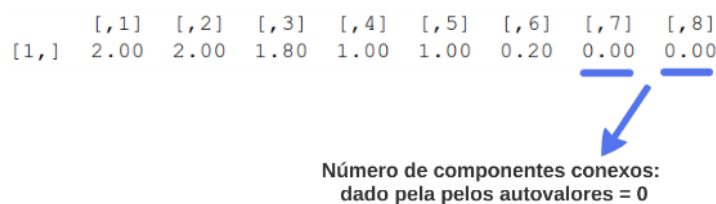
	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]	[, 7]	[, 8]
"E.C42.SG" [1,]	51.44	0.00	0.00	0.00	0.00	0.00	-51.44	0.00
	1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00
"E.C42.SG" [1,]	1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00
"E.H57.CD2" [2,]	0.00	1.00	0.00	0.00	0.00	-0.53	-0.47	0.00
"E.C58.SG" [3,]	0.00	0.00	1.00	0.00	0.00	0.00	-1.00	0.00
"E.L99.CD2" [4,]	0.00	0.00	0.00	1.00	0.00	-1.00	0.00	0.00
"E.N147.CG" [5,]	0.00	0.00	0.00	0.00	1.00	0.00	0.00	-1.00
"I.T17.CG2" [6,]	0.00	-0.43	0.00	-0.57	0.00	1.00	0.00	0.00
"I.E19.CB" [7,]	-0.44	-0.30	-0.27	0.00	0.00	0.00	1.00	0.00
"I.K29.CE" [8,]	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	1.00

7.7.1.4 Passo 4: Decomposição espectral da matriz Laplaciana Normalizada $ACLrw$

Na decomposição espectral, procuramos encontrar os autovalores e autovetores associados à matriz laplaciana normalizada $ACLrw$. Na figura 7.14 são mostrados os autovalores calculados à partir de $ACLrw$.

A indicação da quantidade de componentes conexos de um grafo é uma propriedade fundamental dos autovalores. Essa quantidade é dada pelo número de zeros presentes, ou seja, pela sua multiplicidade. Logo, podemos inferir que para o nosso exemplo, há dois componentes conexos (colunas 7 e 8 da figura 7.14).

Figura 7.14: Autovalores da matriz Laplaciana Normalizada ACL_{rw} do modelo simplificado 1PPF

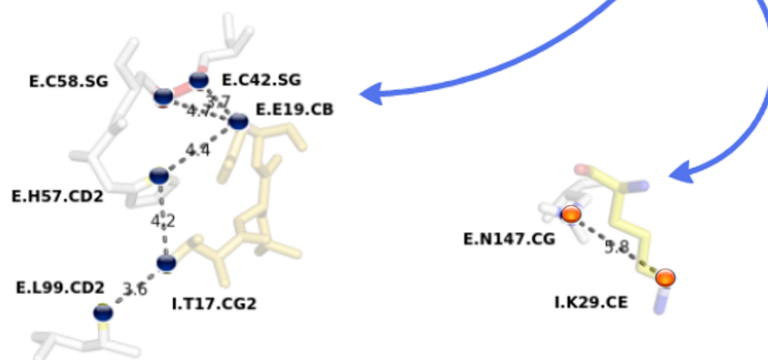


Para identificarmos quais são esses dois componentes conexos, é necessário calcular os autovetores. Estes, associados aos autovalores, são vetores indicativos dos componentes conexos. Os autovetores do nosso exemplo são mostrados na figura abaixo [7.15](#).

Como sabemos quais são os autovetores correspondentes aos dois componentes conexos? Para isso, necessitamos descobrir qual é o autovetor associado a cada autovalor 0. Olhando a matriz de autovetores (figura [7.15](#)), o autovetor é dado pela coluna correspondente à coluna do autovalor 0. Então, para o autovalor 0 da coluna 8 (figura [7.14](#)), temos o autovetor na matriz de autovalores na coluna 8 ([7.15](#)), dado pelos valores 0.41, 0.41, 0.41, 0, 0.41, 0.41, 0. Observa-se que há dois valores repetidos: 0 e 0.41. Esses valores indicam que há dois componentes conexos e mostram quais são os vértices pertencentes a cada componente conexo. Quais são os átomos relacionados à esses valores? Associados ao valor 0, temos os átomos *CG* da asparagina 147 e *CE* da lisina 29 formando um grupo (veja na figura [7.15](#) abaixo). Formando outro subgrupo, associado ao valor 0.41, temos os átomos *SG* da cistina 42, *CD2* da histidina 57, *SG* da cistina 58, *CD2* da leucina 99, *CG2* da treonina 17 e *CB* do glutamato.

Figura 7.15: Matriz de autovetores obtida a partir da matriz Laplaciana Normalizada ACL_{rw} do modelo simplificado 1PPF

		[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]	[, 7]	[, 8]
"E.C42.SG"	[1,]	-0.03	0.41	0.43	0.61	0.16	0.43	0.26	0.41
"E.H57.CD2"	[2,]	-0.03	0.41	-0.09	-0.24	-0.68	-0.09	0.26	0.41
"E.C58.SG"	[3,]	-0.03	0.41	0.43	-0.73	0.50	0.43	0.26	0.41
"E.L99.CD2"	[4,]	-0.03	0.41	-0.56	0.19	0.52	-0.56	0.26	0.41
"E.N147.CG"	[5,]	0.70	0.00	0.00	0.00	0.00	0.00	0.54	0.00
"I.T17.CG2"	[6,]	0.03	-0.41	0.45	0.00	0.00	-0.45	0.26	0.41
"I.E19.CB"	[7,]	0.03	-0.41	-0.34	0.00	0.00	0.34	0.26	0.41
"I.K29.CE"	[8,]	-0.70	0.00	0.00	0.00	0.00	0.00	0.54	0.00



7.7.1.5 Passo 5: Seleção de uma submatriz P_k contendo apenas os k últimos autovetores correspondentes aos k menores autovalores

Para aplicarmos um algoritmo de agrupamento (PAM) e obtermos esses grupos, selecionamos uma submatriz P_k da matriz de autovetores. Para um agrupamento de 3 grupos ($k = 3$), por exemplo, temos a matriz apresentada na figura 7.16, obtida das três últimas colunas da matriz de autovetores da figura 7.15.

Figura 7.16: Seleção da submatriz P_k à partir da matriz de autovetores

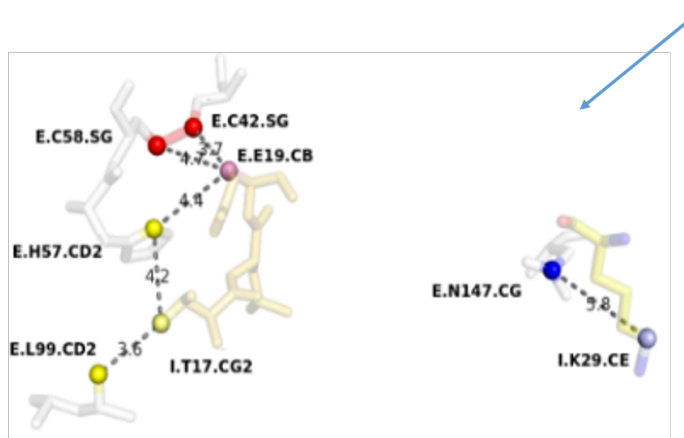
		[, 6]	[, 7]	[, 8]
"E.C42.SG"	[1,]	0.43	0.26	0.41
"E.H57.CD2"	[2,]	-0.09	0.26	0.41
"E.C58.SG"	[3,]	0.43	0.26	0.41
"E.L99.CD2"	[4,]	-0.56	0.26	0.41
"E.N147.CG"	[5,]	0.00	0.54	0.00
"I.T17.CG2"	[6,]	-0.45	0.26	0.41
"I.E19.CB"	[7,]	0.34	0.26	0.41
"I.K29.CE"	[8,]	0.00	0.54	0.00

7.7.1.6 Passo 6: Aplicação de k -medoides para rotular k grupos em P_k

Ao aplicarmos o algoritmo de agrupamento k -medoides em P_k , obtemos os grupos rotulados. Isso pode ser visto no exemplo abaixo (figura 7.17). Ao aplicarmos o k -medoides (PAM) na matriz que contém os três autovetores ($k = 3$), temos como resultado do agrupamento uma outra matriz onde os átomos estão na coluna 1 e na coluna 2 temos o número do grupo a que pertence cada átomo.

Figura 7.17: Processo de obtenção dos grupos com aplicação do k -medoides

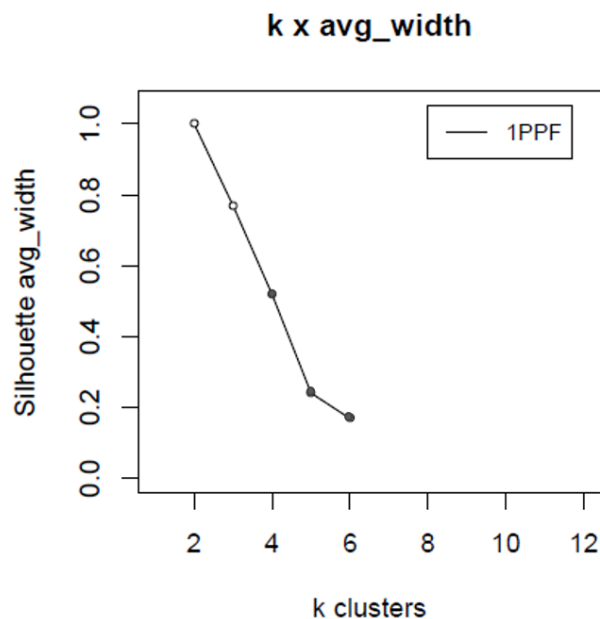
		[, 6]	[, 7]	[, 8]		[, 1]	[, 2]
"E.C42.SG"	[1,]	0.43	0.26	0.41	→	[1,] "E.C42.SG"	"1"
"E.H57.CD2"	[2,]	-0.09	0.26	0.41		[2,] "E.H57.CD2"	"2"
"E.C58.SG"	[3,]	0.43	0.26	0.41		[3,] "E.C58.SG"	"1"
"E.L99.CD2"	[4,]	-0.56	0.26	0.41		[4,] "E.L99.CD2"	"2"
"E.N147.CG"	[5,]	0.00	0.54	0.00		[5,] "E.N147.CG"	"3"
"I.T17.CG2"	[6,]	-0.45	0.26	0.41		[6,] "I.T17.CG2"	"2"
"I.E19.CB"	[7,]	0.34	0.26	0.41		[7,] "I.E19.CB"	"1"
"I.K29.CE"	[8,]	0.00	0.54	0.00		[8,] "I.K29.CE"	"3"



Esse resultado do k -medoides em formato de matriz, refere-se ao modelo estrutural também presente na figura 7.17. Os átomos $E.C42.SG$, $E.C58.SG$ e $E.E19.CB$ pertencem ao grupo 1 (em vermelho). O grupo 2 (amarelo) é formado pelos átomos $E.H57.CD2$, $E.L99.CD2$ e $I.T17.CG2$. Enquanto o grupo 3 (azul) contém os átomos $E.N147.CG$ e $I.K29.CE$.

Para avaliar a qualidade desses agrupamentos formados foi utilizado o coeficiente de silhueta. Observa-se na figura 7.18, que os melhores grupos formados para o nosso exemplo, são obtidos para $k = 2$ e $k = 3$.

Figura 7.18: Gráfico com coeficientes de silhueta - grupos do modelo simplificado 1PPF



Quando $k = 2$ temos o valor ótimo para o coeficiente de silhueta, que é igual a 1. Os grupos formados nessa situação são visualizado na figura 7.19. Quando $k = 3$, o valor aproximado de 0.7 para o coeficiente de silhueta indica que o grupo formado tem uma boa estrutura. Esses três grupos formados são visualizado na figura 7.20.

Figura 7.19: Agrupamento com dois grupos para o modelo simplificado 1PPF

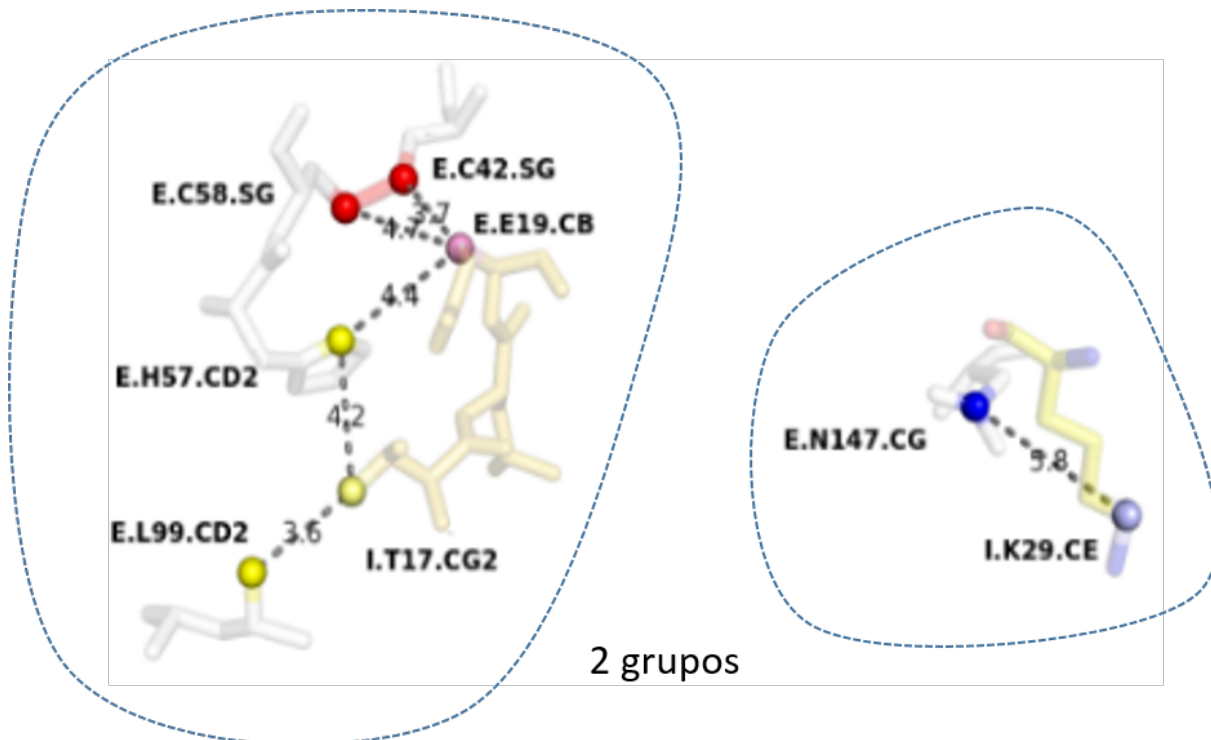
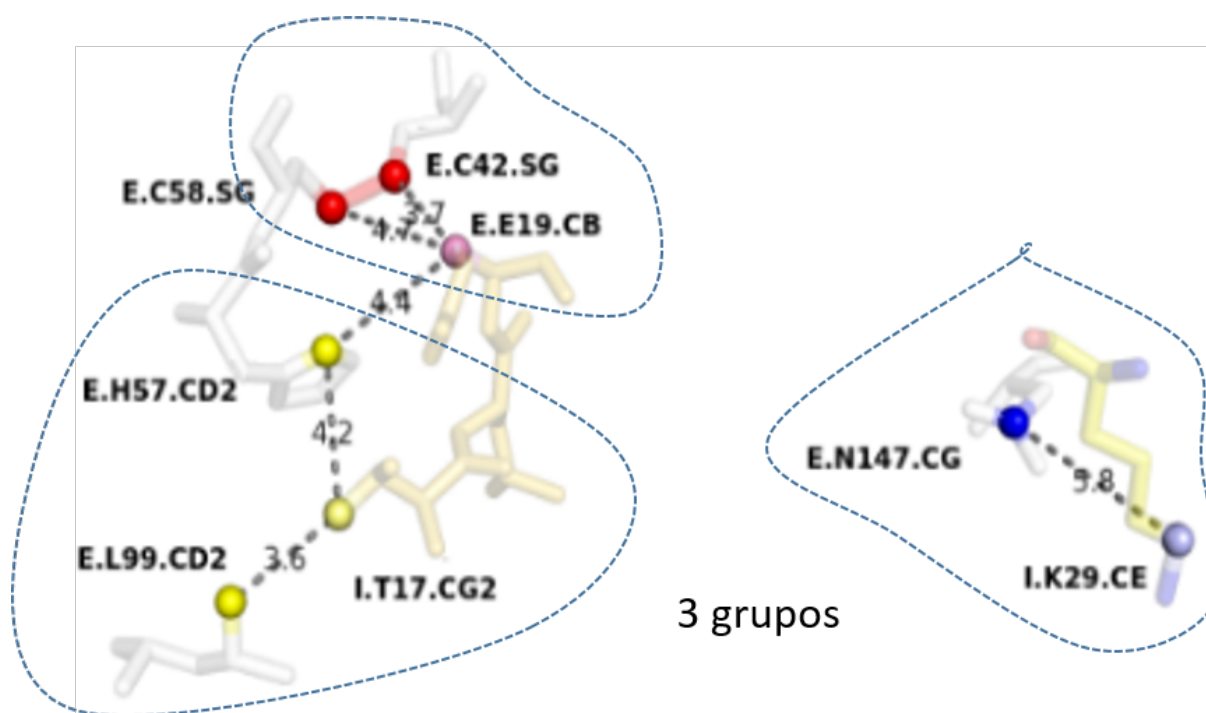


Figura 7.20: Agrupamento com três grupos para o modelo simplificado 1PPF



Capítulo 8

Resultados e discussão

Neste capítulo, são apresentados e discutidos os resultados da aplicação do agrupamento espectral para identificar sítios hidrofóbicos em complexos serino peptidases e inibidores proteicos e a correspondência entre os sítios do lado da enzima e do lado do inibidor. As enzimas dos complexos selecionados são do Tipo Tripsina (tripsina, quimotripsina, elastase, proteinase A e B, Fator Xa) e Tipo Subtilisina (subtilisinas BPN e Carlsberg e Thermitase) interagindo com os inibidores, compreendendo ovomucóide (OMTKY3 e OMTKY2), Eglina C, Ecotina, BPTI, SLPI, Serpina, Elafin, Potato Inhibitor, BBI, MCTI-A e CMTI.

Primeiramente são apresentados gráficos de distribuição da área de contato entre enzima/inibidor calculada para os complexos em termos das características polares e apolares. Em seguida são apresentados os resultados e discussão sobre os mapeamentos das regiões hidrofóbicas de cada complexo.

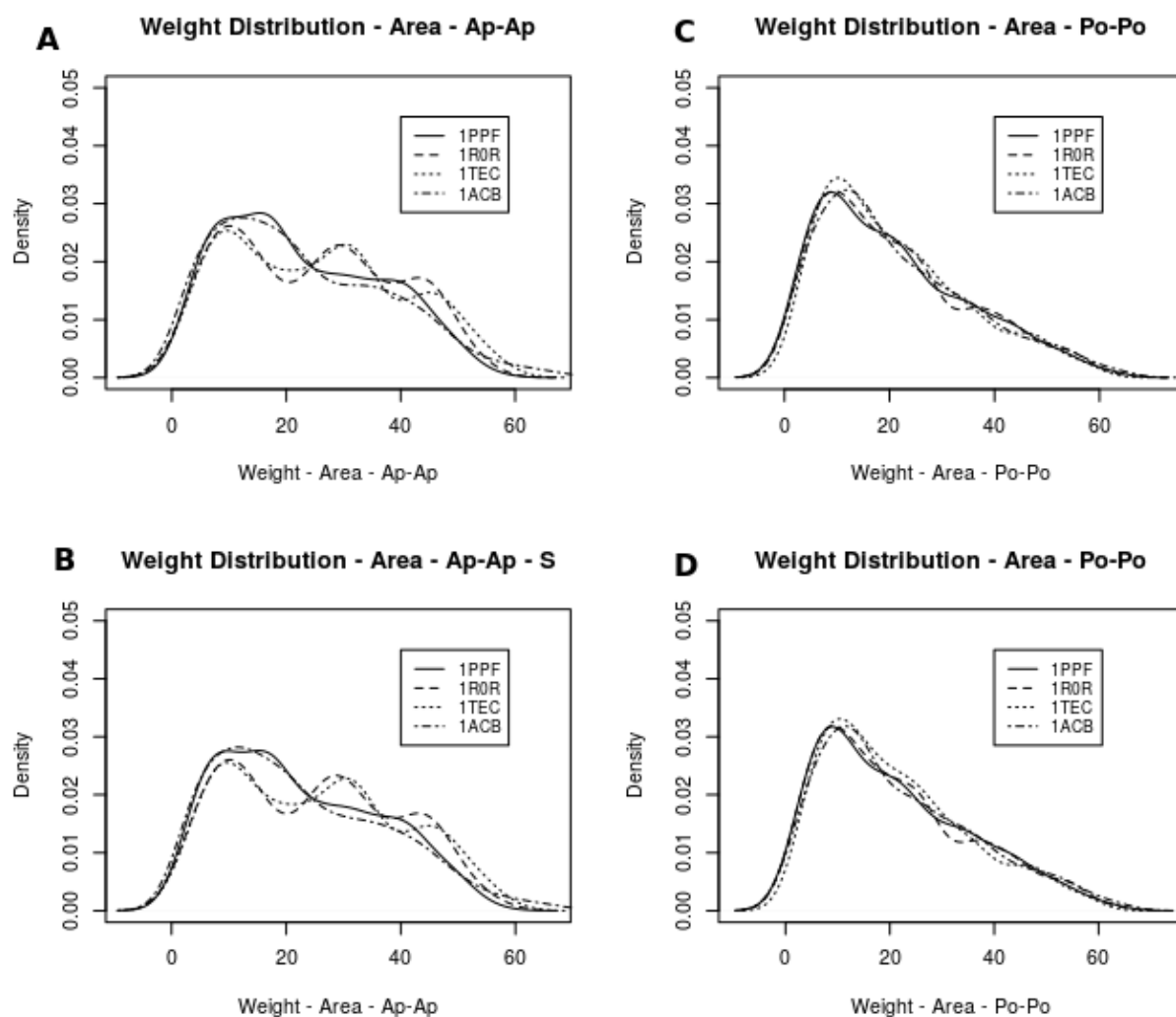
8.1 Análises da Inibição Cruzada

Uma das primeiras ações desta tese, no sentido de dar continuidade ao trabalho publicado na Bioinformatics [[Gonçalves-Almeida et al. \(2012\)](#)], foi uma análise pormenorizada da inibição cruzada, envolvendo: uma tipo tripsina com inibidor ovomucoide (1PPF:TRY-OVO), uma tipo subtilisina com inibidor ovomucoide (1ROR:SBN-OVO), uma tipo tripsina com inibidor eglina (1ACB:TRY-EGL), uma tipo subtilisina com inibidor eglina (1TEC:SBN-EGL). Logo, temos duas classes de peptidases diferentes inibidas por dois inibidores também diferentes numa relação cruzada 2 para 2. Notem que estamos trabalhando não somente com duas enzimas estruturalmente distintas (estão em clãs diferentes na classificação MEROPS), mas também dois inibidores estruturalmente distintos (também em clãs diferentes na MEROPS). Vide tabelas [7.3](#) e [7.4](#).

Um primeiro estudo feito foi das distribuições PDF (*Probability Distribution Functions*) para esses 4 complexos. Para essas análises, foram considerados APOLARES todos

os carbonos, exceto aqueles do backbone. Todos os demais átomos foram tratados como POLARES. O perfil dessa primeira distribuição é visto na figura 8.1 (A) e (B).

Figura 8.1: Distribuição de densidade para as áreas de contatos dos complexos 1PPF, 1ACB, 1R0R e 1TEC em função da classificação dos átomos. (A) e (B): perfil **apolar** para o experimento inicial e o final, respectivamente. (C) e (D): perfil **polar** para o experimento inicial e o final, respectivamente.



Um primeiro fato relevante está em perceber que os valores máximos de A_c em torno de 60\AA^2 são condizentes com os valores estimados pela equação SR para nuvens de van der Waals de carbonos tetraédricos (SP3) em contato, conforme relatado na seção 7.4.2. Mas, algo bem mais curioso se sobressai numa análise visual: o perfil da distribuição POLAR não distingue os 4 complexos, havendo grande similaridade. Isso não se constata para o perfil APOLAR: 1PPF:TRY-OVO que tende a se assemelhar com 1ACB:TRY-EGL, e 1ROR:SBN-OVO a se emparelhar com 1TEC:SBN-EGL.

E a análise estatística corrobora essa percepção visual. Se assumirmos que os dados foram amostrados de PDF contínuas, podemos aplicar o teste de homogeneidade das distribuições de *Kolmogorov-Smirnov* (*KS-Test*), com a hipótese nula H_0 aferindo se

Tabela 8.1: Perfil Apolar

Experimento		Área contato: p-value para ks-test				
ID	Descrição					
1	Experimento inicial: apolar (a): inclusão de todos os carbonos (C), exceto CA e C (da carbonila)	1PPF	1R0R	1TEC	1ACB	
		1PPF	1.00	0.11	0.16	0.83
		1R0R	0.11	1.00	0.97	0.03
		1TEC	0.16	0.97	1.00	0.07
2	Experimento final: apolar (a): todos os carbonos não meso + todos os enxofres (S). Carbonos meso (m)	1ACB	0.83	0.03	0.07	1.00
		1PPF	1R0R	1TEC	1ACB	
		1PPF	1.00	0.01	0.04	0.87
		1R0R	0.01	1.00	0.97	0.00
1TEC	0.04	0.97	1.00	0.01		
1ACB	0.87	0.00	0.01	1.00		

Tabela 8.2: Perfil Polar

Experimento		Área contato: p-value para ks-test				
ID	Descrição					
1	Experimento inicial: polar (p): todos os polares (N e O) + C da carbonila e CA (i)	1PPF	1R0R	1TEC	1ACB	
		1PPF	1.00	0.98	0.94	0.89
		1R0R	0.98	1.00	0.90	0.97
		1TEC	0.94	0.90	1.00	0.95
2	Experimento final: polar (p): todos os não apolares	1ACB	0.89	0.97	0.95	1.00
		1PPF	1R0R	1TEC	1ACB	
		1PPF	1.00	0.85	0.91	1.00
		1R0R	0.85	1.00	0.92	0.77
1TEC	0.91	0.92	1.00	0.86		
1ACB	1.00	0.77	0.86	1.00		

as amostragens produzem oscilações estatísticas aceitáveis de uma mesma distribuição. O resultado do KS pode ser visto nas tabelas 8.1 e 8.2, ID=1. Vemos que as diferenças não são significativas (p -value alto) para os pares de curvas que nosso olho considera como assemelhadas.

A conclusão possível que se pode chegar é que os contatos POLARES não são discriminativos para as distribuições de Ac para esses 4 pares peptidase-inibidor em inibição cruzada. Mas, os contatos APOLARES são. E o são de tal forma que é o lado ENZIMA que impõe o seu perfil hidrofóbico como discriminador. Em outras palavras, os inibidores, mesmo tão diferentes quanto ovomucoide e eglina, são obrigados a se adaptar à superfície hidrofóbica das peptidases.

E mais: o fato de estarmos constatando que as distribuições POLARES são indiferenciadas podem implicar que a quantidade e a intensidade dos contatos polares sob a métrica Ac são muito semelhantes nos 4 complexos. Algo que faz certo sentido, dado que a maior parte dos contatos POLARES envolverão pontes de hidrogênio e pontes salinas, cujas distâncias ótimas de contato tendem a variar pouco. O inusitado aqui seria constatar que nos 4 complexos presenciamos uma quantidade média similar desses contatos,

de tal forma que as distribuições são similares. O mesmo não ocorre com os contatos APOLARES. Cada grupo de enzimas (pelo menos, nesses 4 complexos avaliados) parece ter sua especificidade.

Do ponto de vista termodinâmico, isso pode indicar que fatores entrópicos (aqueles ligados à hidrofobicidade) teriam certa primazia sobre os fatores entálpicos na diferenciação dos valores do ΔG de *binding*. Infelizmente, não tivemos tempo para confirmar essa hipótese nesta tese, e será algo que pretendemos avaliar em trabalhos futuros. Seja como for, a dissecação energética feita por [Baker and Murphy (1997)] (e discutida na Introdução) parece ir nessa direção também.

As figuras 8.96 e 8.100 mostram os *subclusters* desses 4 complexos no contexto estrutural em imagens construídas a partir do Pymol.

Há um último ponto a destacar. Apesar de ovomucoide (OVO) e eglina (EGL) serem estruturalmente tão diferentes, em suas alças inibitórias elas apresentam um padrão marcante:

SUBCLUSTER	[1]	[2]	[3]	[4]	[5]	[6]	[7]
1PPF:	TRY-OVO: [PA]	[T]	[CL]	[E]	[R]	[Y]	[]
1ACB:	TRY-EGL: [P]	[T]	[VL]	[D]	[R]	[LY]	[]
1ROR:	SBN-OVO: [PA]	[T]	[L]	[]	[R]	[Y]	[]
1TEC:	SBN-EGL: [P]	[T]	[L]	[D]	[R]	[L]	[Y]

Algumas pré-conclusões notáveis:

- 1º SUBCLUSTER: OVO tem uma *A* entre *P* e *T*, mas sua inserção ainda a mantém no mesmo subcluster. Se olharmos nas figuras 8.96 e 8.100 constatamos que os contatos feitos por *PA* em OVO têm equivalência com *P* em EGL;
- 2º SUBCLUSTER: é *T* em ambas, e a grande contribuição hidrofóbica vem do seu carbono gama (CG2), aquilo que *T* tem a mais em relação à *S*;
- 3º SUBCLUSTER: há presença de uma *L* que se encaixa no bolsão de especificidade, mas nas tipo tripsinas, os inibidores OVO e EGL oferecem a mais, contatos hidrofóbicos de um *C* e um *V*, respectivamente.
- 4º e 5º SUBCLUSTERES: é marcante a presença de resíduos carregados aqui, sendo negativos no primeiro (*D* ou *E*) e positivos no segundo (*R*). Mas, eles também oferecem contatos hidrofóbicos através de seus grupos metis antecedendo os carboxilatos e aminos. Em 1R0R, há a presença de um *E* negativo no 4º subcluster, mas ele não foi flagrado fazendo contatos hidrofóbicos no cristal, o que não quer dizer que isso não possa acontecer se a 1ROR estiver livre no solvente.
- 6º SUBCLUSTER: temos uma situação parecida com o que analisamos no 1º subcluster, com um *LY* nas EGL, mas apenas *Y* nas OVO, mas ambos fazem contatos

equivalentes. Na 1TEC, um Y também está presente, mas ele faz contatos num subcluster separado (não mostrado nas figuras).

- GERAL: de forma geral, as redes de contatos hidrofóbicos em tipo tripsinas tendem a ser mais complexas e numerosas que em tipo subtilisinas.

Vemos, portanto, que na estranha álgebra biológica, $PA = P$ e $Y = LY$. O que importa é que ambos tenham no cômputo final intensidade de interações equivalentes no nível atômico. Isso não deixa, de certa forma, de justificar os “*gaps*” que são usados nos alinhamentos de sequências. Mas, também revela que análises em granulosidade no nível dos resíduos podem ser imprecisas, dado que as interações operam no nível atômico. Voltaremos a discutir isso mais adiante.

8.2 Definição dos tipos de átomos

A análise desses 4 complexos em inibição cruzada também serviu para nos auxiliar a encontrar a melhor partição de tipos de átomos POLAR e APOLAR. Isso pode ser aferido nas tabelas 8.1 e 8.2. A metodologia que utilizamos foi tentar uma partição que maximizasse os valores de *p-values* entre os pares (1PPF:TRY-OVO, 1ACB:TRY-EGL) e (1ROR:SBN-OVO, 1TEC:SBN-EGL) e minimizasse (1PPF:TRY-OVO, 1ROR:SBN-OVO) e (1ACB:TRY-EGL, 1TEC:SBN-EGL).

Ajudou-nos também a selecionar quais carbonos considerar como MESO, dentro daquela ideia de vizinhança POLAR. Acabaram definidos como: Q.CD, S.CB, N.CG, E.CD, D.CG, K.CE. T.CB, R.CZ, Y.CZ. Exceção aos carbonos da histidina, que não foram considerados MESO, mesmo estando ligados a átomos POLARES. Se isso fosse feito, apenas o carbono beta seria considerado APOLAR. A histidina é um resíduo versátil, sendo bem aceita tanto na superfície (em contato com a água) quanto no interior das proteínas (ao lado de outros hidrofóbicos).

Dentre as classificações apresentadas na tabela 7.8, após inúmeros testes, a “melhor” definição do que considerar APOLAR ou POLAR pode ser vista na tabela 8.3. Melhor no sentido dos objetivos perseguidos por esta tese, de ressaltar as correspondências hidrofóbicas entre enzima e inibidor. Para isso, procuramos uma definição mais estrita de APOLAR, colocando nesse conjunto apenas os átomos com alta chance de serem realmente apolares.

Seja como for, a análise comparativa dos *p-values* nas tabelas 8.1 e 8.2 e dos traçados das distribuições PDF nos gráficos das figuras 8.1 (A a D) revela que há uma

Tabela 8.3: Classificação dos átomos - classes

Classe atômica	Subclasses
Apolares	a
Polares	b, c, i, m, p

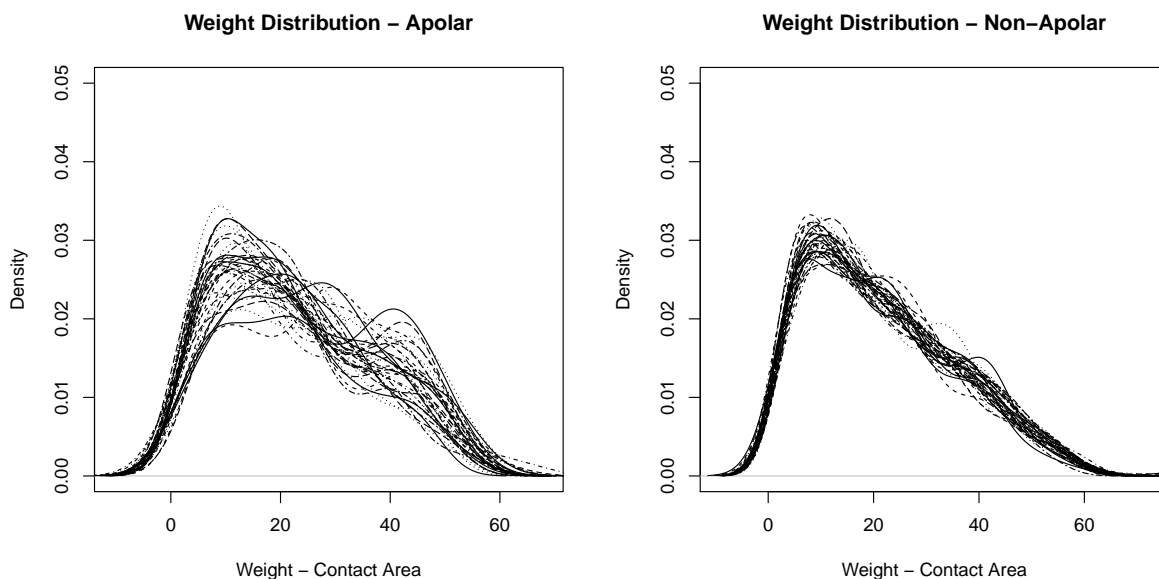
certa robustez em se usar uma definição mais abrangente de APOLAR (incluindo os MESOS) ou não (sem os MESOS).

8.3 Distribuição estendida para os 36 complexos

A mesma tendência da distribuição das áreas de contato (Ac) enzima/inibidor se mantém para os demais 36 complexos da base de dados. A Figura 8.2 mostra separadamente os perfis APOLAR e POLAR. O perfil POLAR mantém-se ainda homogêneo, mas vemos uma oscilação maior para alguns casos. O perfil APOLAR continua com seu poder discriminatório, mas não fica tão evidente a separação em dois grandes subgrupos (um para tipo tripsina e outro para tipo subtilisina) conforme havíamos previsto com o estudo da inibição cruzada com os 4 complexos feita acima.

Uma análise mais pormenorizada de quais seriam esses subgrupos ficou para trabalhos futuros a esta tese. Em função do tempo escasso, preferimos avançar no estudo dos padrões do agrupamento espectral nos 36 complexos, até porque eles poderiam nos dar *insights* sobre como melhor explicar a complexidade das distribuições de Ac polares e apolares evidenciados na figura 8.2.

Figura 8.2: Gráficos de distribuição das áreas de contato apolar e polar (36 complexos).



8.4 Regiões hidrofóbicas em cada complexo

Nesta seção, serão apresentadas as regiões hidrofóbicas complementares enzima/inibidor encontradas em cada complexo da base de dados estudada, com ênfase para o lado do inibidor, visto que já foi mostrado em [Gonçalves-Almeida et al. \(2012\)](#) a existência das regiões hidrofóbicas conservadas no lado enzima. Em alguns casos, essas regiões dos complexos são comparadas com as regiões do complexo de referência 1PPF e em outros com complexos que tenham o mesmo inibidor ou que seja da mesma família.

Três tipos de imagens ajudam a entender cada complexo. A primeira, é uma imagem da interface feita no Pymol, a partir das coordenadas atômicas da estrutura resolvida depositada no PDB. Na visão 3D a enzima é representada no formato "surface" e parte do inibidor (alça inibitória e outros resíduos) em *sticks*. Um padrão de cores identifica cada subcluster no lado enzima, acompanhado também de um id numérico.

A segunda é constituída por um gráfico com o coeficiente médio de Silhueta dos subclusters. Na ordenada, temos esse coeficiente médio de silhueta, e na abscissa o número k de subclusters. Pontos em branco indicam que nenhum subcluster teve silhueta negativo; pontos em cinza, que houve subclusters em que a silhueta de algum ponto ficou perto de zero (entre -0.5 e $+0.5$); pontos em preto, que algum ponto teve silhueta negativo. Lembrando que uma silhueta perto de zero indica que há dois ou mais subclusters disputando aquele ponto (provavelmente um *outlier*). Uma silhueta negativa, que aquele ponto poderia ter sido melhor alocado a outro subcluster.

A fim de facilitar o entendimento dos padrões, foi montando modelos de visuali-

zação mais simplificados que os das estruturas tridimensionais no Pymol, como projeções 2D das interfaces, onde cada subcluster é representado por elipses, e as letras indicam os resíduos envolvidos, tanto do lado enzima, quanto inibidor. Esta é a terceira imagem. Neste modelo, é calculado o centro geométrico dos átomos do resíduo envolvido em cada subcluster, e projetado em 2D no plano $z=0$; ou seja, das coordenadas (x,y,z) , considera-se $(x,y,0)$. A elipsoide mínima que cerca todos esses átomos projetados em cada subcluster é traçada, usando a função do R *ellipsoidhull*. O padrão de cores aqui é o seguinte: letras em **vermelho negrito** indicam resíduos do lado enzima; letras em **azul negrito** resíduos da alça inibitória; letras em **azul simples**, resíduos do inibidor, mas que não são da alça. São considerados resíduos da alça inibitória aqueles contíguos na sequência que se alojam ou assentam na região da interface da enzima, formando o que a literatura costuma chamar de *loop*. A imagem do modelo também mostra, no subtítulo: o silhueta médio, o número k de subclusters e o código de 3 letras que identifica a peptidase e o inibidor.

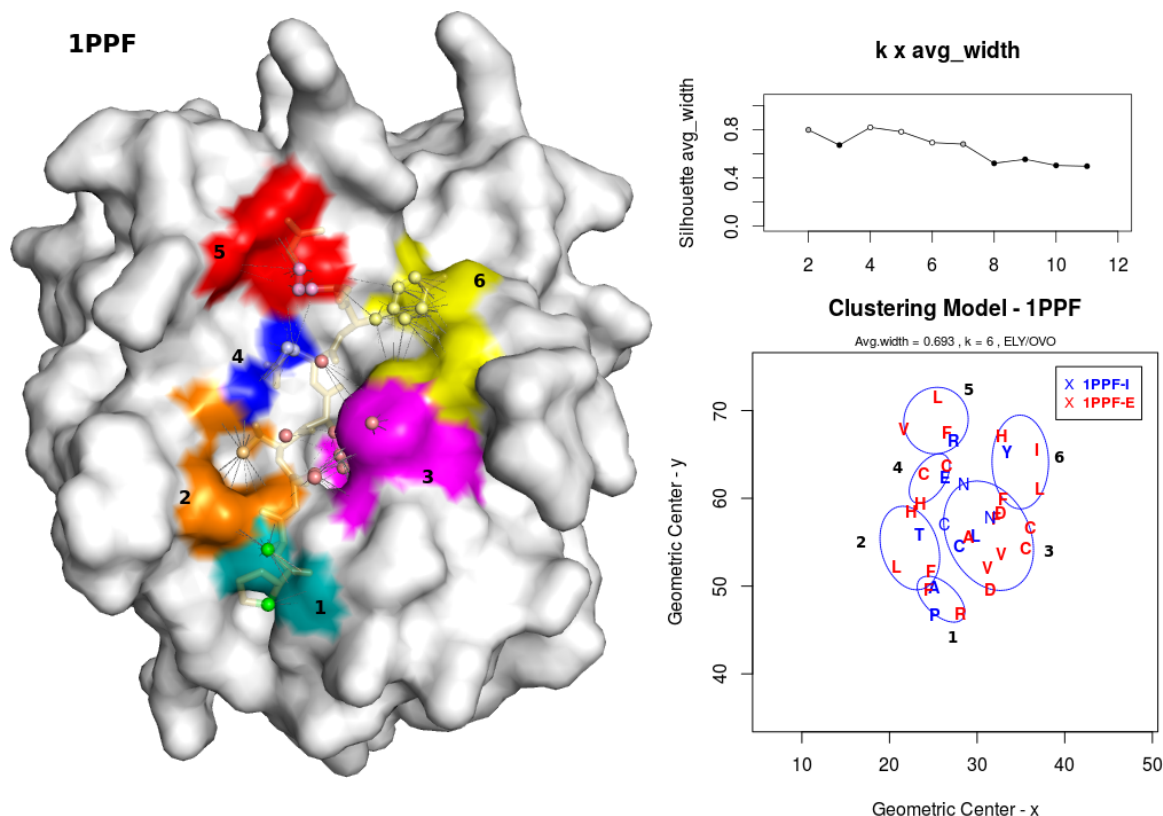
Como nossa granulosidade é no nível atômico, é permitido que átomos diferentes de um mesmo resíduo pertençam a subclusters diferentes. Como na figura a visualização é no nível de resíduos, pode acontecer que duas letras próximas nos limites de cada elipse sejam do mesmo resíduo. Por exemplo, na Figura 8.3, o resíduo **H**, em vermelho, nas subregiões 2 e 4 da representação gráfica intitulada *Clustering Model - 1PPF* são da mesma histidina. De certa forma, enquanto resíduo, essa histidina comporta-se como um elemento de conexão entre os subcluster 2 e 4 na figura.

8.4.1 Complexos Tipo Tripsinas e Inibidores

Analisamos primeiro aqui os complexos envolvendo peptidases tipo tripsinas e respectivos inibidores. Dos 36 complexos, 27 complexos (75 %) são desse tipo.

8.4.1.1 Complexo 1PPF

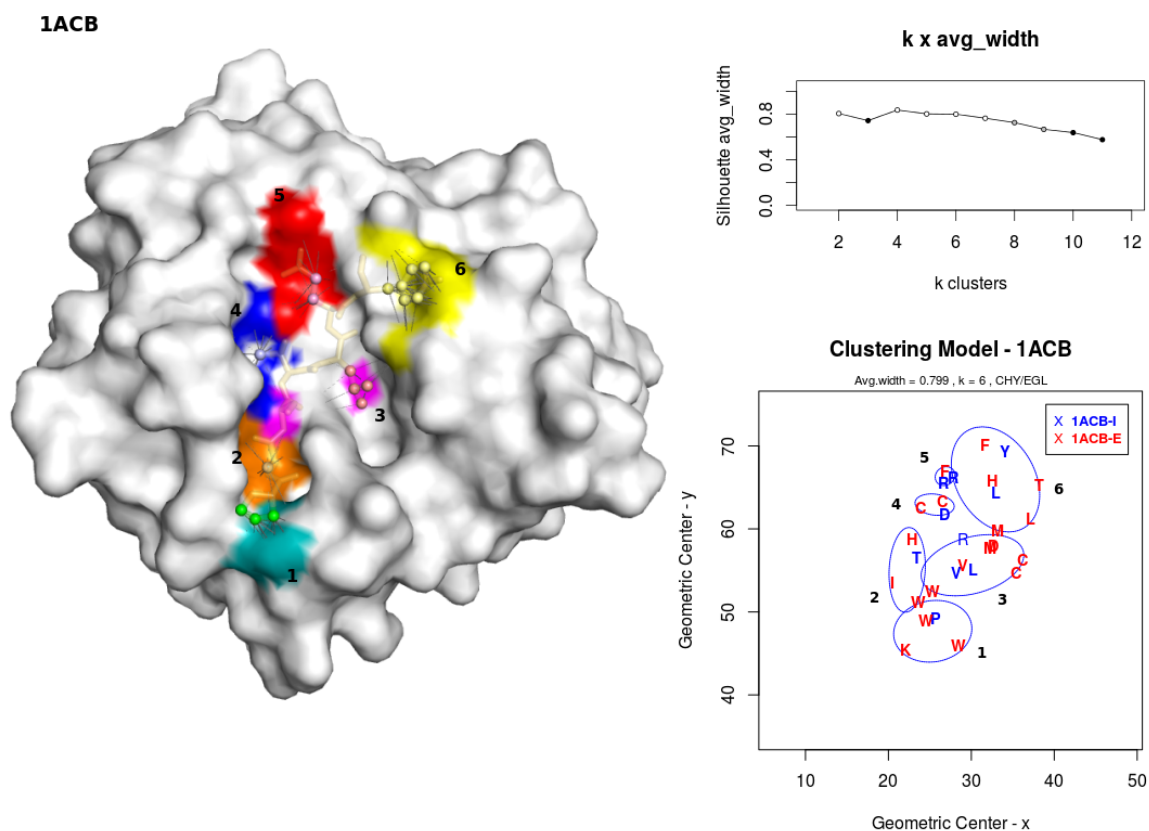
Figura 8.3: 1PPF (Elastase de leucócitos humanos - ELY e OMTKY3): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



Para o complexo 1PPF 6 regiões hidrofóbicas envolvendo resíduos da enzima e do inibidor foram delineadas como o maior número de grupos que podem ser formados com boa qualidade (Figura 8.3), pois os valores médios do coeficiente de Silhueta estão entre 0.6 e 0.8 ($0.6 \leq s_m \leq 0.8$) para $4 \leq k \leq 6$. O resíduo Leu18I está na posição *P1* do sítio de ligação e fica entre os seguimentos 214-216 e 191-192 da cadeia principal do bolsão de especificidade da ELY [Bode et al. (1986)]. O subcluster 3, contém a Leu18I e resíduos desses segmentos como Val216E, Val190E, Cys191E, Phe192E, Asp192E, Asp194E, além de Ala213E, Ala220E e Asp226E. Logo, a formação da região reflete as características presentes na estrutura em termos de contatos estabelecidos inibidor/enzima.

8.4.1.2 Complexo 1ACB

Figura 8.4: 1ACB (α -Quimotripsina e Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



Para $k = 6$, os subclusters hidrofóbicos obtidos para 1ACB foram comparados com o modelo de 1PPF. As enzimas α -Quimotripsina e ELY têm 35% de identidade sequencial, mas têm alto grau de similaridade estrutural [Frigerio et al. (1992)]. Os inibidores eglina (EGL) e ovomucoide (OMTKY3) são dois inibidores não homólogos (estão em clãs diferentes, conforme dito).

EGL é um forte inibidor de α -Quimotripsina e também de elastases e subtilisinas, com constante de inibição K_i variando entre 10^9 e $10^{11} M_1$. Quando se liga a β -tripsina bovina e elastase pancreática suína K_i está no intervalo de 10^5 a $10^6 M_1$.

Figura 8.5: Exemplo do modelo geral para os complexos 1PPF e 1ACB.

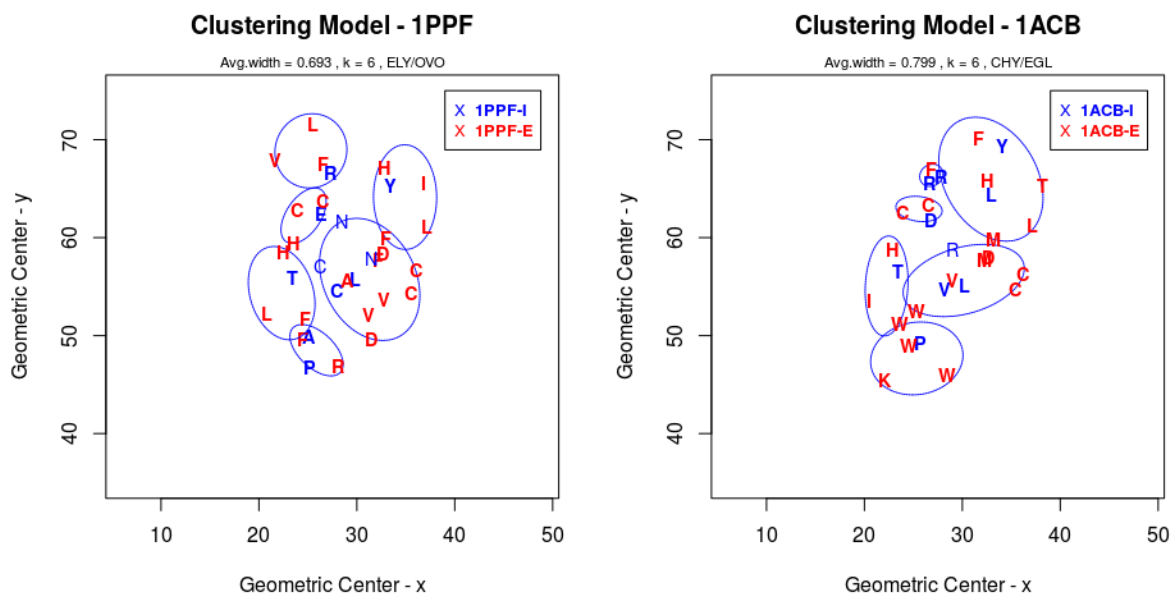
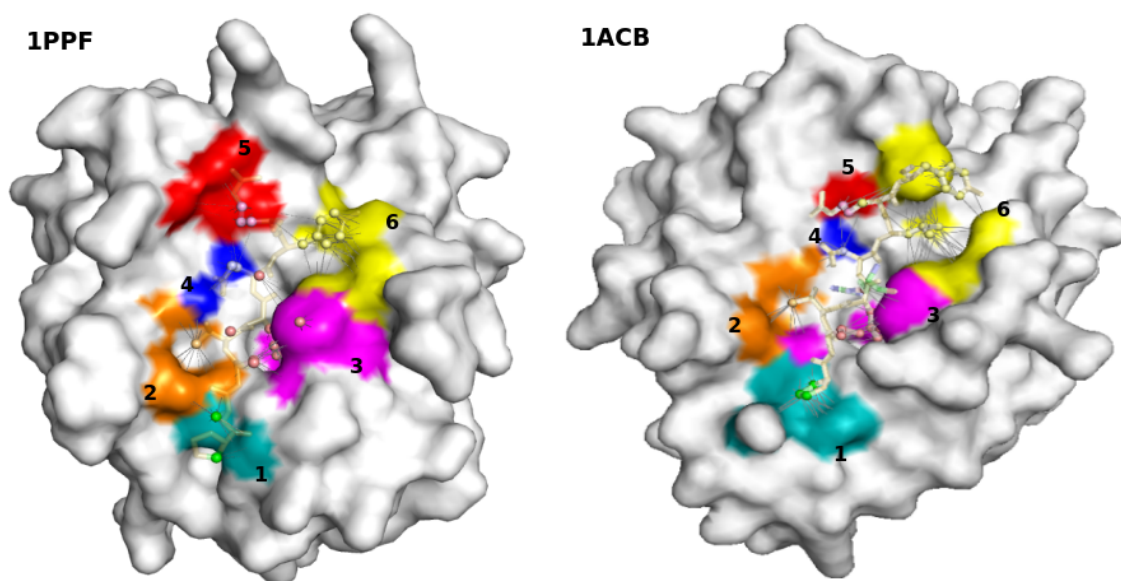


Figura 8.6: Visão 3D de 6 (k=6) regiões hidrofóbicas correspondentes (enzima-inibidor) para complexos 1PPF e 1ACB.



8.4.1.3 Complexo 4B2B

Figura 8.7: 4B2B (Tripsina com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

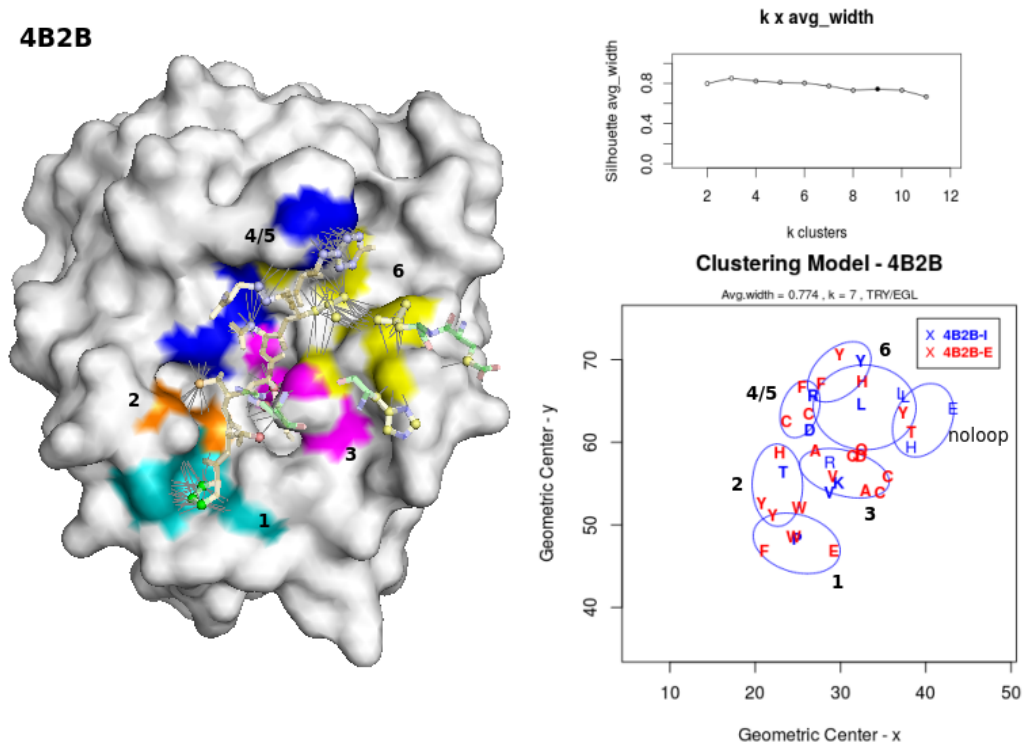
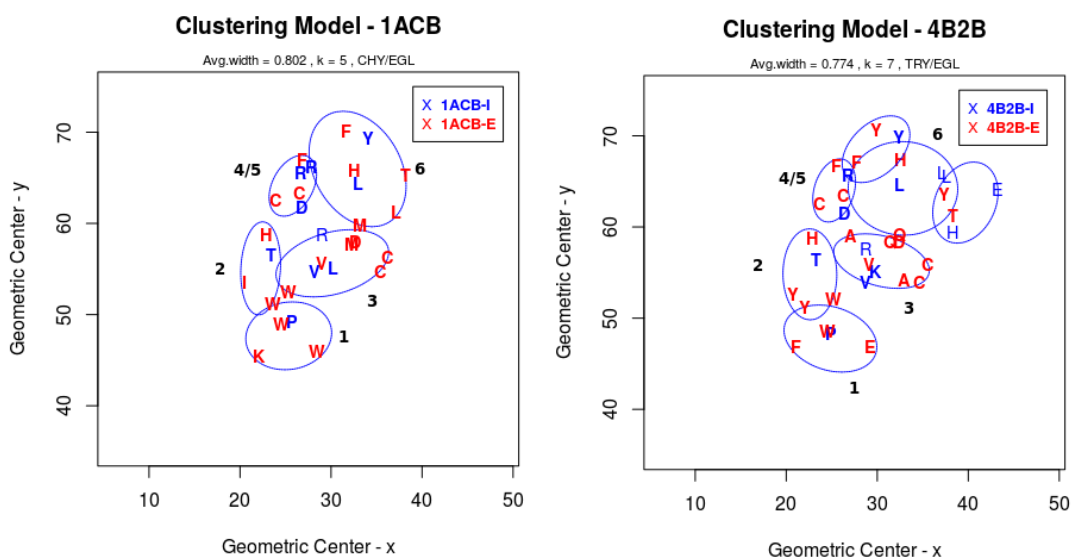


Figura 8.8: 1ACB e 4B2B: comparação de clusters.

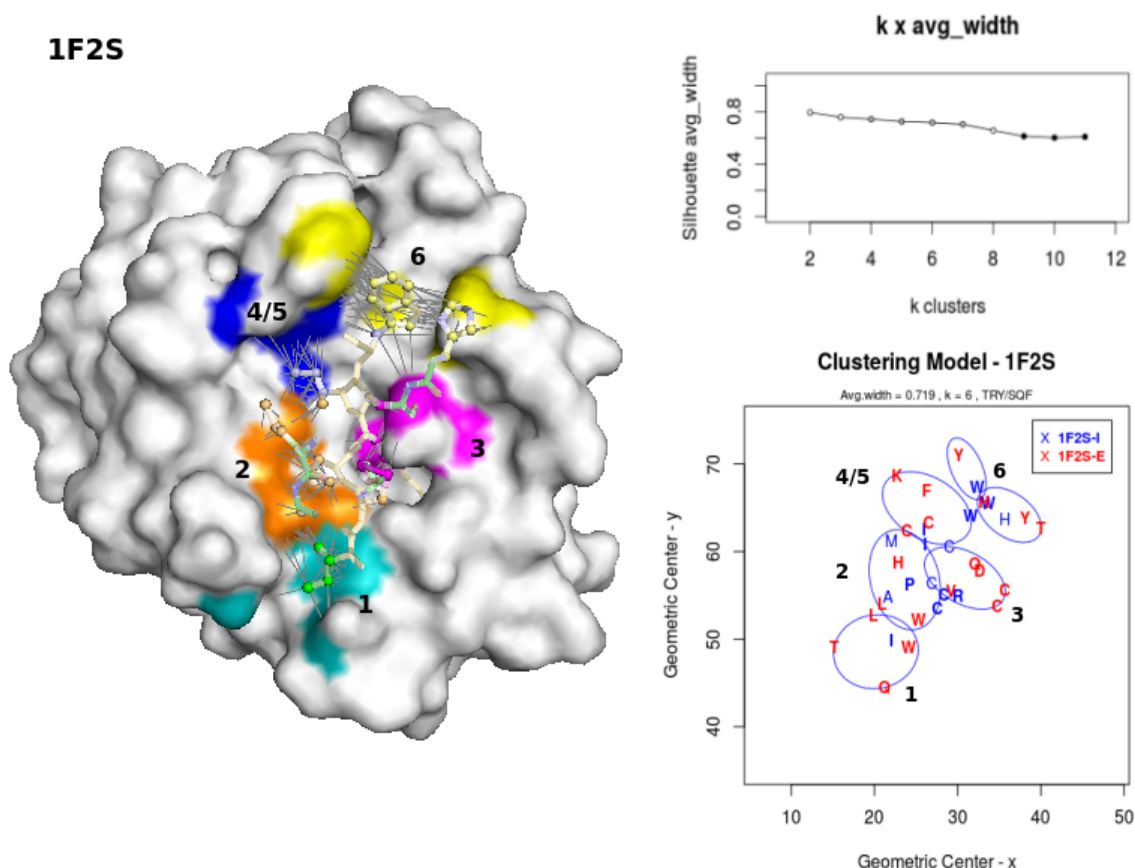


Uma variante de tripsina complexada com a Eglina mutante (L45K) formam o complexo 4B2B. As regiões hidrofóbicas para esse complexo são muito semelhantes às

obtidas para 1ACB (CHY/EGL), com ilustra a Figura 8.8. A região 3 em termos de composição dos resíduos dos inibidores refletem a diferença na sequência da alça devida a mutação L45K. Na vizinhança da região 6 de 4B2B, um subcluster de resíduos *noloop* está presente, em contraste com 1ACB. A região 6 de 1ACB se sobreposta à região 6 de 4B2B mantém a correspondência L-L, Y-Y.

8.4.1.4 Complexo 1F2S

Figura 8.9: 1F2S (β -tripsina e MCTI-II): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O complexo 1F2S tem o inibidor MCTI-II da família I7 (SQF) que difere de OVO e EGL. A alça inibitória de MCTI-II é formada por *ICPRIW* em contraste com *PACTLERY* do OMTKY3 de 1PPF e com *PVTLDLRY* da EGL de 1ACB. Ao se comparar os subclusters destes complexos (Figuras 8.10 e 8.11), para $k = 6$, a região 2 tem o resíduo P para os inibidores de 1PPF e 1ACB e T para o inibidor de 1F2S. Quando se compara as regiões de 1F2S com 1PPF (ambos contendo enzima do tipo Tripsina), há muita semelhança entre os posicionamentos dos clusters. No entanto, as regiões 4 e

5 em 1F2S se fundem (8.9). Além disso, considerou-se que a região 6, é composta por **YHYTWH**, embora no modelo sejam mostrados dois subgrupos **YW** e **HYTWH**. O **W** é o mesmo resíduo que está na fronteira entre os dois subgrupos.

Figura 8.10: 1PPF e 1F2S: comparação de clusters.

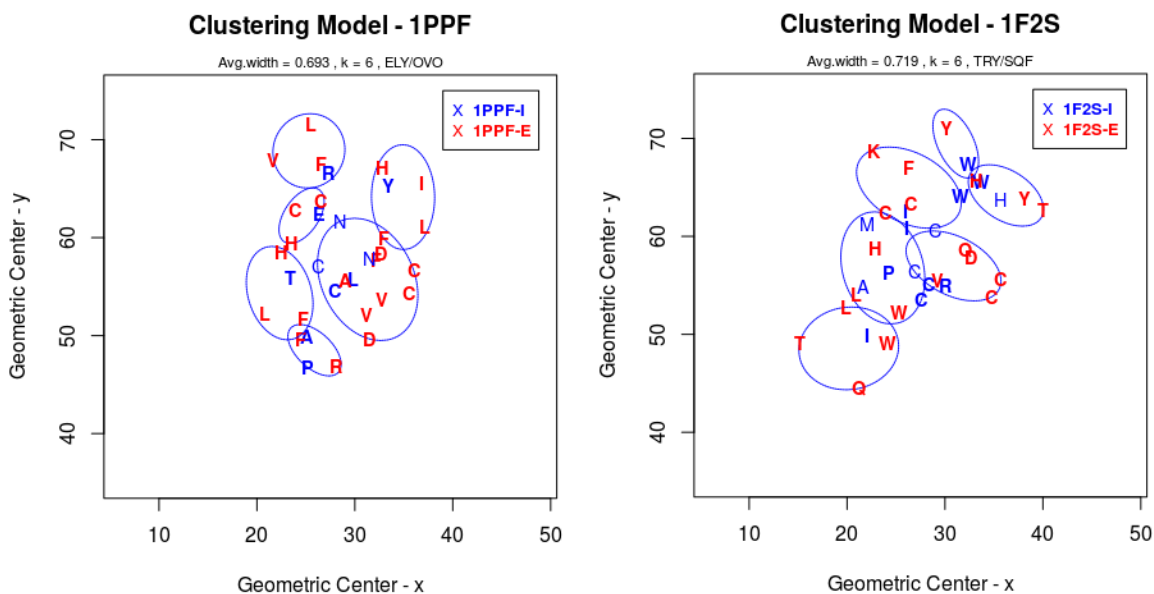
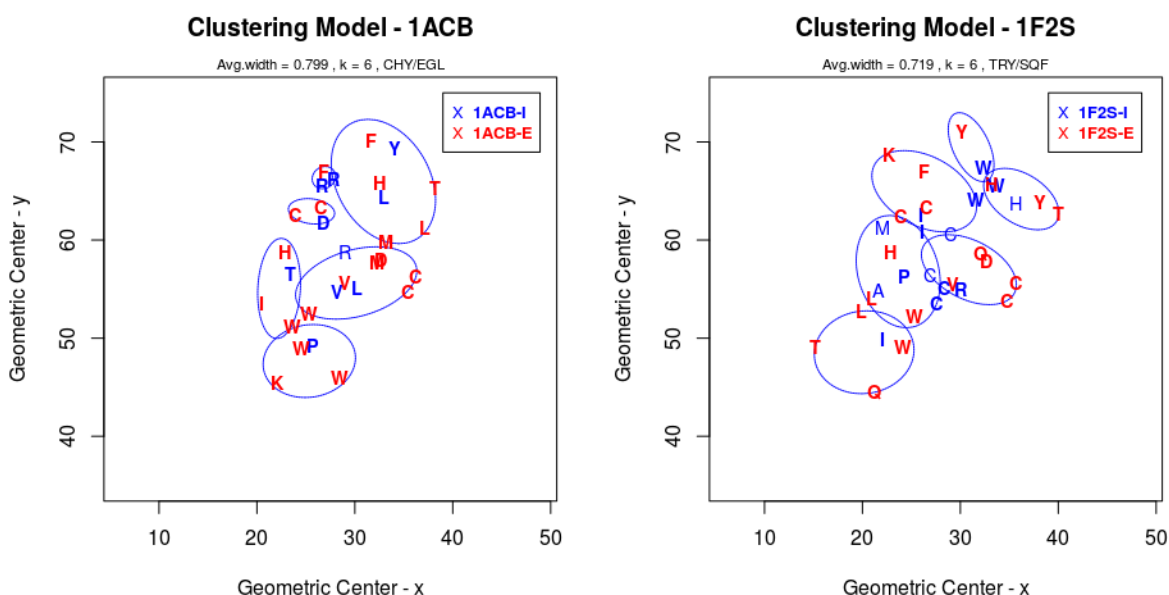
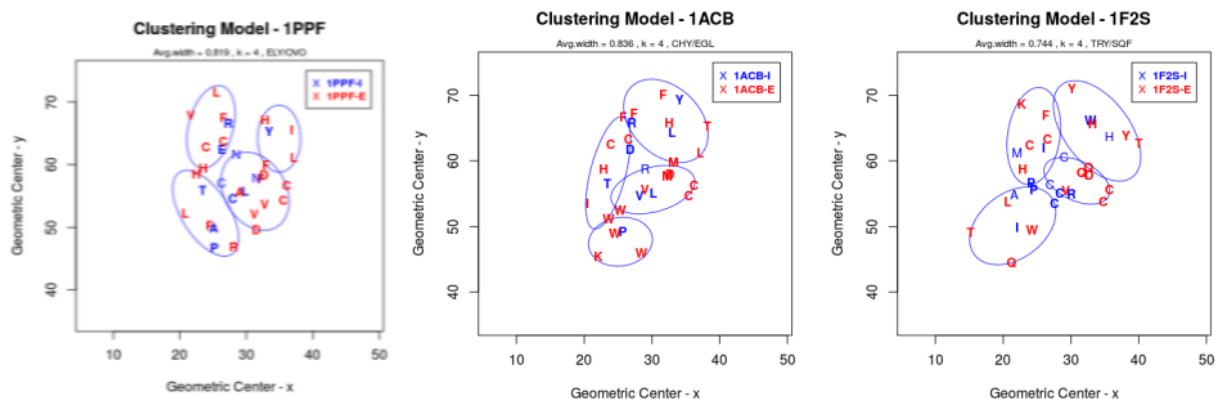
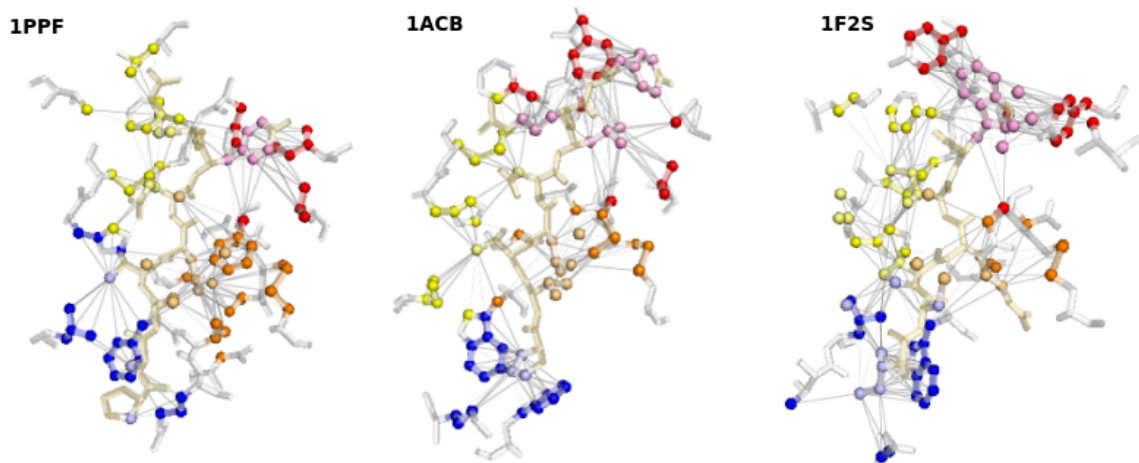


Figura 8.11: 1ACB e 1F2S: comparação de clusters.



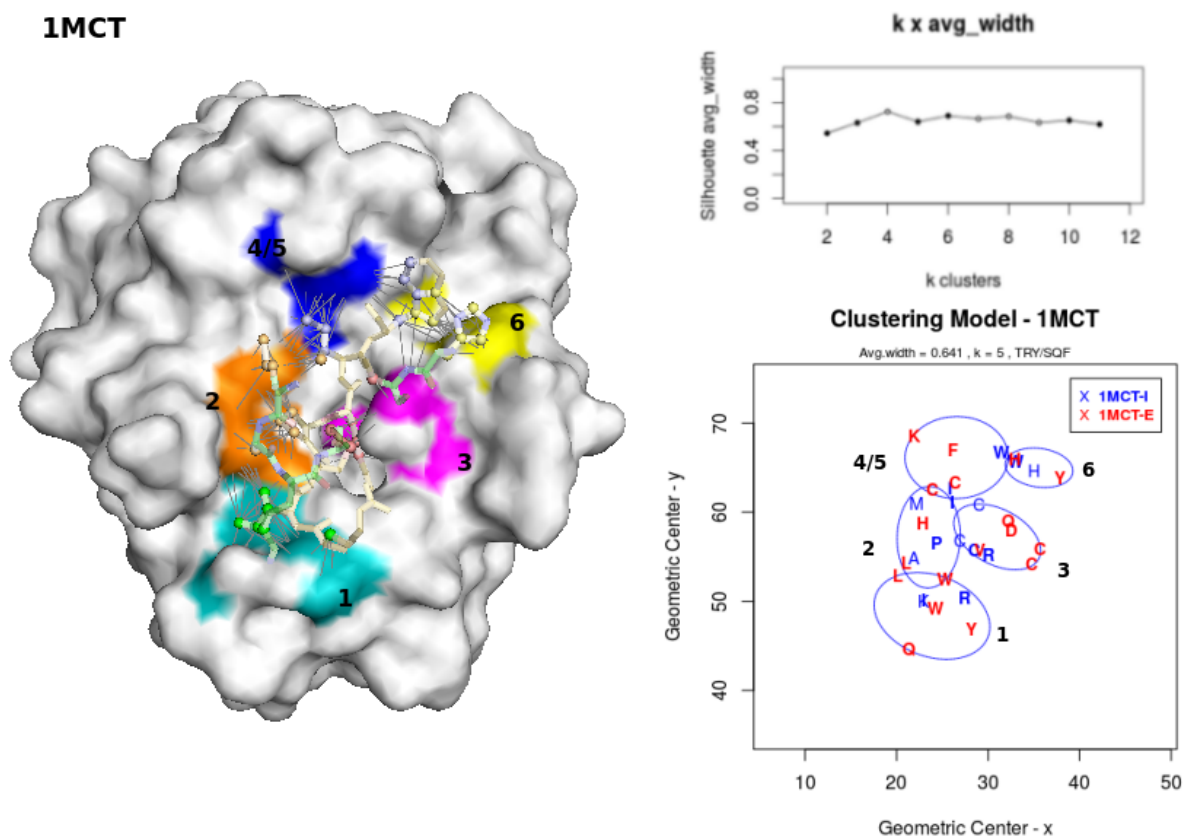
Para $k = 4$ subclusters, 1F2S tem um arranjo mais parecido com 1ACB do que com um 1PPF (Figuras 8.12 e 8.13).

Figura 8.12: 1PPF, 1ACB e 1F2S: comparação de clusters ($k=4$).Figura 8.13: 1PPF, 1ACB e 1F2S: visão de grafos e estrutura tridimensional ($k=4$).

Ao realizar essa varredura no número de clusters e considerando-se os coeficientes médios de Silhueta que caracterizam a formação de clusters de boa qualidade, aplicou-se a mesma abordagem para os demais complexos apresentados abaixo e construiu-se um modelo contendo de 4 a 5 regiões hidrofóbicas conservadas tanto para tripsinas quanto subtilisinas, na maioria dos casos.

8.4.1.5 Complexo 1MCT

Figura 8.14: 1MCT (Tripsina e MCTI-II): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



Os complexos 1F2S e 1MCT têm o mesmo tipo de enzimas (TRY) e inibidores (da mesma família (I7=SQF)). Comparando os clusters dos modelos de ambos, tem-se um padrão bem semelhante como pode ser visto na Figura 8.15. Logo, ao se comparar 1PPF com 1MCT (8.16), as mesmas considerações são válidas quando se comparou 1PPF com 1F2S. O padrão se mantém.

Figura 8.15: 1F2S e 1MCT: comparação de clusters.

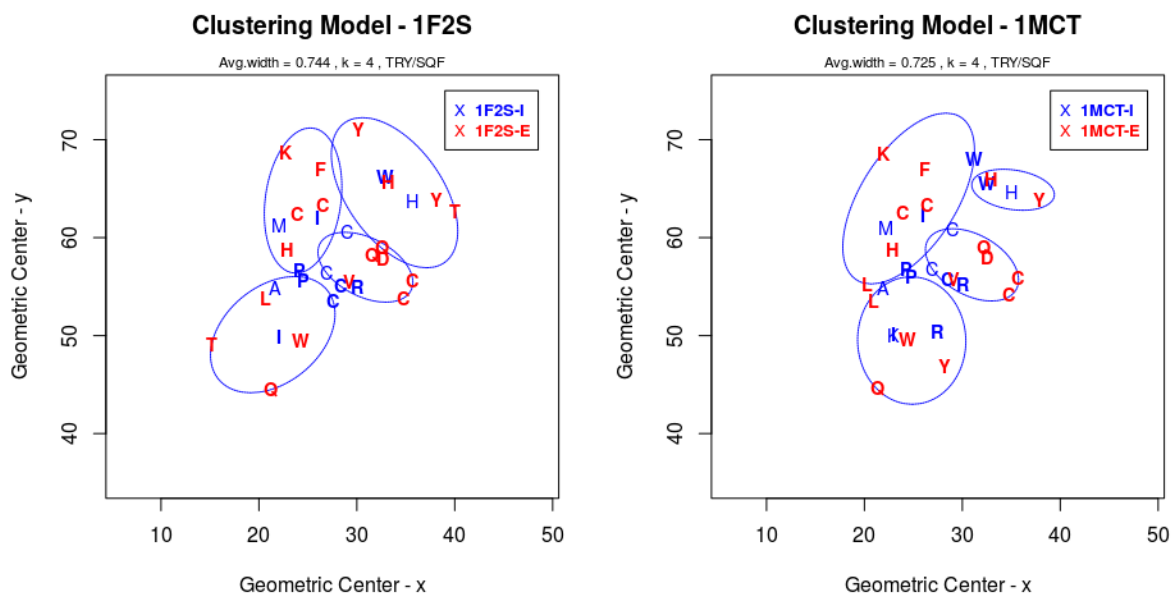
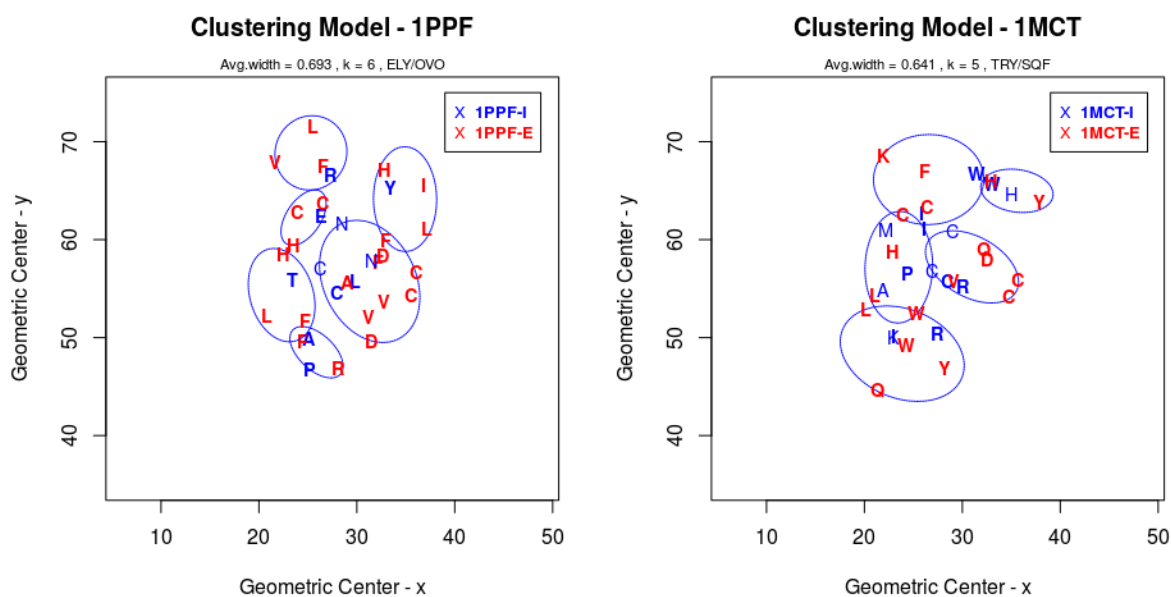
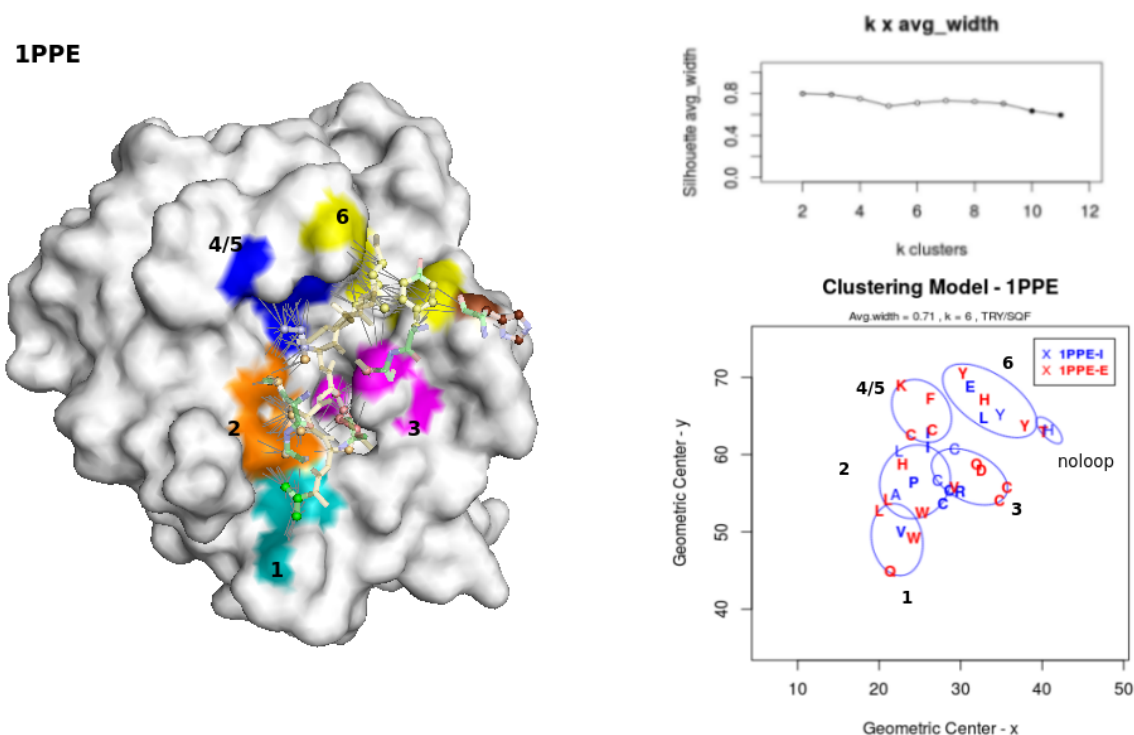


Figura 8.16: 1PPF e 1MCT: comparação de clusters.



8.4.1.6 Complexo 1PPE

Figura 8.17: 1PPE (Tripsina e CMTI-I): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



Como o complexo 1PPE tem o inibidor CMTI-I pertencente à família I7 (SQF), foi realizada uma comparação dos subclusters obtidos com os subclusters de 1F2S e 1MCT (Figuras 8.18 e 8.19, respectivamente). Na região 6, o **W** dos inibidores de 1F2S e 1MCT é substituído pelos dois resíduos **LE** da alça do inibidor CMTI-I. Isso é compreensível, dado o tamanho maior do triptofano que ocupa sozinho essa região nos dois outros complexos. Sobrepondo os três modelos (Figura 8.20), a similaridade fica mais evidente, a despeito da presença de uma região secundária formada por resíduos *noloop* em 1PPE (veja a região em marrom na Figura 8.17).

Figura 8.18: 1F2S e 1PPE: comparação de clusters.

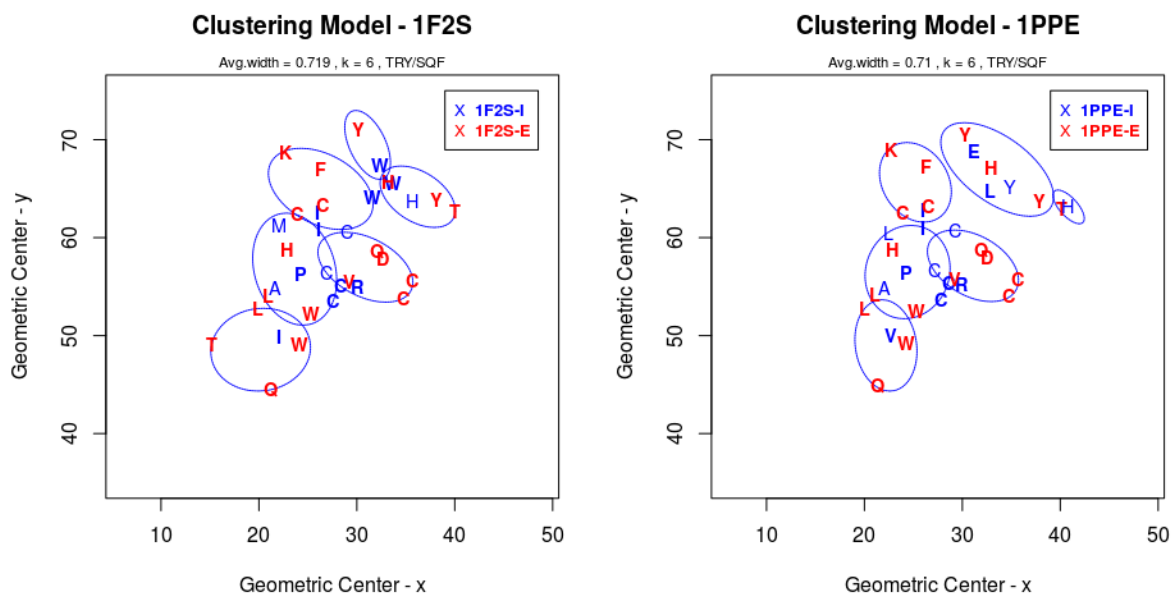


Figura 8.19: 1MCT e 1PPE: comparação de clusters.

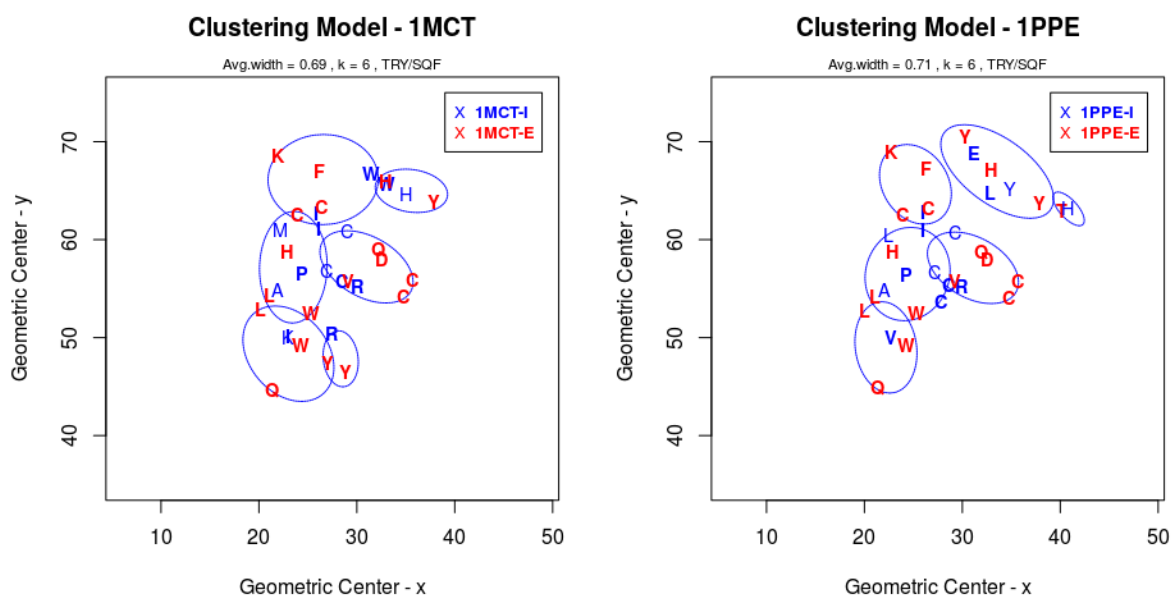
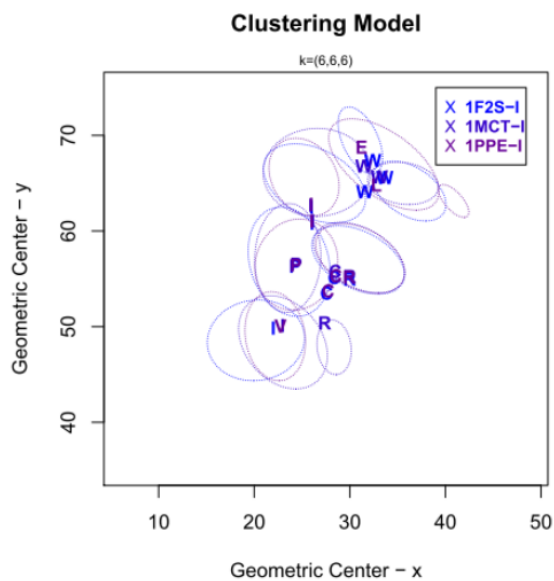


Figura 8.20: 1MCT, 1F2S e 1PPE: sobreposição de clusters - Inibidor



8.4.1.7 Complexo 1HJA

O complexo 1HJA, composto pela Quimotripsina C e EGL, também apresenta regiões hidrofóbicas semelhantes às encontradas em 1PPF (TRY/OVO), conforme se verifica na Figura 8.22. A exceção ocorre em 1HJA pela formação de um subcluster composto exclusivamente por resíduos *noloop*, a saber: Tyr31I e Lys29I (Figura 8.21, região marrom). Embora essa região apareça no modelo 2D como sendo aparentemente uma interseção entre as regiões 3 e 6, não o é. Essa é uma limitação do modelo ao mapear a visão 3D para 2D.

Figura 8.21: 1HJA (Quimotripsina C e Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

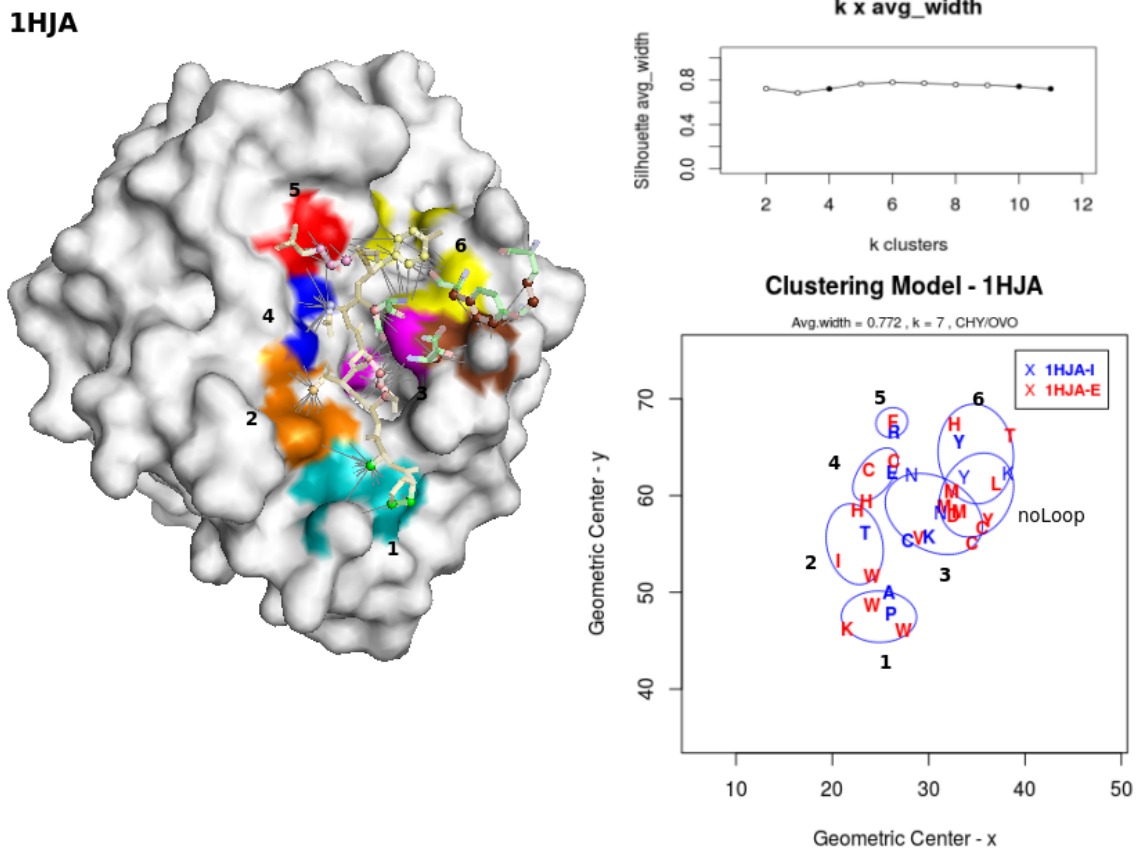
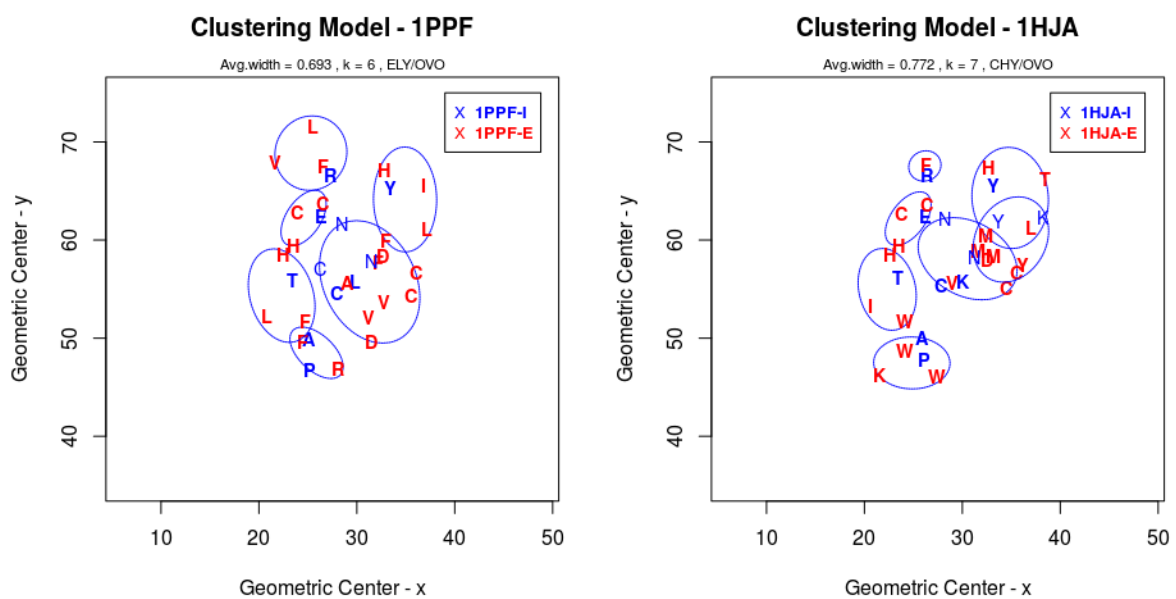
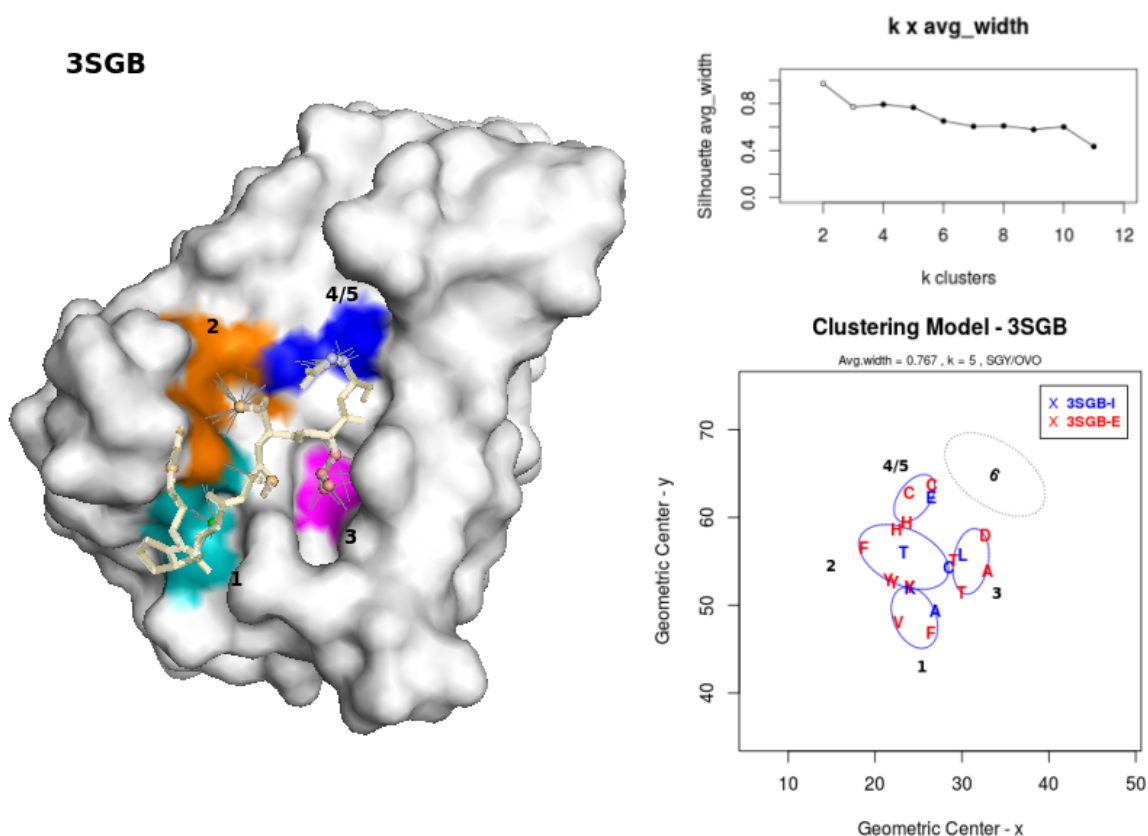


Figura 8.22: 1PPF e 1HJA: comparação de clusters (k=6 e k=7, respectivamente).



8.4.1.8 Complexo 3SGB

Figura 8.23: 3SGB (*Streptomyces griseus* proteinase B e OMTKY3): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



3SGB é composto pela enzima SGY e OMTKY3. Na Figura 8.23 a região do subcluster 6, presente nos outros modelos analisados até aqui, está vazia, como se ele não existisse. Mas esse subcluster existe na 3SGB. Acontece que estamos vendo nesta figura 8.23 o maior subcluster conexo. Se olharmos o modelo da figura 8.25 com todos os subgrafos conexos, percebe-se que há um elemento conexo menor, envolvendo **RTY**. Outros elementos conexos com apenas um nó aparecem em **N** e o **T**, e por formarem somente uma aresta com a enzima, não é possível desenhar uma elipse.

Percebemos neste caso algo muito importante: a formação dos contatos hidrofóbicos não necessariamente forma um único grafo todo conexo. É possível ter subregiões desconexas do ponto de vista da sua representação como um grafo. Mas, do ponto de vista termodinâmico, tais regiões desconexas vão fazer suas devidas contribuições para o ΔG de *binding*. Isso pode ser visto mais claramente na sobreposição da figura 8.26.

Figura 8.24: 1PPF e 3SGB: comparação de clusters considerando apenas o maior componente conexo.

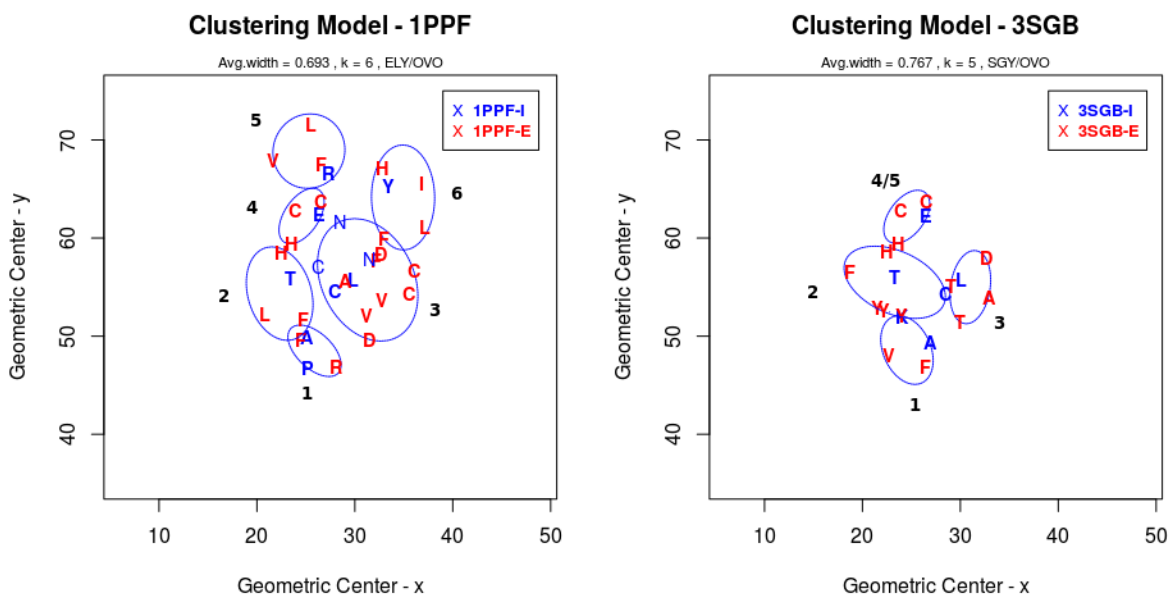


Figura 8.25: 1PPF e 3SGB: comparação considerando todos os componentes conexos.

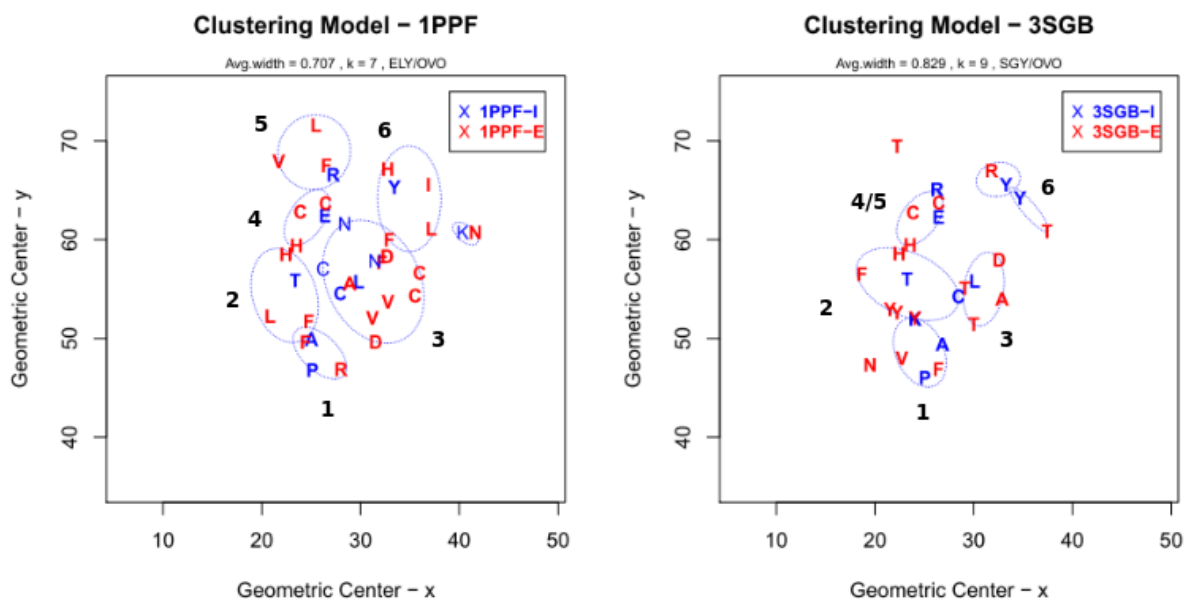
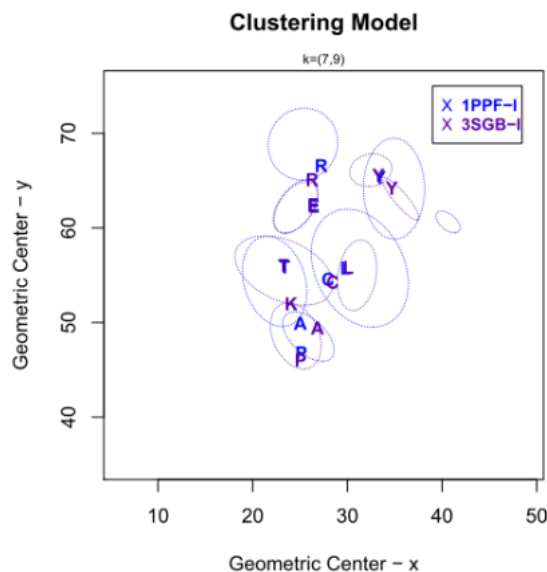


Figura 8.26: 1PPF e 3SGB: sobreposição - Inibidores.



8.4.1.9 Complexo 4SGB

4SGB é um complexo composto de uma SGY com o inibidor PCI-I da família I20 (POT). Esse complexo foi comparado com o 3SGB (SGY/OMTK3). As estruturas moleculares de 4SGB e 3SGB são muito próximas. Um dos *loops* de SGY (Ser35 a Gly40) difere na conformação nos dois complexos em mais de 2.0 Å para os átomos da cadeia principal. A Thr39 tem grandes diferenças com o átomo de carbono da carbonila com um desvio de 3.6 Å. Esta conformação alternativa é um resultado das diferenças nas estruturas moleculares dos resíduos posteriores a *P4'* do sítio reativo nos dois inibidores [Greenblatt et al. (1989)].

Comparando-se os modelos de clusters dos dois complexos em termos de regiões presentes, o elemento desconexo presente na região 6 de 3SGB, **RTY**, não está presente em 4SGB (Figura 8.28). Em relação à composição dos subclusters (Figura 8.29), é observada uma boa sobreposição entre os resíduos dos inibidores. Algumas diferenças ocorrem nas regiões 2, 4/5 e 6. Na região 2, o **T** de OVO é substituído por **P** em POT. Na região 4/5, o **RC** por o **N** e na região 6 o **Y** por **C**. Na região 3 o resíduo da posição *P1* se mantém o mesmo em ambos (L). Em 4SGB, também ocorre uma presença maior de *noloops*, associados às regiões 2 e 3: Cys57I (2) e Ala8I, Asn5I e Cys53I (3) (Figura 8.27). O inibidor PCI-1 tem 4 pontes dissulfeto, cujas cisteínas estão também visíveis no modelo, que presumivelmente têm influência na conformação do sítio reativo [Greenblatt et al. (1989)].

Figura 8.27: 4SGB (*Streptomyces griseus* proteinase B e Potato Inhibitor I (PCI-1)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

4SGB

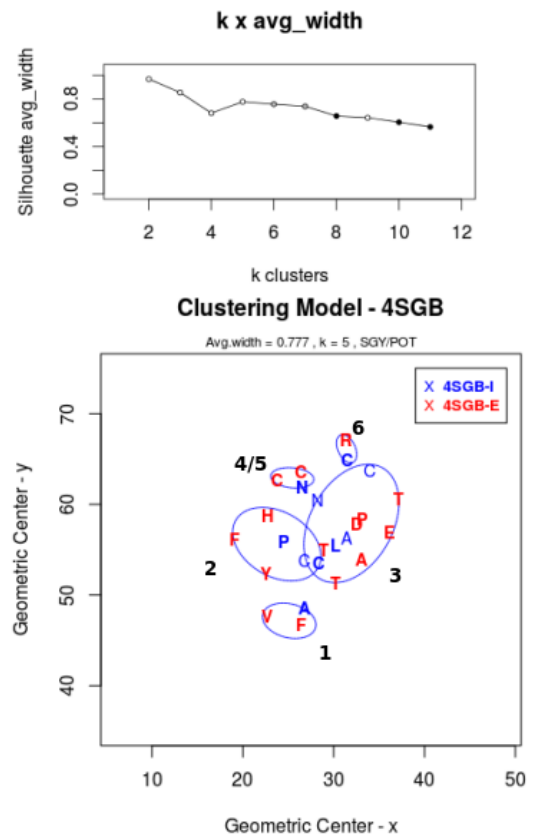
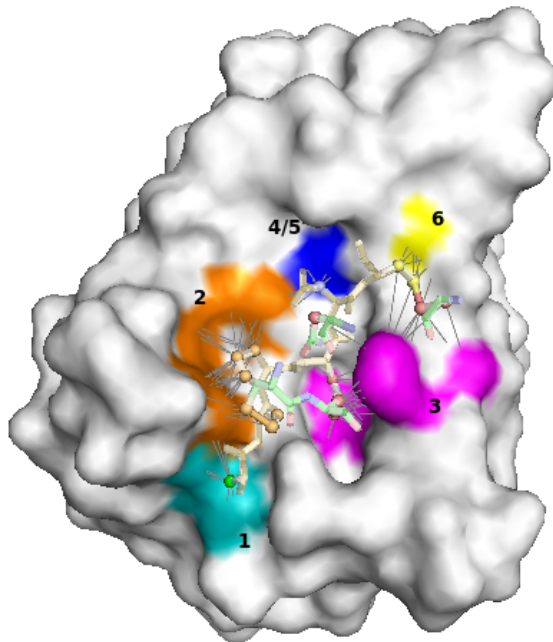


Figura 8.28: 3SGB e 4SGB: comparação de clusters.

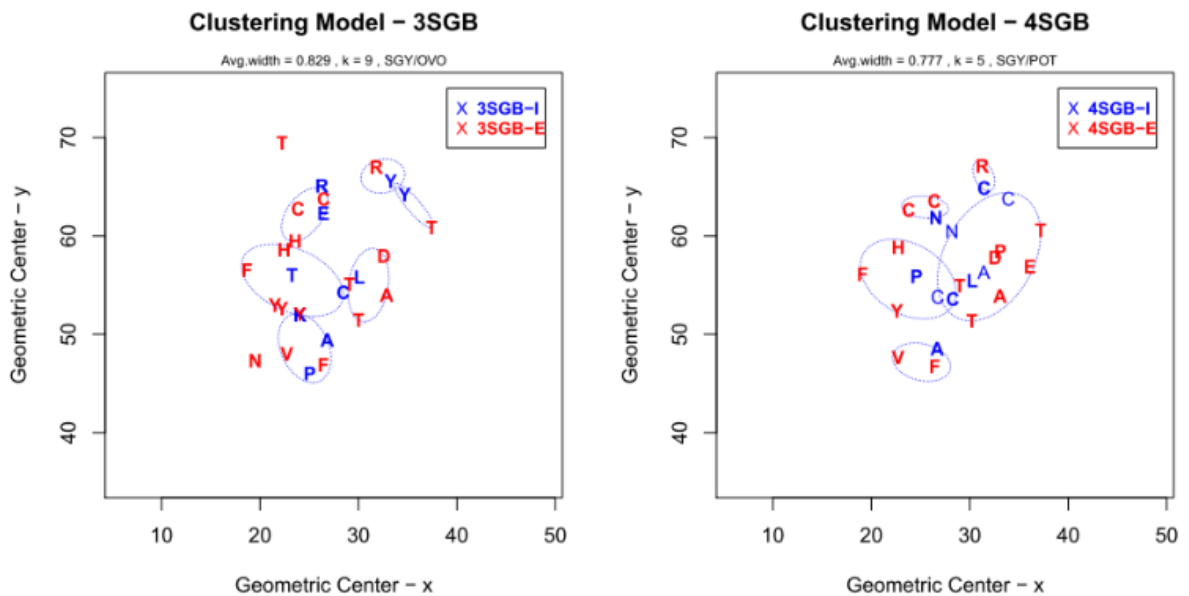
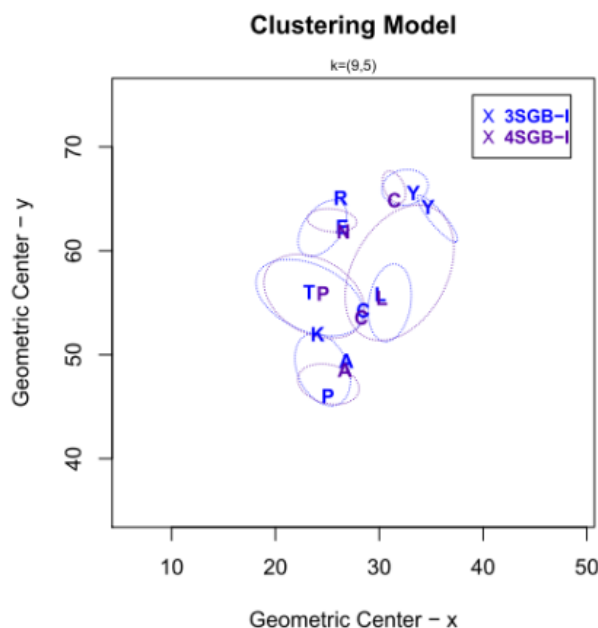


Figura 8.29: 3SGB e 4SGB: sobreposição de clusters - Inibidor



8.4.1.10 Complexo 1Z7K

O complexo 1Z7K é composto pela β -Tripsina suína e OMTKY2 (ovomucoide do segundo domínio, família I1). O inibidor OMTK (*Turkey egg*) é um inibidor de “cabeça dupla” (*double-headed*), onde o segundo domínio é específico para tripsina e o terceiro para quimotripsina e elastase. A conexão peptídica entre esses domínios pode ser facilmente hidrolisada e os domínios resultantes são independentemente reativos, sendo chamados de (OMTKY2) e (OMTKY3). Na família Kazal, o resíduo localizado em *P1* tem um papel vital na inibição da enzima e é responsável por 50% de todos os contatos na interface de ligação proteína-inibidor. Os resíduos localizados nessa região de *P3* a *P4'* têm conformação e interações conservadas [Ibrahim and Patabhi (2004)].

O OMTKY2 tem um *fold* canônico do tipo Kazal com uma α -hélice central e uma pequena folha- β antiparalela. O carbono do sítio reativo prefere uma geometria trigonal. A geometria do *loop* do sítio reativo é complementar à superfície e carga do sítio de ligação da β -Tripsina. OMTKY2 e OMTKY3 tem identidade sequencial em torno de 31% no ClustalW e apresentam na região *P4* a *P3'* somente a posição *P3* conservada (Tabela 8.4).

Figura 8.30: 1Z7K (β -Tripsina suína e OMTKY2): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

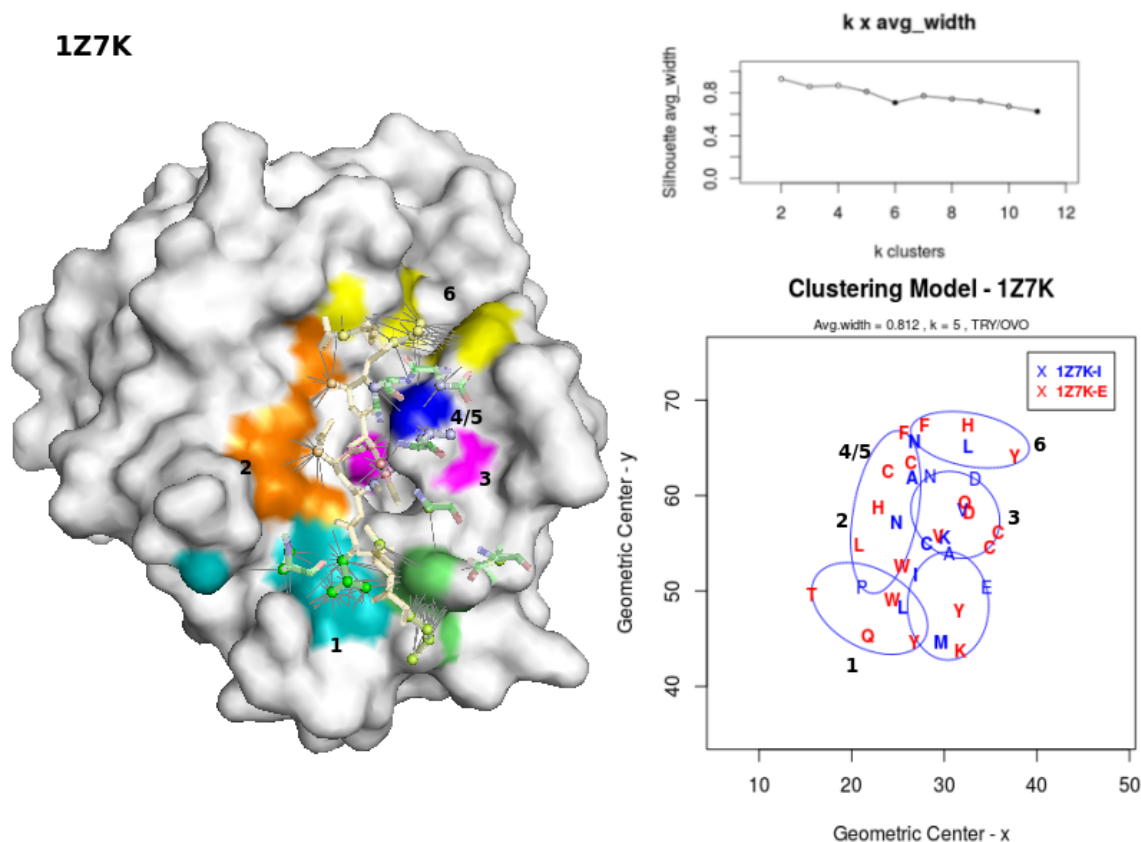


Tabela 8.4: Sequências $P4$ a $P3'$ da região do sítio ativo de OMTKY2 e OMTKY3. $P3$ é a posição conservada entre esses domínios.

$P4$	$P3$	$P2$	$P1$	$P1'$	$P2'$	$P3'$	
L	C	N	K	A	L	N	Segundo domínio
A	C	T	L	E	Y	R	Terceiro domínio

Os modelos de clusters de 1Z7K (TRY-OMTKY2) e 1PPF (TRY-OMTK3) não são parecidos (Figuras 8.31). Porém, na sobreposição dos modelos somente contendo os resíduos dos inibidores (Figura 8.32), há uma boa correspondência entre C-C, L-K, F-A, R-N, Y-L e talvez T-N (o primeiro resíduo é de 1PPF e o segundo de 1Z7K). Constata-se também que uma L em 1Z7K pode estar fazendo o papel de uma dupla P-A em 1PPF. Há a indicação de um cluster a mais para 1Z7K, colorido em verde na Figura 8.30, com centroide numa metionina (M) e os subclusters 5 e 6 parecem que sofreram fusão, embora sejam exibidos na estrutura tridimensional separadamente.

Figura 8.31: 1PPF e 1Z7K: comparação de clusters.

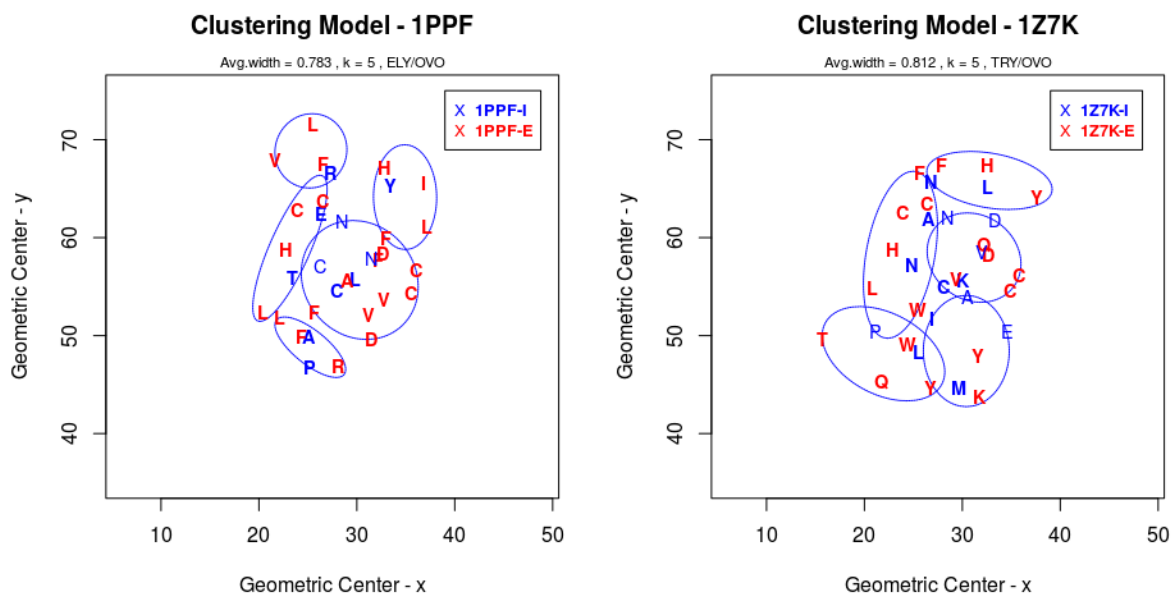
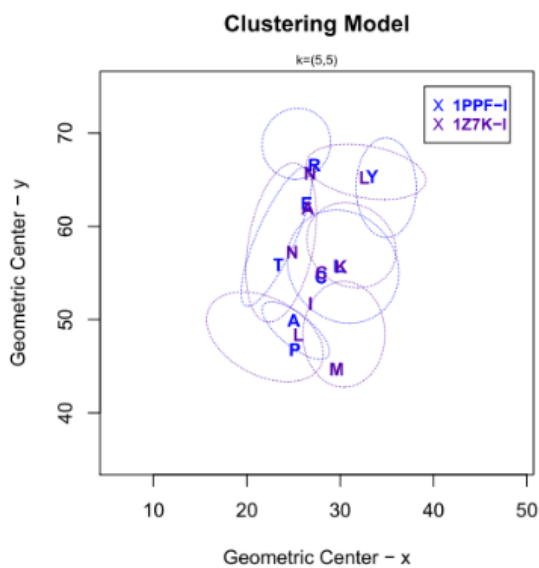


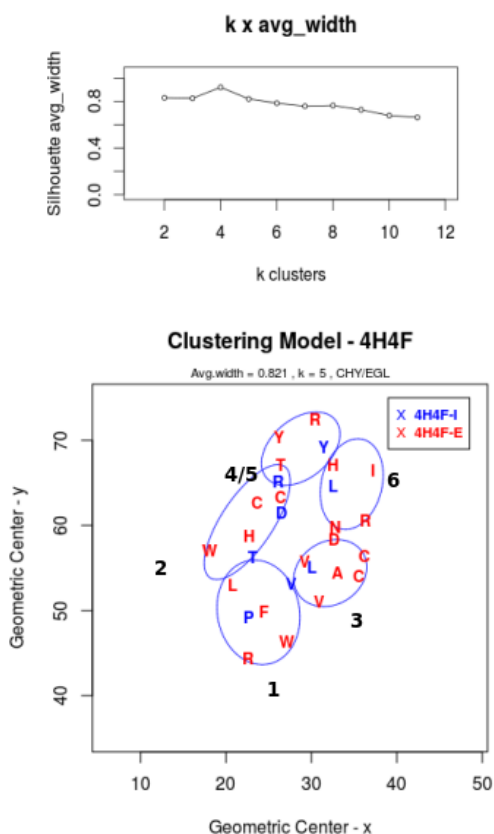
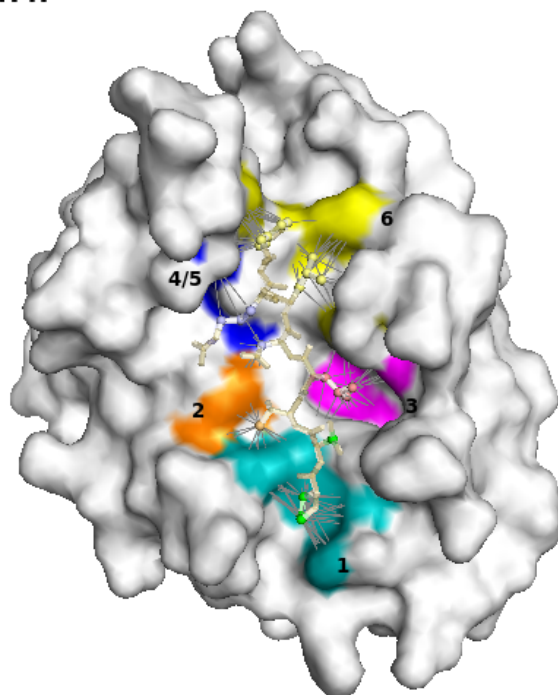
Figura 8.32: 1PPF e 1Z7K: sobreposição de clusters - Inibidores.



8.4.1.11 Complexo 4H4F

Figura 8.33: 4H4F (Quimotripsina C e Eglina C): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

4H4F



Uma enzima CHY e o inibidor ELG são os componentes de 4H4F. Na comparação dos subclusters deste complexo com 1ACB (α CHY/EGL), há muitas semelhanças (Figuras 8.34 e 8.35). Porém, enquanto em 1ACB, LY tendem a formar o subcluster 6, em 4H4F eles tendem à separação, com o Y mais próximo do subcluster 5 (Figura 8.34). O maior afastamento entre Y e L em 4H4F provavelmente explique a tendência de segregação do subcluster 6.

Figura 8.34: 1ACB e 4H4F: comparação de clusters.

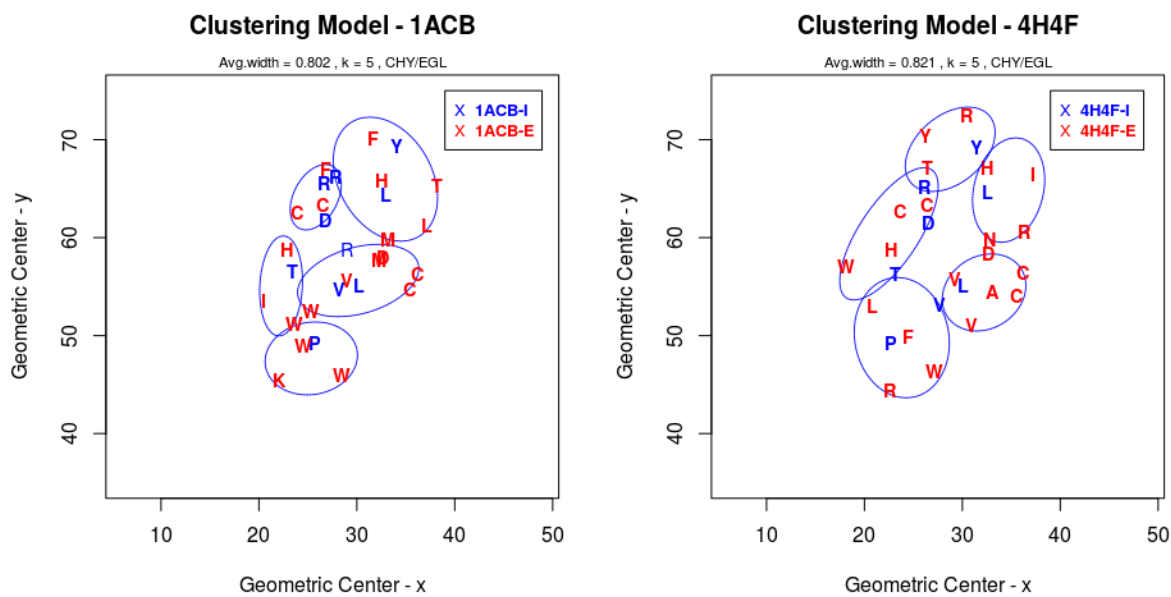
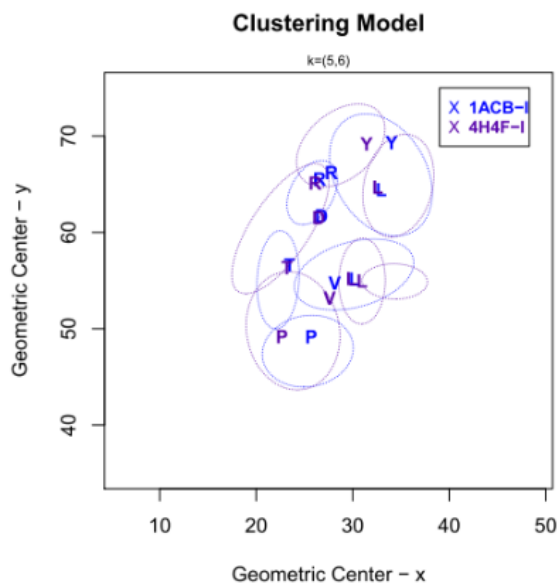
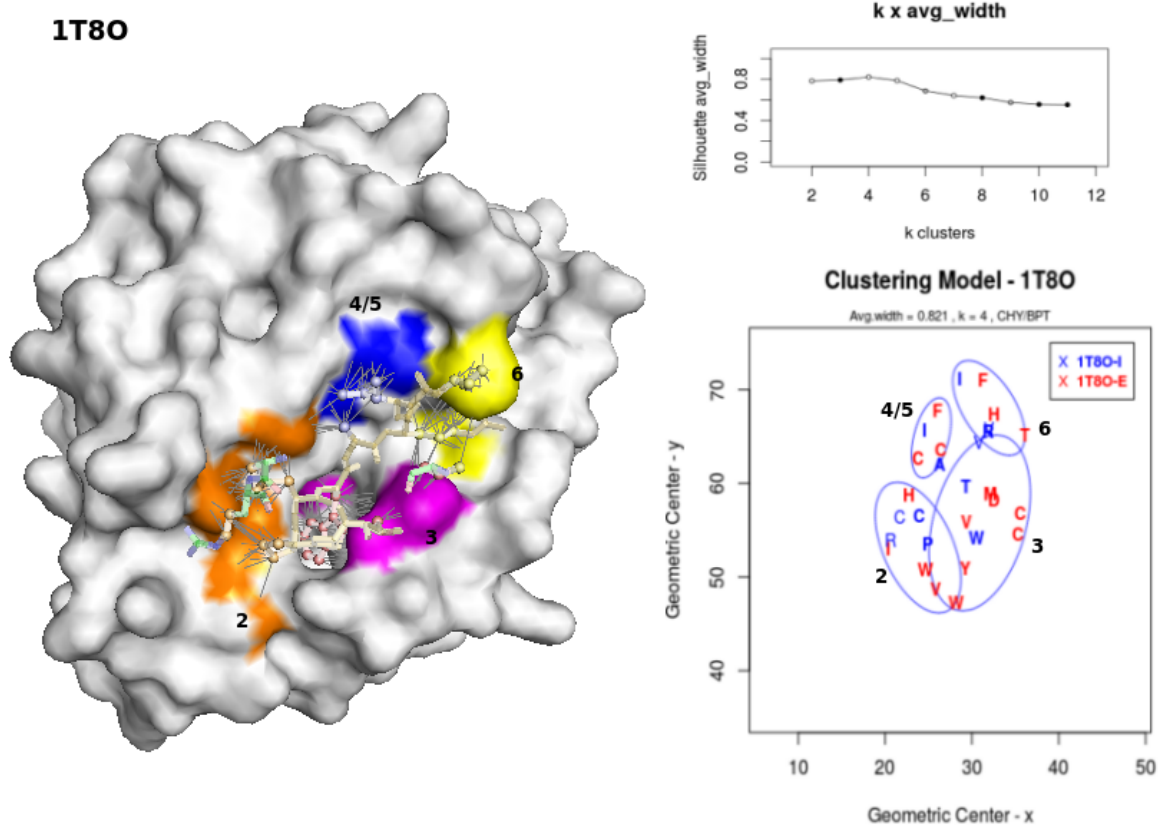


Figura 8.35: 1ACB e 4H4F: comparação de clusters - Inibidores.



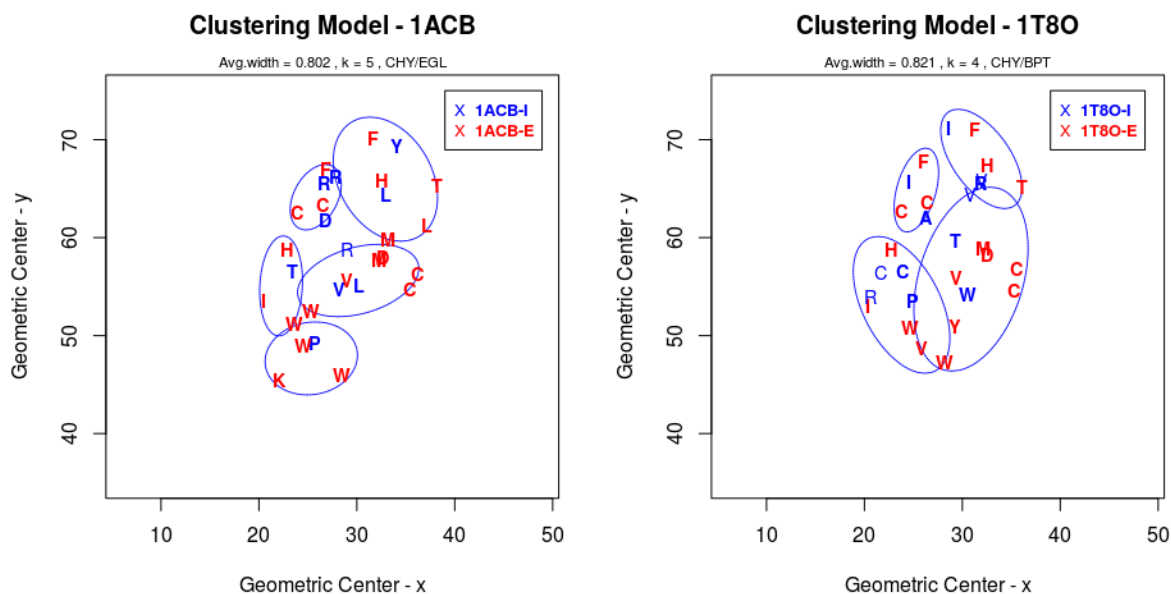
8.4.1.12 Complexo 1T80

Figura 8.36: 1T80 (Quimotripsina A e BPTI): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



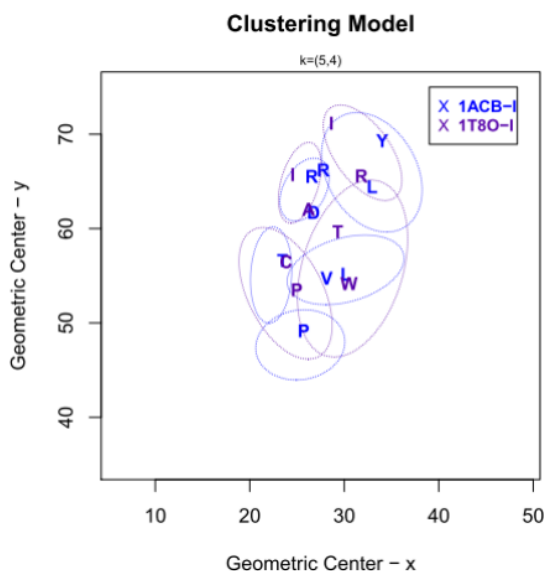
Como o complexo 1T80 é uma CHY com inibidor BPTI (família I2), foi realizada uma comparação dos suclusters obtidos para 1ACB cuja enzima é uma CHY (Figura 8.37). Observa-se que o BPT apresenta um padrão novo, pois a região 1 ou está vazia ou sofreu uma fusão com a região 2.

Figura 8.37: 1ACB e 4H4F: comparação de clusters.



Da sobreposição do conjunto de subclusters (lado inibidor), nota-se que as prolinas de ambos inibidores (BPTI/ELG) estão realmente em regiões distintas (Figura 8.38). Há boa sobreposição entre T-C, L-W, D-A, e mais parcial entre R-R, e R-L. Ressalta-se que o R pode fazer papel hidrofóbico de um L.

Figura 8.38: 1ACB e 1T80: sobreposição de clusters - Inibidores.



Comparando-se 1HJA (CHY/OVO) e 1T80 (Figura 8.39) e pela sobreposição dos clusters (Figura 8.40), novamente são observadas as sobreposições mais robustas T-C, E-A e mais parciais: K-W, Y-R.

Figura 8.39: 1HJA e 1T8O: comparação de clusters - Inibidores.

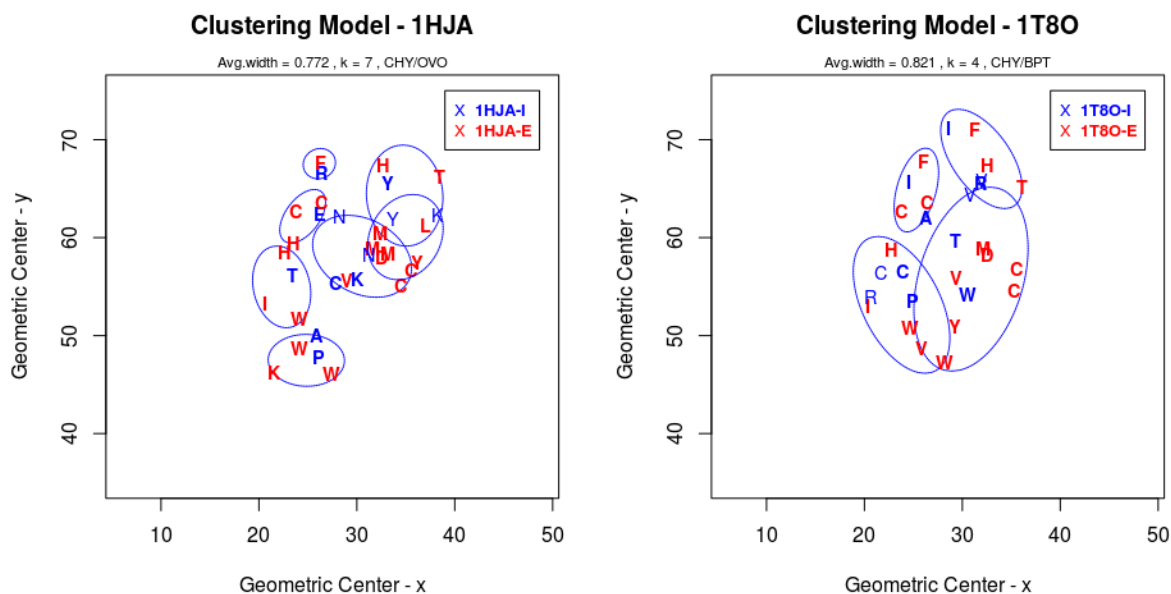
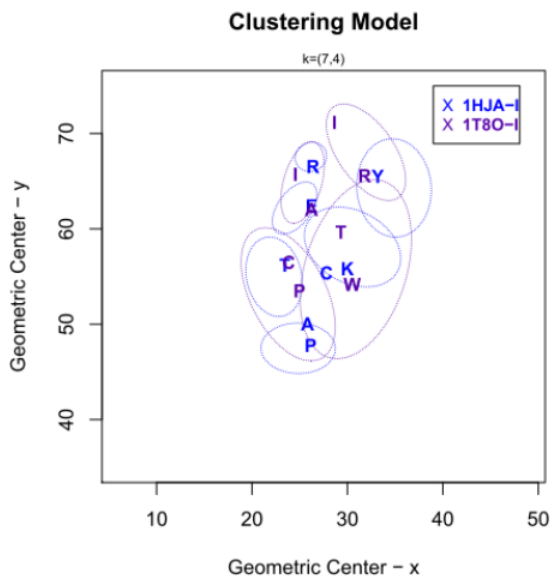
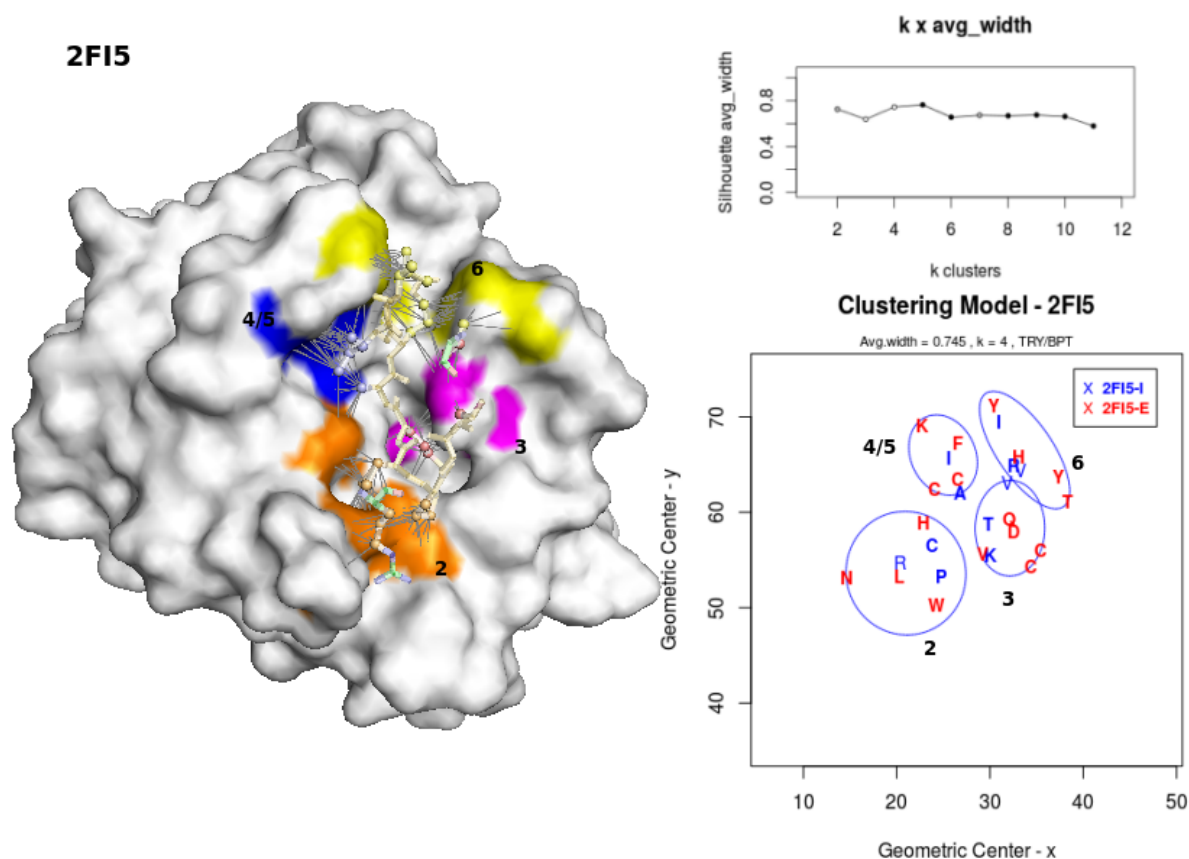


Figura 8.40: 1HJA e 1T8O: sobreposição de clusters



8.4.1.13 Complexo 2FI5

Figura 8.41: 2FI5 (Tripsina A e BPTI): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



2FI5, composto de uma Tripsina A e o inibidor BPTI (TRY/BPT), assim como 1T8O (TRY/BPT) não apresenta a região 1 observada para a maioria dos complexos acima (Figura 8.41).

A correspondência é evidente entre os subclusters de 1T8O e 2FI5, conforme observado na Figura 8.42 e na sobreposição dos subclusters com resíduos do inibidor (Figura 8.43). Todos os resíduos são praticamente os mesmos, conforme esperado para o mesmo inibidor: P-P, C-C (região 2), T-T (região 3), A-A, I-I (região 4/5), I-I, R-R (região 6). Na região 3, entretanto, o triptofano de 1T8O tem correspondência com a lisina de 2FI5.

Figura 8.42: 1T8O e 2FI5: comparação de clusters

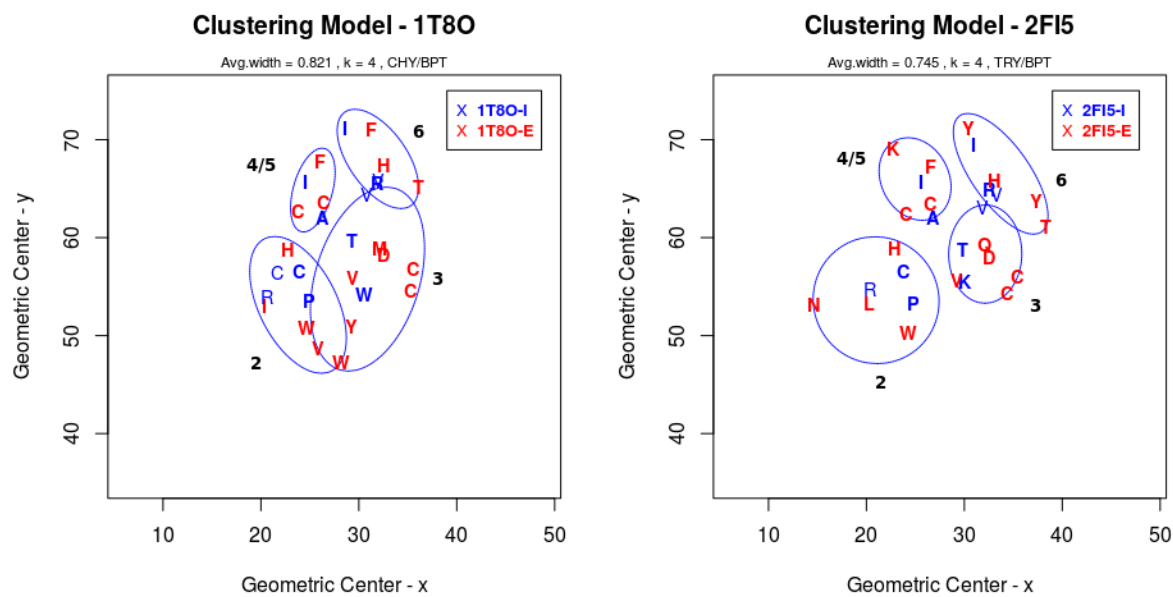
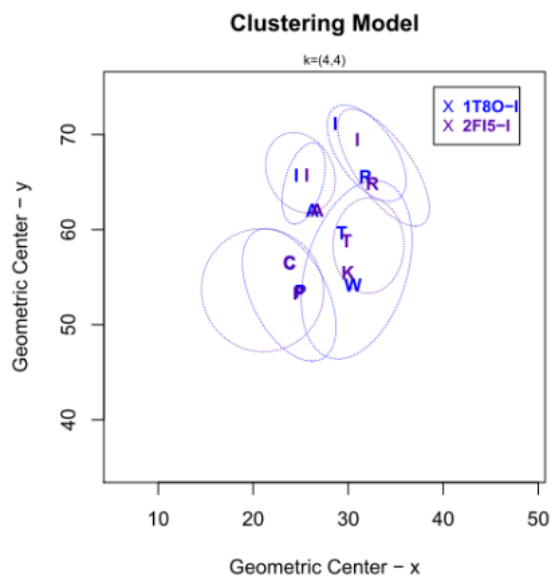


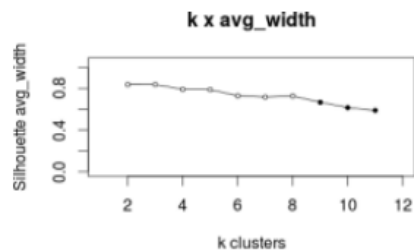
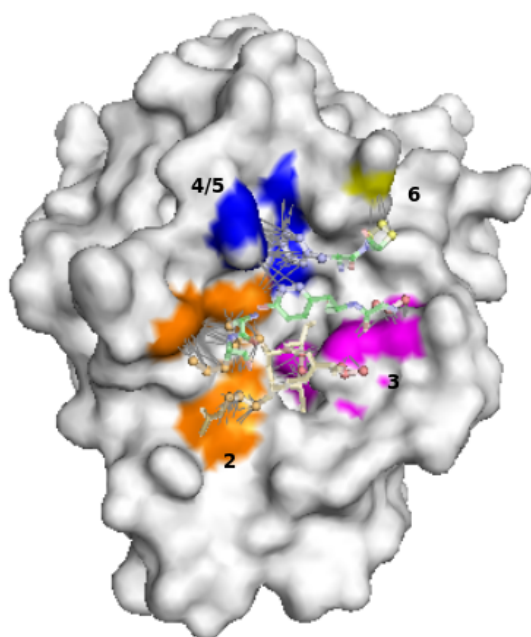
Figura 8.43: 1T8O e 2FI5: sobreposição de clusters - Inibidores.



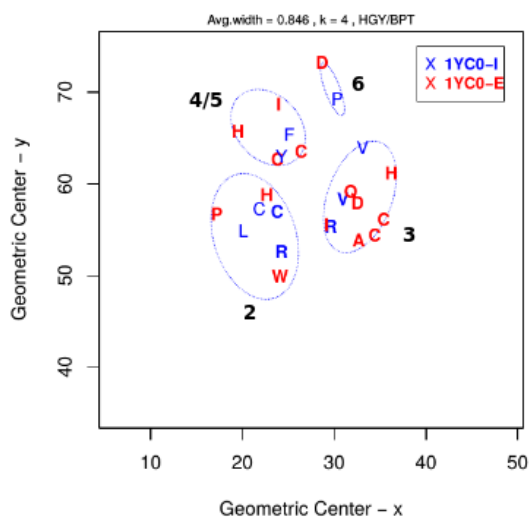
8.4.1.14 Complexo 1YC0

Figura 8.44: 1YC0 (*Hepatocyte growth factor activator* (HGFA) e *Kunitz-type protease inhibitor 1* (HAI-1)): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

1YC0

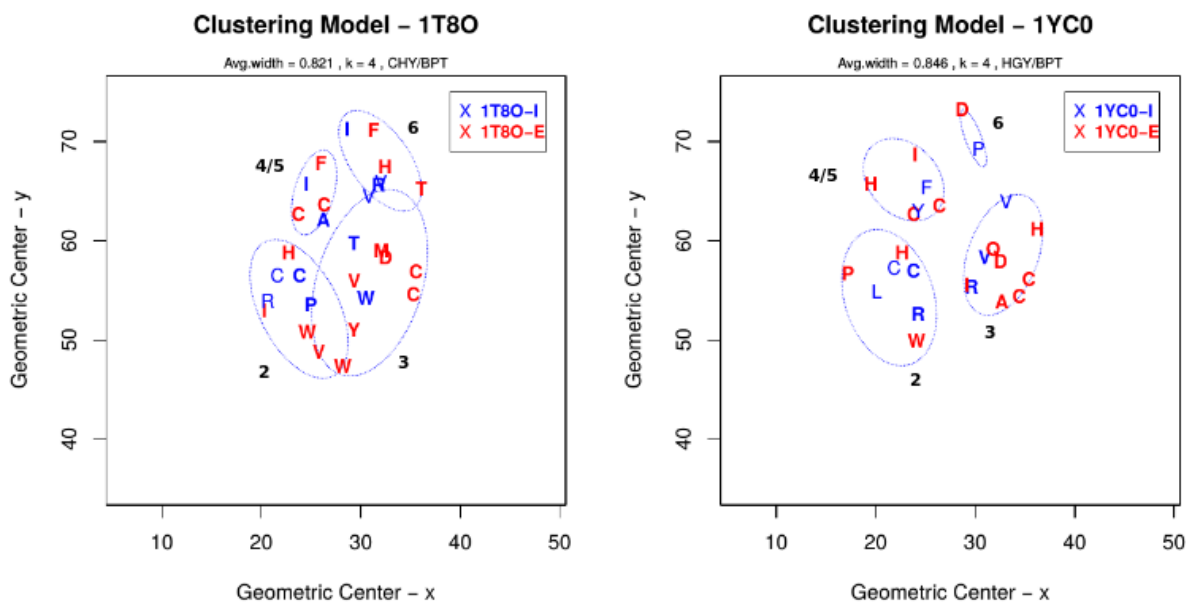


Clustering Model - 1YC0



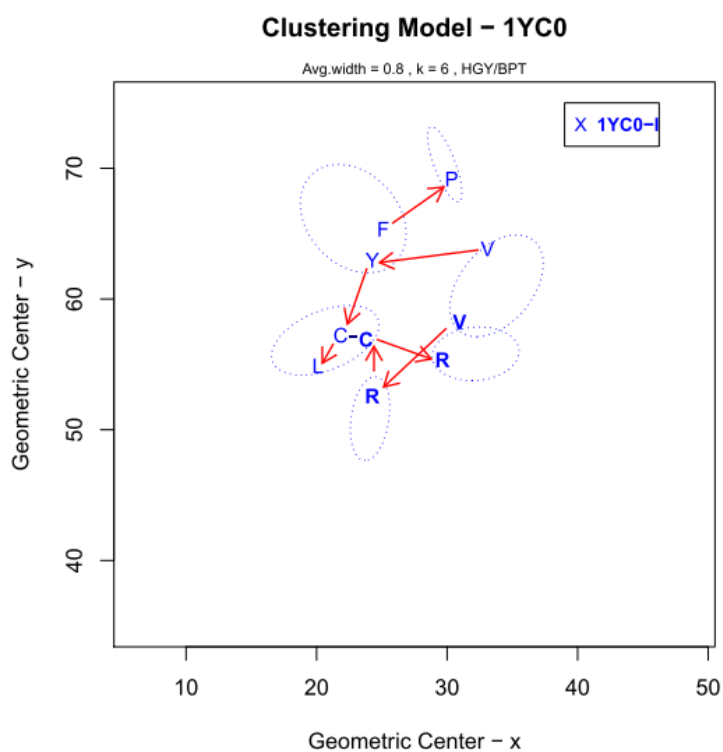
O complexo 1YC0 contém a enzima HGFA complexada com primeiro domínio Kunitz do inibidor HAI-1 (KD1). HGFA (Ativador de fator de crescimento de hepatócitos) é uma serino peptidase, secretada pelo fígado, que converte o fator de crescimento do hepatócito [Shia et al. (2005)]. KD1 é um inibidor da família I2 (Kunitz-BPT).

Figura 8.45: 1T80 e 1YCO: comparação de clusters.



A base de comparação do complexo 1YCO (HGY/BPT) é o 1T80 (CHY/BPT), pois ambos têm o mesmo inibidor. Foi observado que embora com o mesmo tipo de inibidor, os subclusters apresentam algumas diferenças intrigantes (Figura 8.45) que estão relacionadas à presença de uma ponte dissulfeto no inibidor que lhe confere flexibilidade de interações (Cisteínas da região 2).

Figura 8.46: 1YCO: 3 alças desconexas. 1a:V(G)RCR; 2a: VY(GG)CL; 3a: FP.



Isso pode ser melhor entendido a partir da figura 8.46. Vemos que o inibidor BPT na 1YCO não tem uma única alça inibitória “linear”, mas 3 subalças que buscam complementação hidrofóbica: a primeira, formada por $V_{256}(G_{257})R_{258}C_{259}R_{260}$; a segunda por $V_{279}Y_{280}(G_{281}G_{282})C_{283}L_{284}$; a terceira, por $F_{263}P_{264}$. Há algumas glicinas que não aparecem nos modelos por não terem átomos apolares. Percebam, pelas sequências das numerações, que as 3 subalças são de fato descontínuas. Apesar disso, esses 3 subalças buscaram complementaridades hidrofóbicas como se fossem uma só.

Para a interface, pouco importa se os casamentos hidrofóbicos estão vindo de uma alça formada por resíduos contíguos na sequência ou não. O que interessa é o pareamento das hidrofobicidades. Trata-se de mais uma forte evidência de que os inibidores e enzimas buscam complementaridades no nível atômico, mesmo que para isso tenham que resolvê-la em múltiplas alças. É improvável que qualquer algoritmo que operasse apenas com sequências de resíduos dessa conta desse alinhamento não-local.

Na 1T8O, a alça é contínua: $T_{11}(G_{12})P_{13}C_{14}W_{15}A_{16}R_{17}I_{18}I_{19}$. Mas, vejam na figura 8.48 que 1T8O e 1YCO, a despeito dessas diferenças nas continuidades, sobrepõe-se muito bem.

Em termos de sequência, essa sobreposição poderia ficar assim:

1T8O: [**PC**] [**TW**] [**AI**] [**IR**]
 1YCO: [**RC**] [**VR**] [**YF**] [**PV**]

Figura 8.47: 1T8O e 1YCO: estruturas tridimensionais com regiões hidrofóbicas. Inibidor em *sticks*.

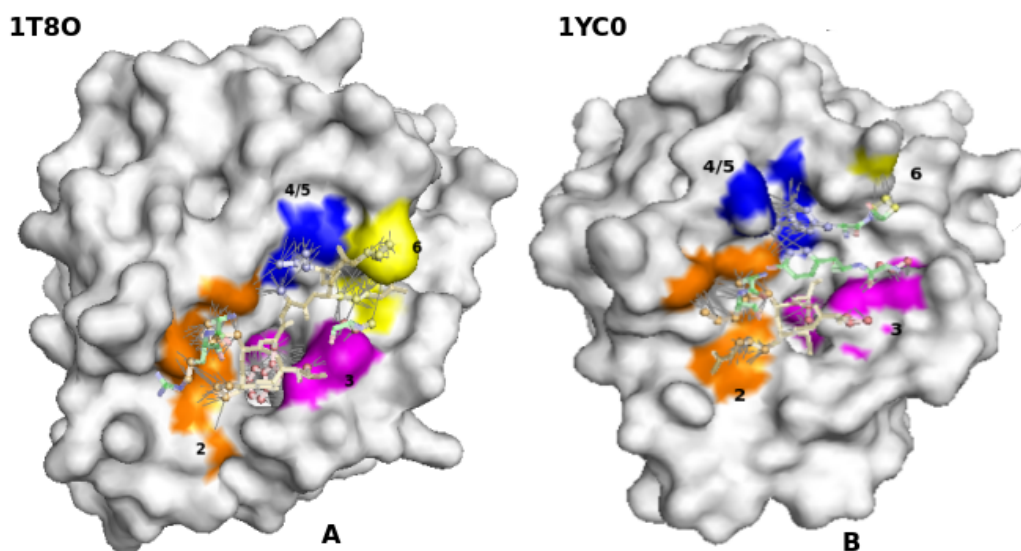
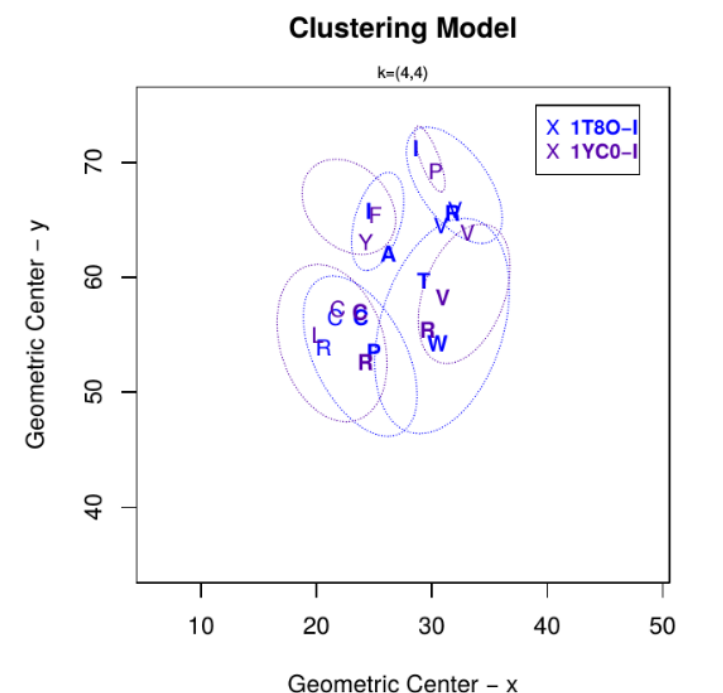


Figura 8.48: 1T8O e 1YCO: comparação de clusters - Inibidores.



8.4.1.15 Complexo 4DG4

Um tipo TRY (mesotripsina) e um inibidor BPT formam o complexo 4DG4. Para este foram caracterizadas 4 regiões hidrofóbicas, conforme ilustradas na Figura 8.49. A região 1 presente em outros complexos, aparentemente não existe e a região 2 se mostra bastante extensa. No entanto, esse resultado obtido está relacionado com a ausência dos contatos da Gly69 da alça com o inibidor com a enzima. Situação também observada para 2Z7F, 1K9O e 3MYW.

O modelo resultante foi comparado com os modelos dos complexos 2FI5 e 1YCO que têm enzimas tipo tripsina e inibidor BPT. Com base nas Figuras 8.50, 8.51 e 8.52 observa-se que o BPT da 4DG4 segue mais o padrão de 2FI5 do que de 1YCO. Com a sobreposição é possível notar que os subgrupos estão bem posicionados para quase todos os resíduos (Figura 8.52). Entretanto, apesar dessa semelhança, há uma divergência na rede de contatos enzima-inibidor dos subclusters 4/5 e 6. Por exemplo, a região 4/5 é composta pelos resíduos **CFY AI** em 4DG4 e **CFK AI** em 2FI5.

Figura 8.49: 4DG4 (Tripsina III-mesotripsina e BPT): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

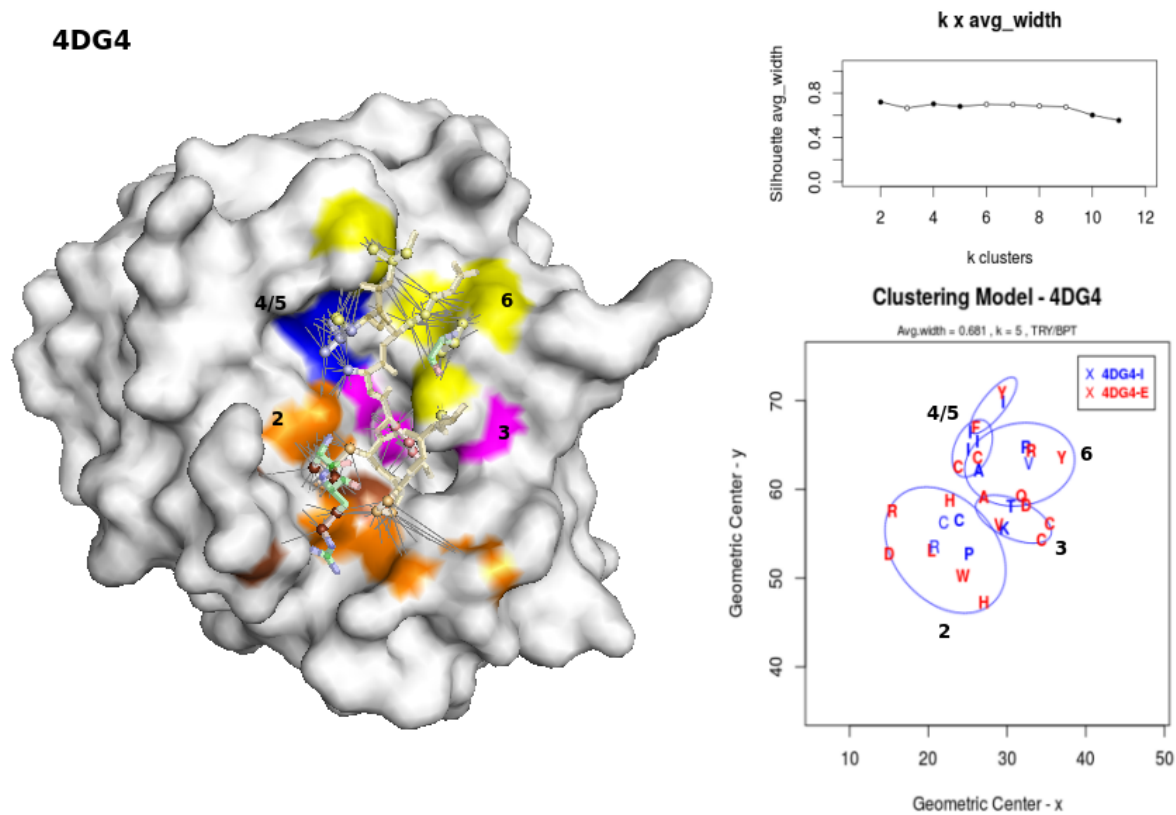


Figura 8.50: 2FI5 e 4DG4: comparação de clusters.

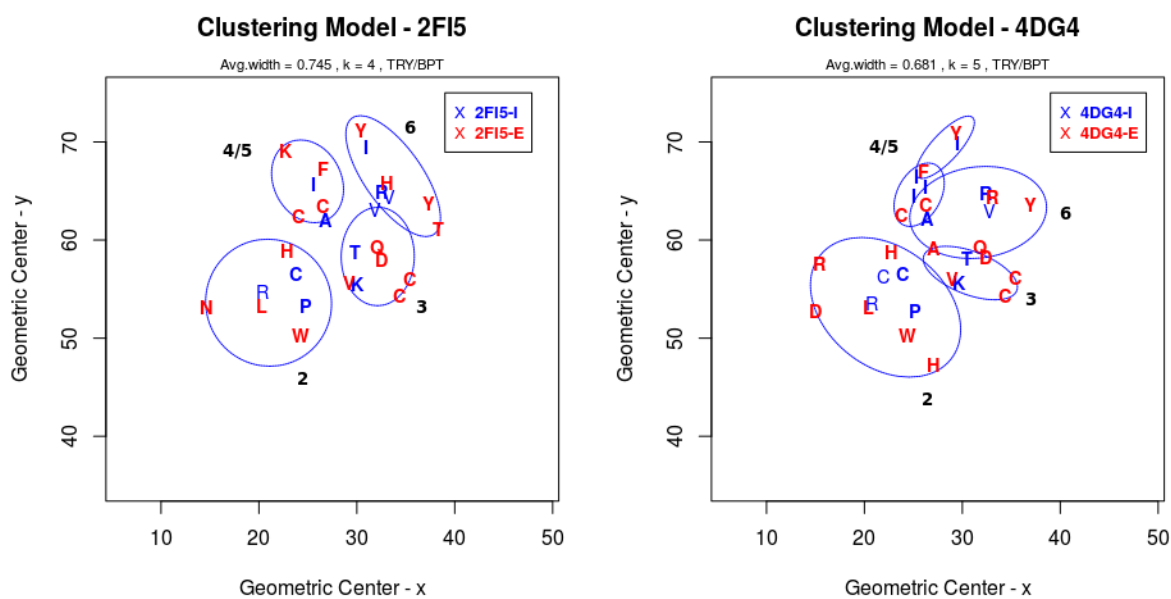


Figura 8.51: 1YCO e 4DG4: comparação de clusters.

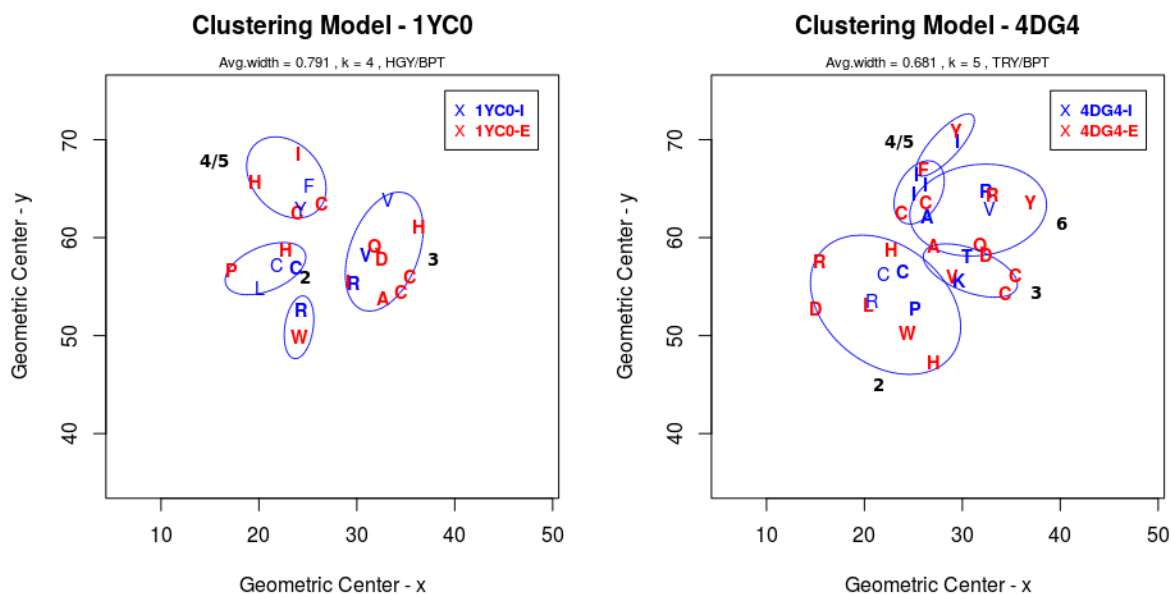
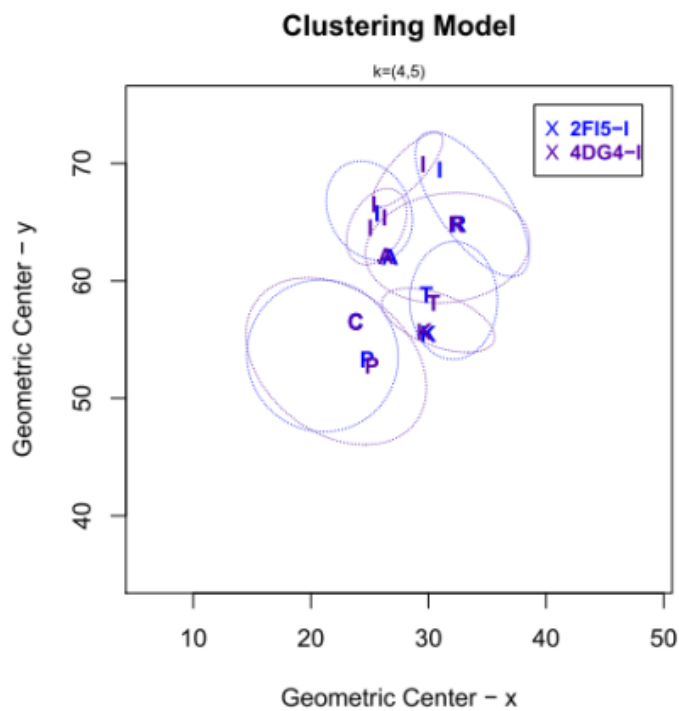
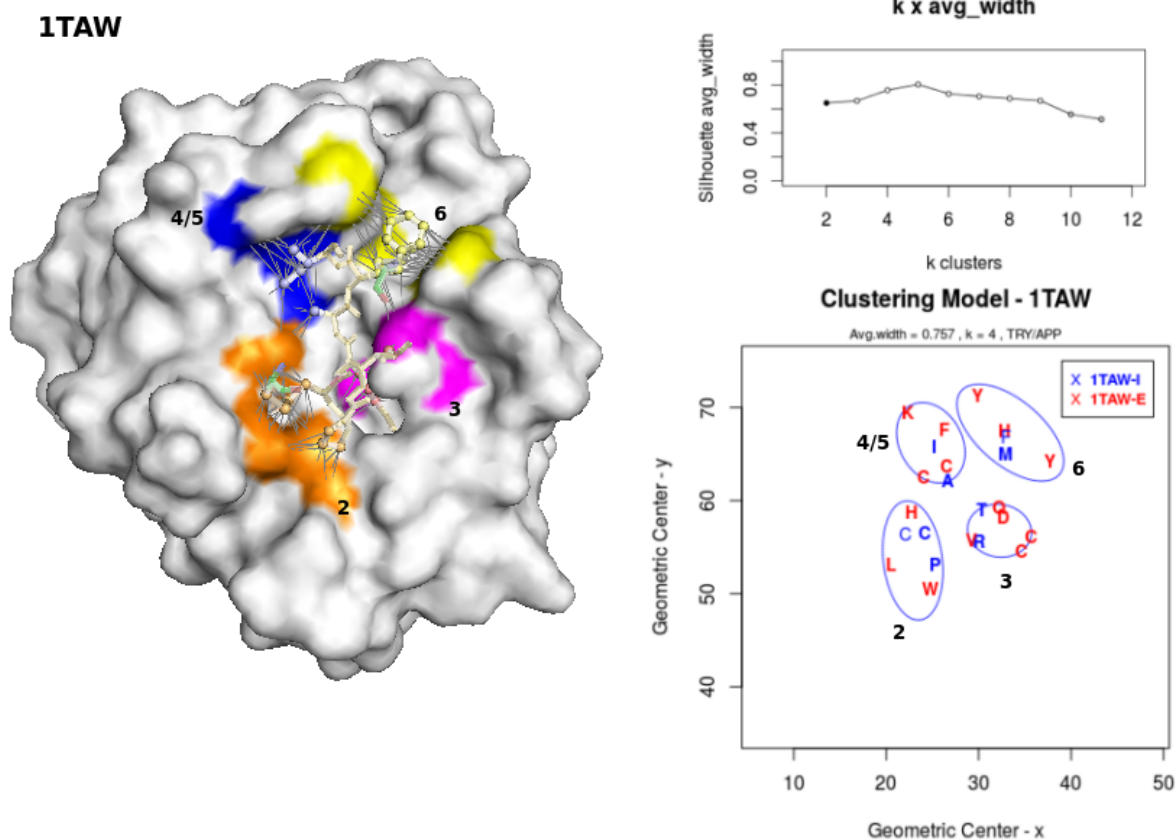


Figura 8.52: 2FI5 e 4DG4: sobreposição de clusters.



8.4.1.16 Complexo 1TAW

Figura 8.53: 1TAW (Tripsina e APP): 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O inibidor do complexo 1TAW é o APP da família I2:Kunitz-BPT. Comparando-se, então o modelo de clusters do complexo 1T8O (CHY/BPT) com o modelo de 1TAW tem-se muita similaridade. (Figura 8.54). Com a sobreposição dos clusters, lado inibidor, isso fica mais claro ainda. (Figura 8.55). Na região, **R** de 1T8O foi substituído por **M** de 1TAW.

Figura 8.54: 1T8O e 1TAW: comparação de clusters.

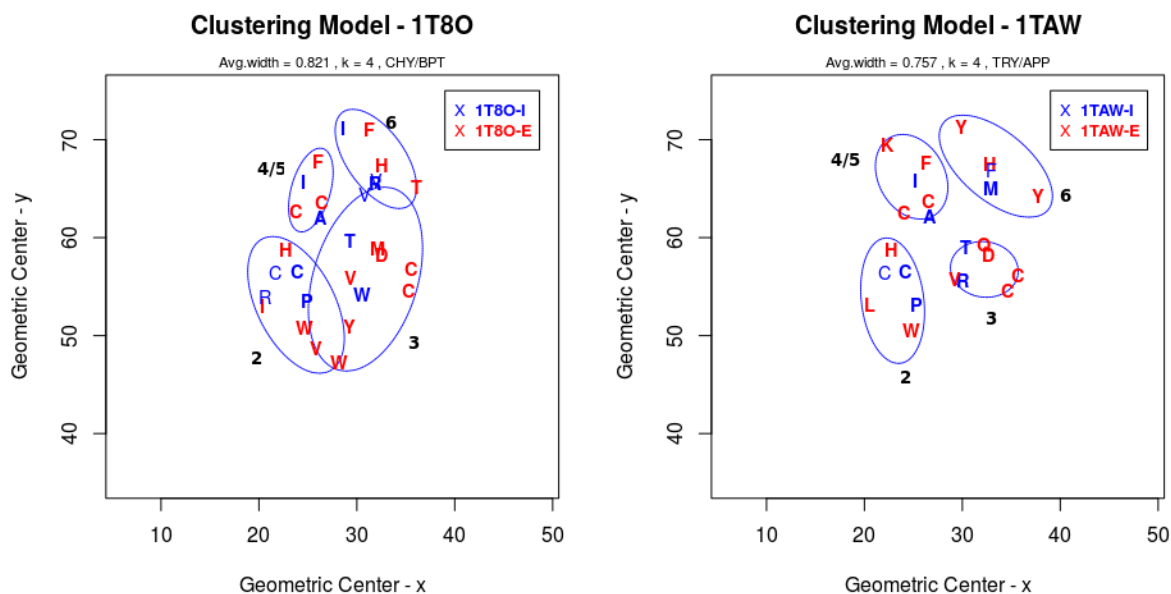
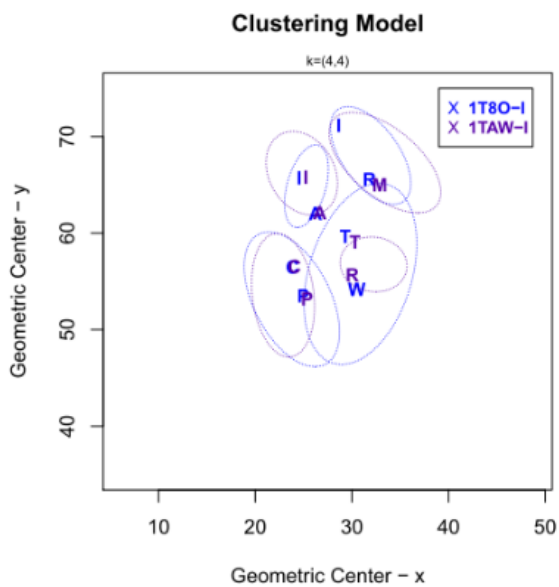
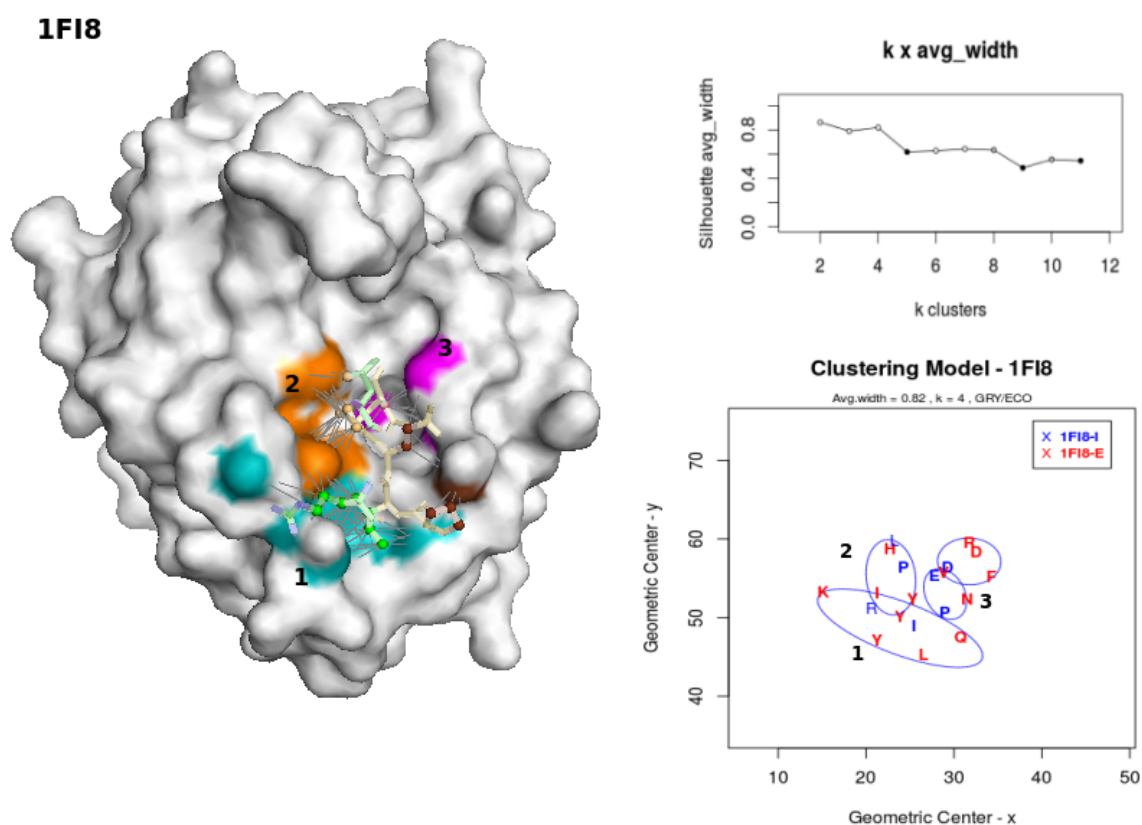


Figura 8.55: 1T8O e 1TAW: sobreposição de clusters.



8.4.1.17 Complexo 1FI8

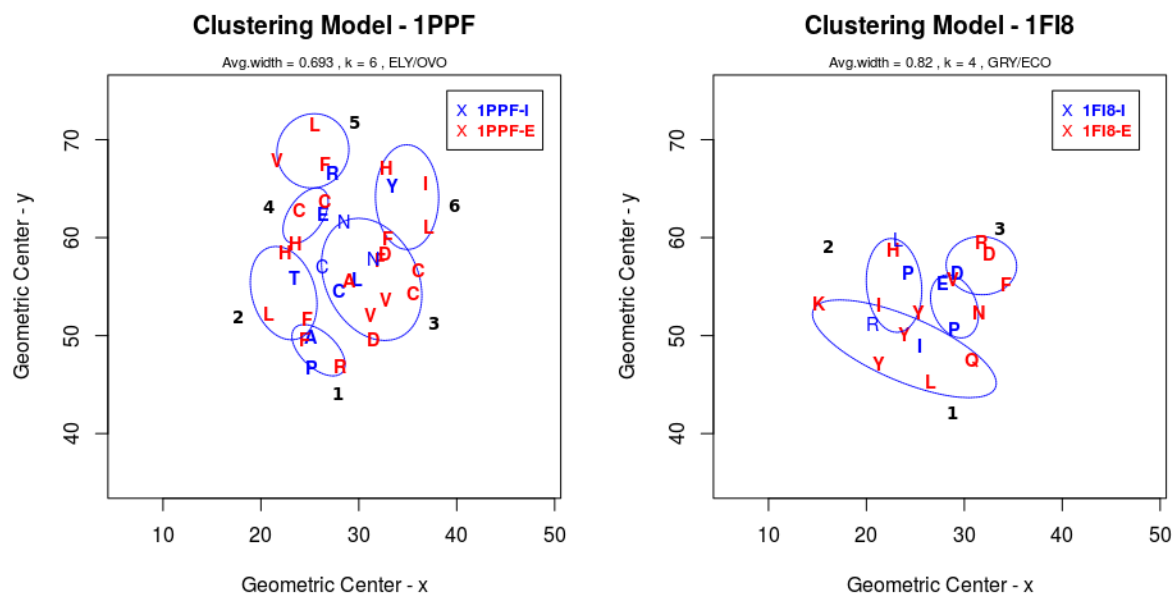
Figura 8.56: 1FI8 (Granzima B e ECO): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



A Granzima B (GRY) é uma serino peptidase com *fold* da Quimotripsina que em conjunto com os linfócitos citotóxicos está relacionada com processos de apoptose da célula. Tem uma estrita preferência por um Asp na posição P1 de seus substratos. Essa especificidade primária pelo Asp ocorre por meio de interações com a Arg226 que tem a cadeia lateral praticamente enterrada [Waugh et al. (2000)]. A ECO nesse complexo é um variante da Ecotina *wild-type*. A alça inibitória é composta nas posições de P4 a P1 por IEPD (81-84). A sequência nativa da alça é PVSTMMACP (P4 : 80 a P3' : 88) [Jin et al. (2005)]. A porção C-terminal da alça não foi resolvida na estrutura (P1': Met85 e P2':Ala86). Na estrutura tridimensional da Figura 8.56, se vê o tamanho reduzido essa alça.

Em razão disso, apenas 3 regiões hidrofóbicas foram encontradas para esse complexo. Além de uma região secundária formada pelo *noloop* Pro80I, Glu82I da alça e Asn218 da enzima (Figura 8.56). Essa região, aparentemente não mostra correspondência com alguma das 6 regiões de 1PPF (8.57).

Figura 8.57: 1PPF e 1FI8S: comparação de clusters



8.4.1.18 Complexo 1EZS

O inibidor ECO de 1EZS é um mutante com as substituições M84R, W67A, G68A, Y69A [Gillmor et al. (2000)]. R84I é parte da alça desse inibidor: *PVSTRMACP* ($P5 : 80$ a $P4' : 88$). Como já foi mencionado acima, a ECO do complexo 1FI8 apresenta na alça somente os resíduos *IEPD*. Essas diferenças são refletidas na formação dos subclusters desses complexos, conforme é mostrado nas Figuras 8.59 e 8.60.

1EZS apresenta as regiões hidrofóbicas complementares (enzima/inibidor) 1, 2, 3, 4/5 e 6, como vários complexos acima apresentados e 1FI8, somente as regiões 1, 2 e 3. Ambos os complexos têm uma sétima região formada por P80I e resíduos da enzima. 1EZS também apresenta regiões hidrofóbicas formadas por *noloops* representas em marrom na Figura 8.58.

Figura 8.58: 1EZS (Tripsina II e ECO): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

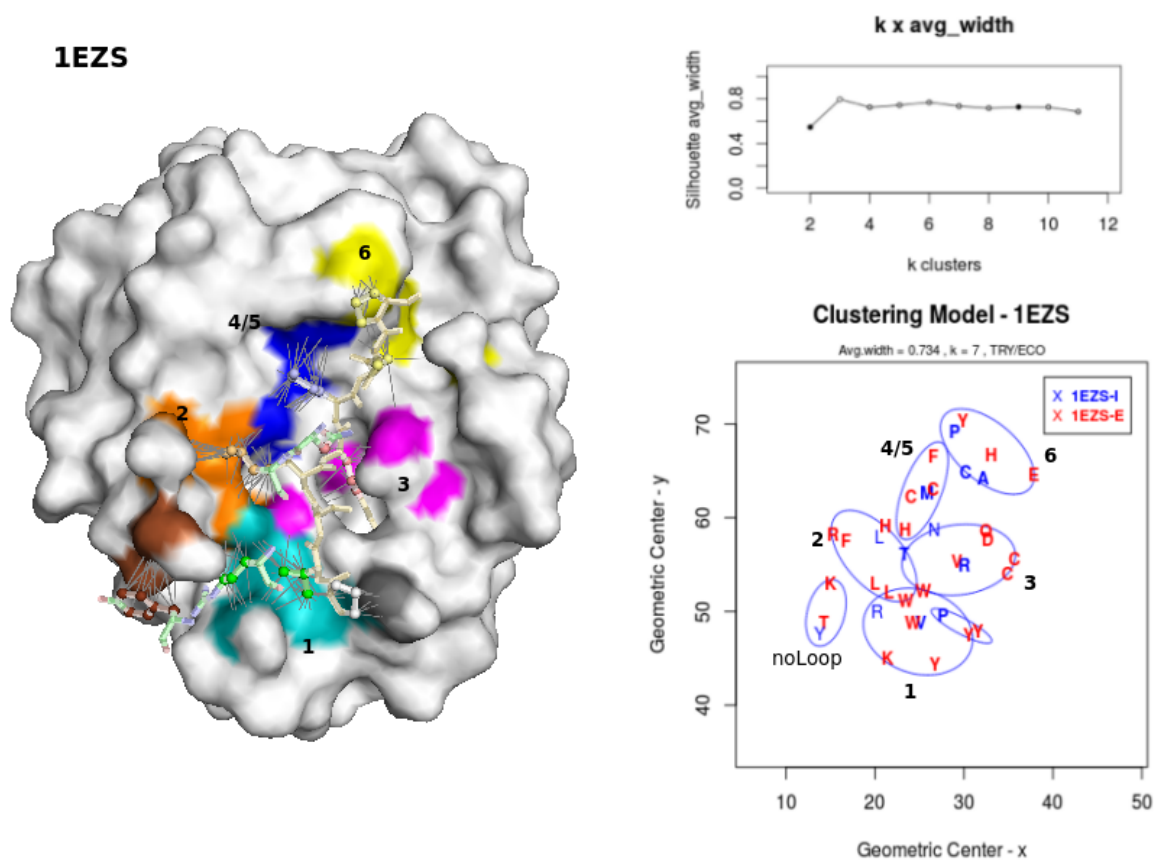


Figura 8.59: 1FI8 e 1EZS: comparação de clusters.

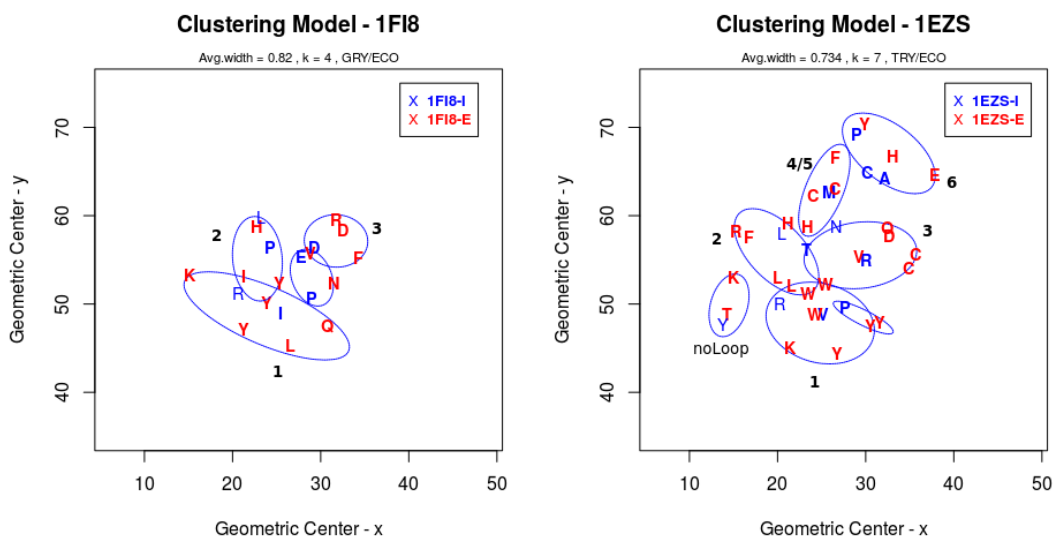
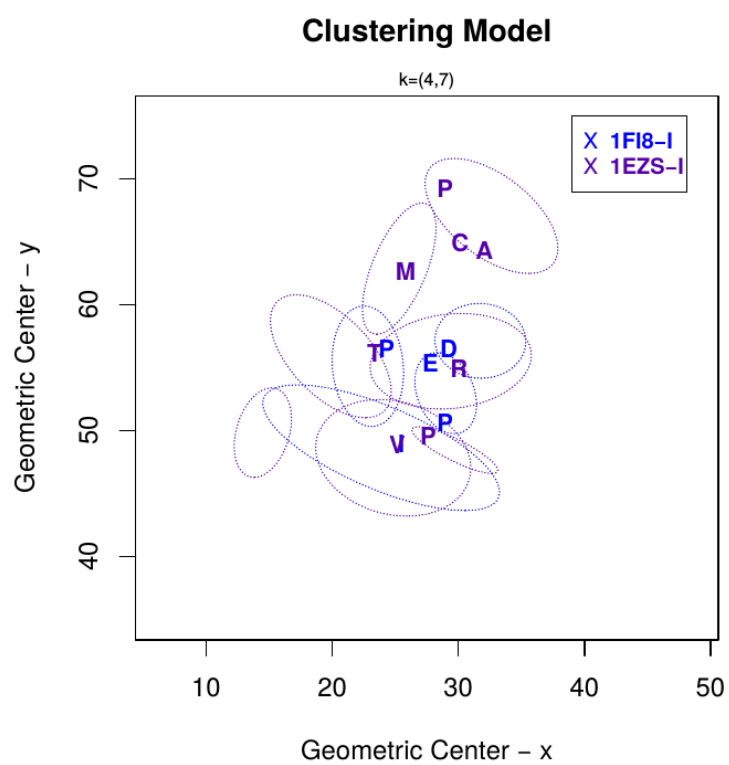
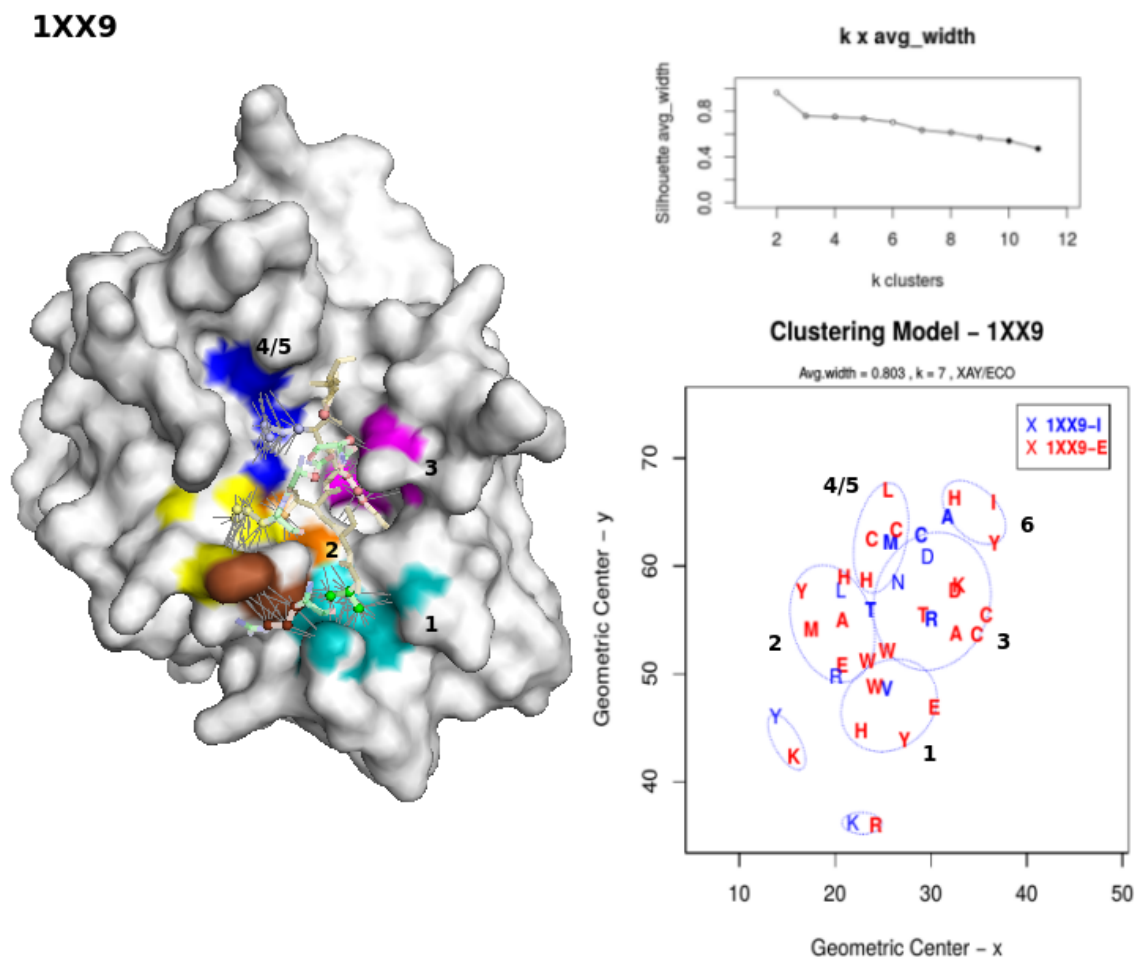


Figura 8.60: 1FI8 e 1EZS: sobreposição de clusters - Inibidores



8.4.1.19 Complexo 1XX9

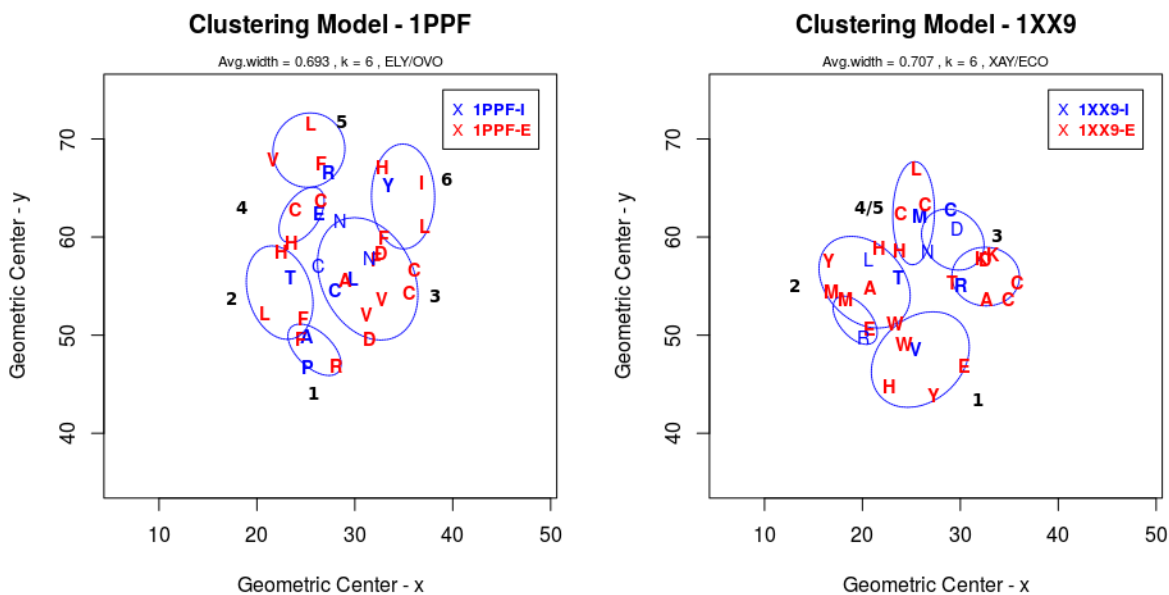
Figura 8.61: 1XX9 (FXIa e ECO): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O Fator XI ativado (FXIa) é uma enzima importante na cascata de coagulação sanguínea. Para a formação do complexo FXIa/ECO, foi realizada a mutação M84R na sequência nativa da ecotina [Jin et al. (2005)], resultando na alça *PVSTRMACP* ($P4 : 80$ a $P4' : 88$) que é a mesma de ECO mutante em 1EZS.

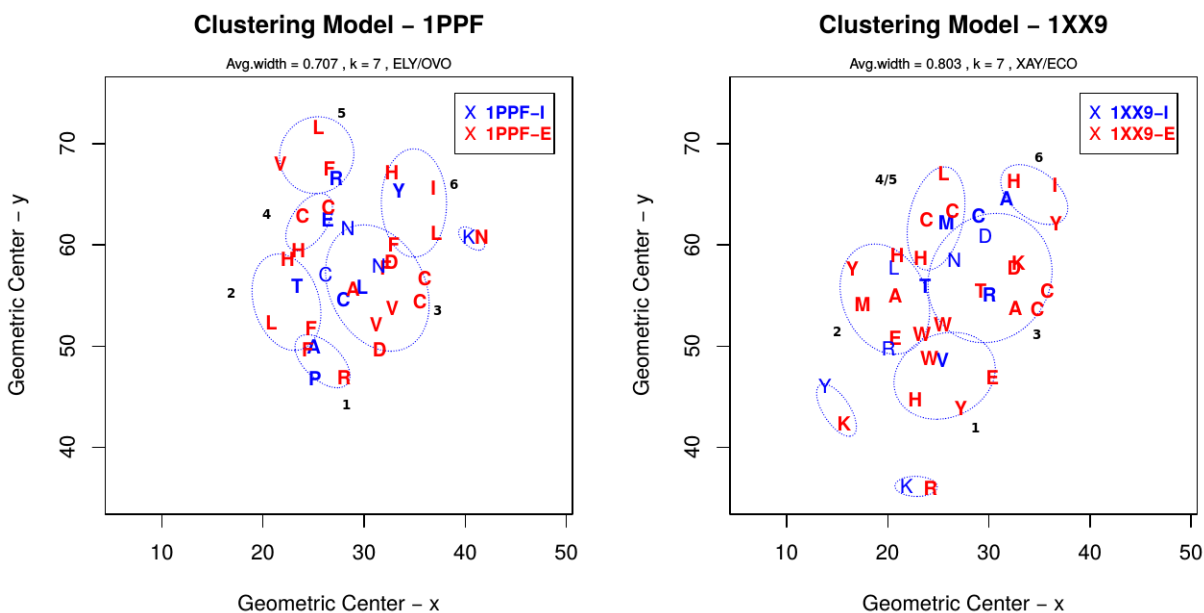
Inicialmente foi realizada uma comparação de 1XX9 com 1PPF, por meio de uma varredura no número de clusters e gerando o modelo de clusters somente para o maior componente conexo. Para $k = 6$, aparentemente 1XX9 não ocupa as regiões 5 e 6 (Figura 8.62).

Figura 8.62: 1PPF e 1XX9: comparação de clusters (k=6).



Quando a exibição é mais detalhada, mostrando também outros elementos não conexos ao maior, obteve-se o modelo abaixo (Figura 8.63):

Figura 8.63: 1PPF e 1XX9: comparação de clusters (k=7).



São vistos 3 elementos conexos menores para 1XX9: dois na parte de baixo da figura (**K-Y** e **R-K**) e o mais significativo na parte superior ocupando exatamente a região 6 **HYI-A**. Portanto, a 1XX9 tem todas as regiões de 1PPF bem representadas. Portanto, por razões de “conectividade”, há um subcluster desconexo na região 6.

1E2S e 1XX9, com inibidores ECO mutantes M84R, têm subclusters cuja sobreposição é muito boa para os resíduos V,T,R,M,A da alça (Figuras 8.64 e 8.65). Também

apresentam regiões compostas por resíduos *noloop*, como pode ser visto para 1XX9 em marrom na Figura 8.61. A região 6 não foi representada nessa estrutura 3D por ser um elemento desconexo.

Figura 8.64: 1PPF e 1XX9: comparação de clusters (k=7).

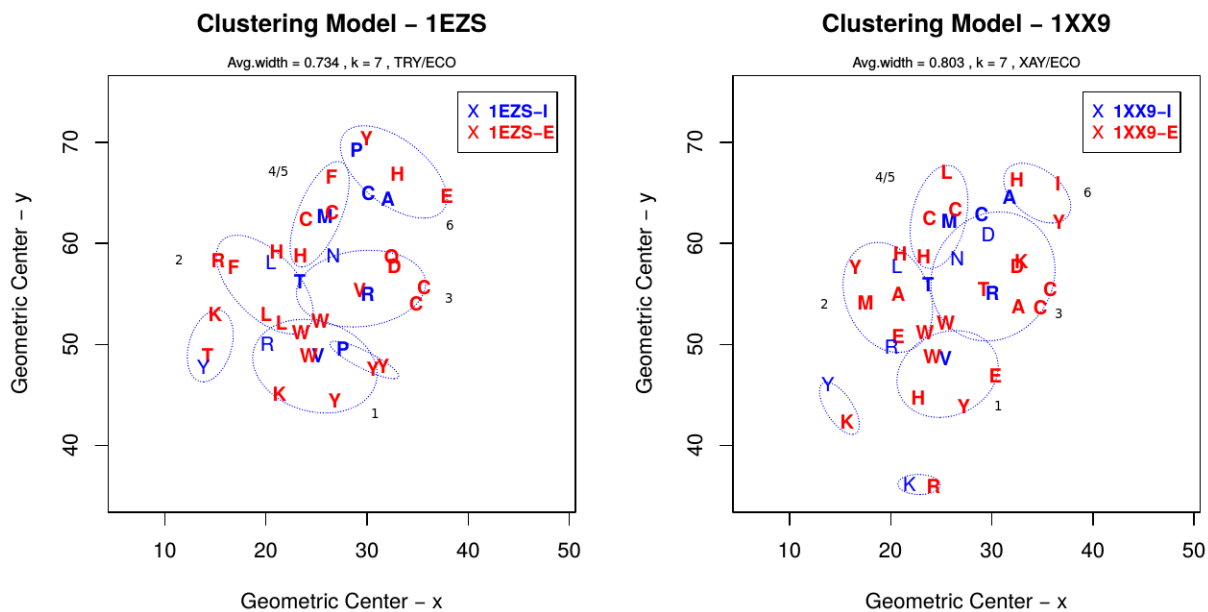
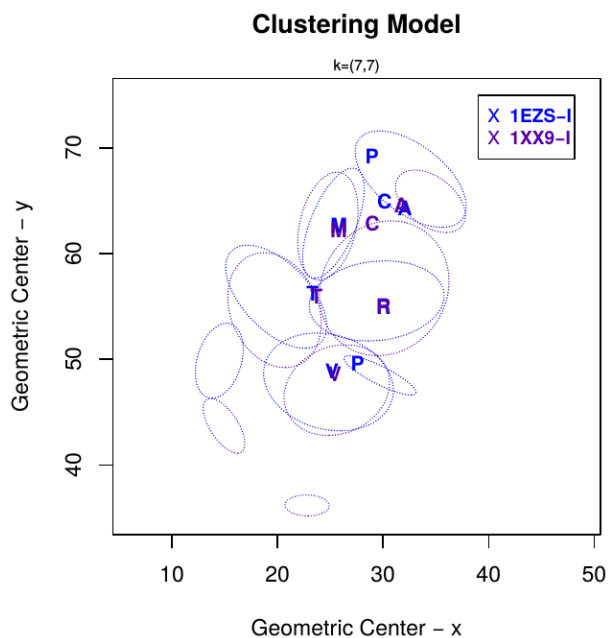


Figura 8.65: 1EZS e 1XX9: sobreposição de clusters - Inibidores.



8.4.1.20 Complexo 1FLE

Figura 8.66: 1FLE (Elastase Pancreática - PPE e Elafin): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters. Na estrutura tridimensional, a região marrom é composta somente por átomos de resíduos *noloop* (Pro29 e Arg31).

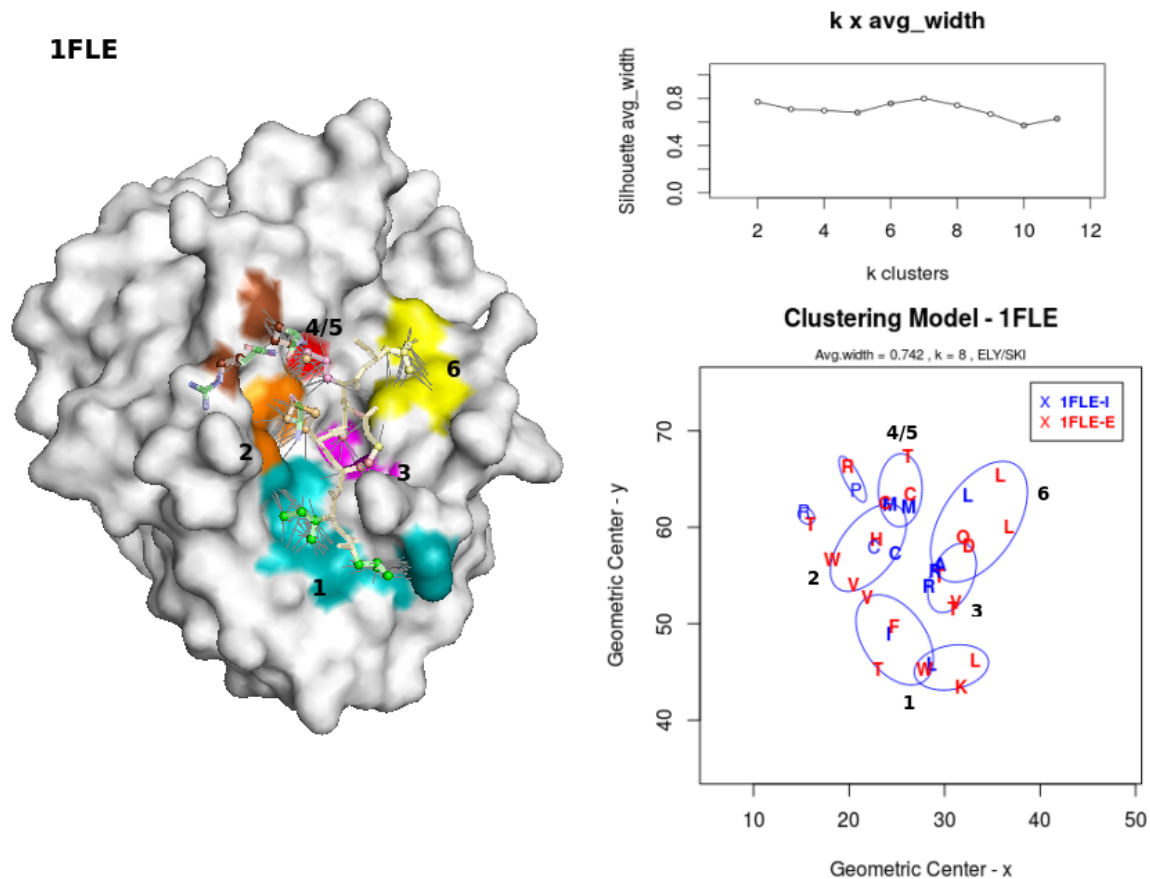
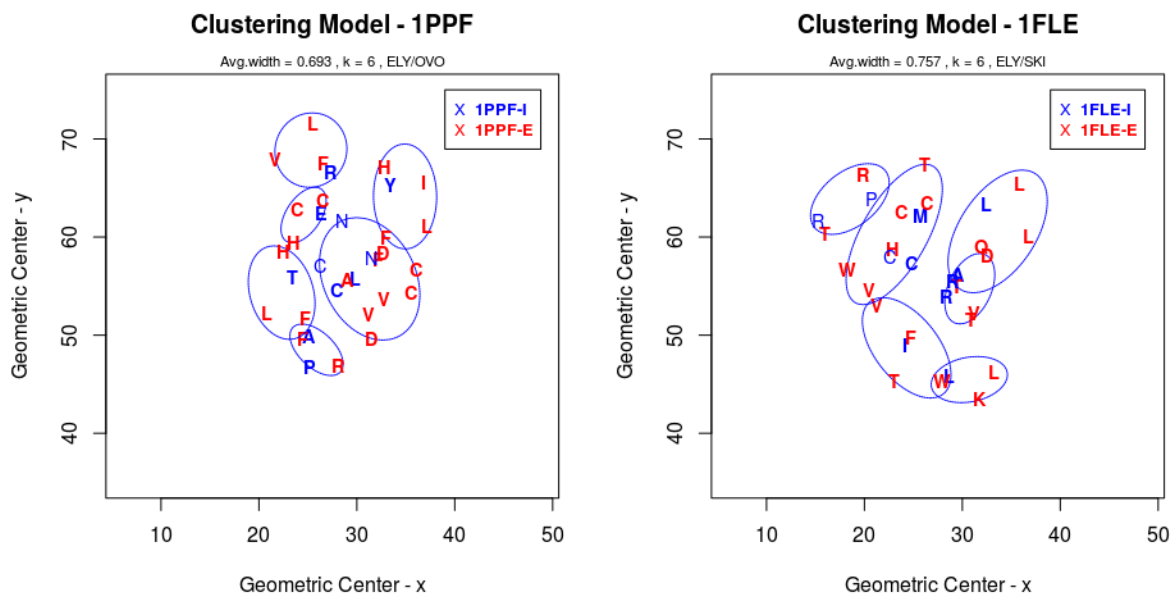
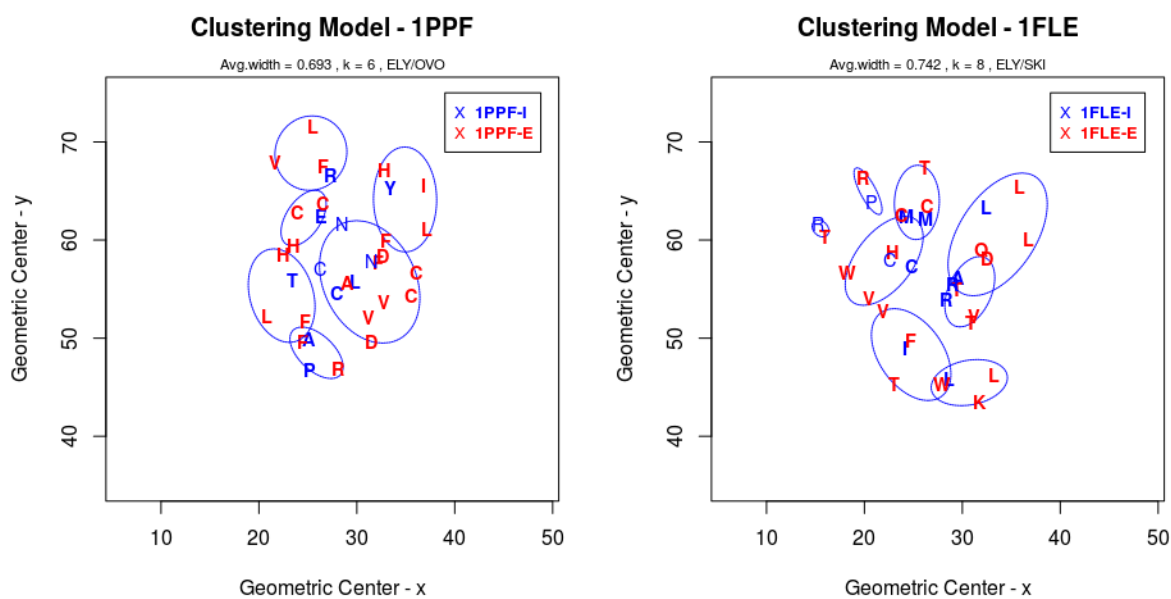


Figura 8.67: 1PPF e 1FLE: comparação de clusters ($k=6$).Figura 8.68: 1PPF e 1FLE: comparação de clusters ($k=6$ e $k=8$, respectivamente).

O inibidor elafin (SKI) de 1PPF, pertencente à família I17 (WAP), é um potente e específico inibidor de elastase pancreática e elastase de leucócitos humanos.

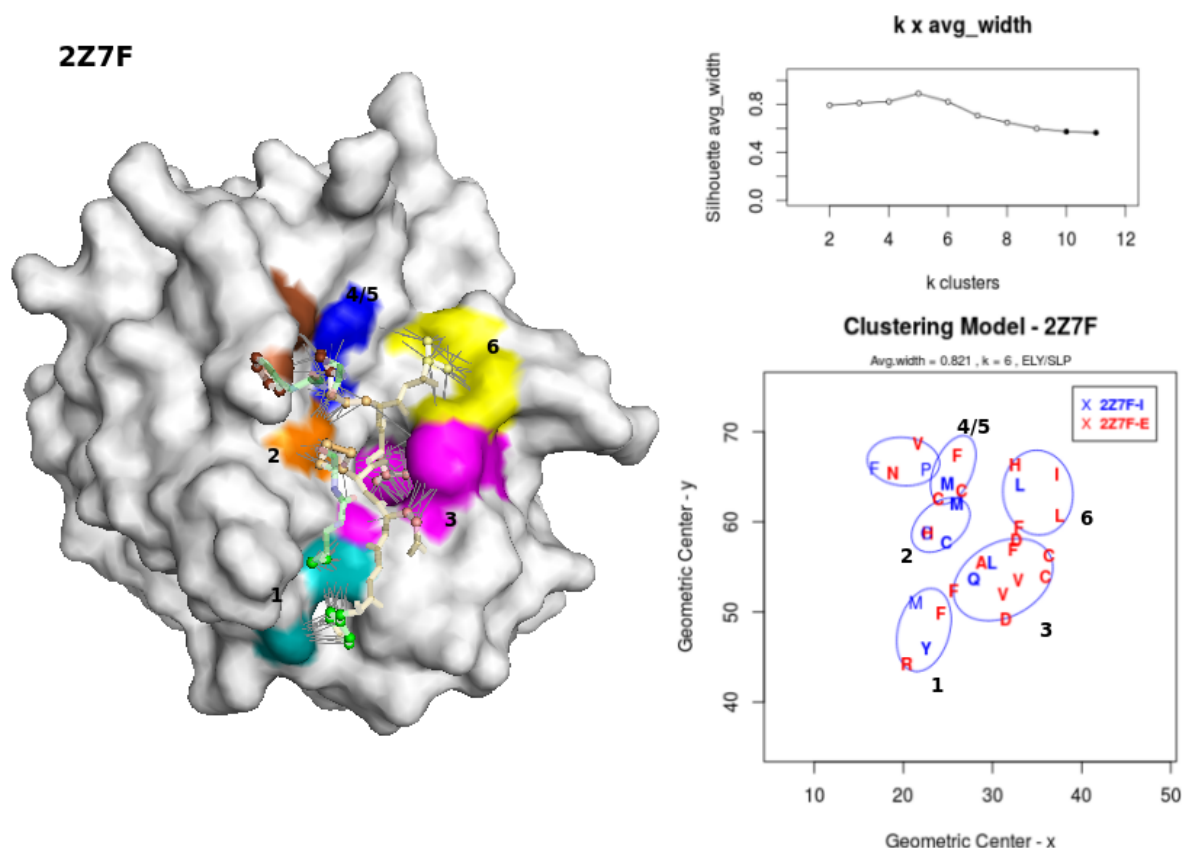
O complexo 1FLE apresenta um padrão muito próximo de 1PPF para os subclusters, apesar de terem inibidores diferentes (SKI e OVO) (Figura 8.68). A enzima de 1FLE é a elastase pancreática e de 1PPF a elastase de leucócitos humanos. Na 1FLE, para $k = 6$, o coeficiente médio de silhueta $s_m = 0.757$ é ligeiramente melhor que para $k = 8$ onde $s_m = 0.742$. Nesses dois casos, os resíduos *noloops* RP são agrupados em

único conjunto quando $k = 6$ e divididos em dois quando $k = 8$.

Conclui-se disso, que subclusters noloop, mesmo que integrem um único elemento conexo, podem interferir no avg.w “ótimo” (coeficiente de silhueta) deslocando o k para valores mais altos (por exemplo, $k=6$). Quando se está usando o coeficiente de silhueta como métrica de qualidade em cluster, isso deve ser feito com a devida cautela. Por isso nesta tese, priorizamos olhar mais o conjunto desses coeficientes, quando se faz a varredura dos agrupamentos em k , do que enfatizar apenas o cluster com o melhor k .

8.4.1.21 Complexo 2Z7F

Figura 8.69: 2Z7F (Elastase de leucócitos humanos e 1/2SLPi): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O complexo 2Z7F é composto de uma SLY, a Elastase de leucócitos humanos, inibida por 1/2SLPi. Este inibidor é um recombinante de metade do domínio C-terminal de SLPI ou SLP e pertence à família I17 (WAP).

Na análise das regiões hidrofóbicas de 2Z7F (Figura 8.70), foi notada a presença de um cluster secundário formado pelos *noloops* FP. Essa região está representada em

marrom na Figura 8.69. Os clusters 1 e 2 no modelo de visualização estão mais distantes um do outro.

A sobreposição dos subclusters de 2Z7F (ELY/SLP) e 1PPF (ELY/OVO), como já era esperado, apresenta dissimilaridades quanto à composição das regiões, pois os inibidores têm alças mais distintas (Figura 8.71).

Figura 8.70: 1PPF e 2Z7F: comparação de clusters.

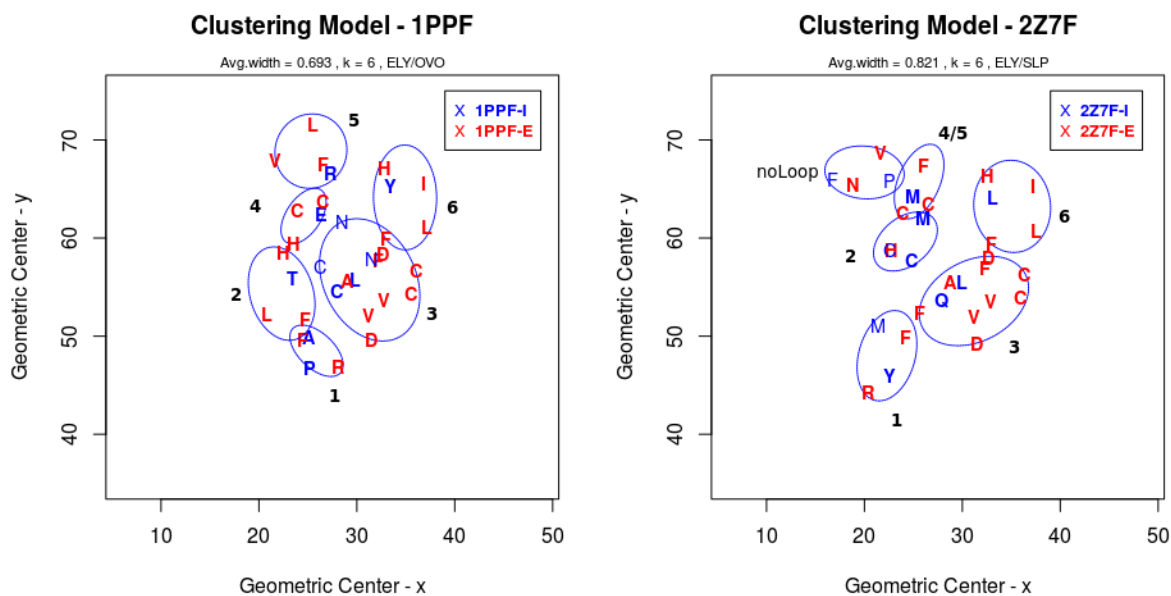
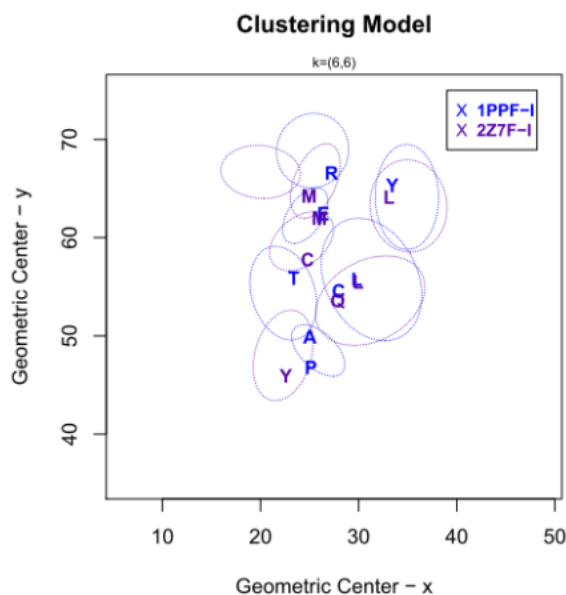
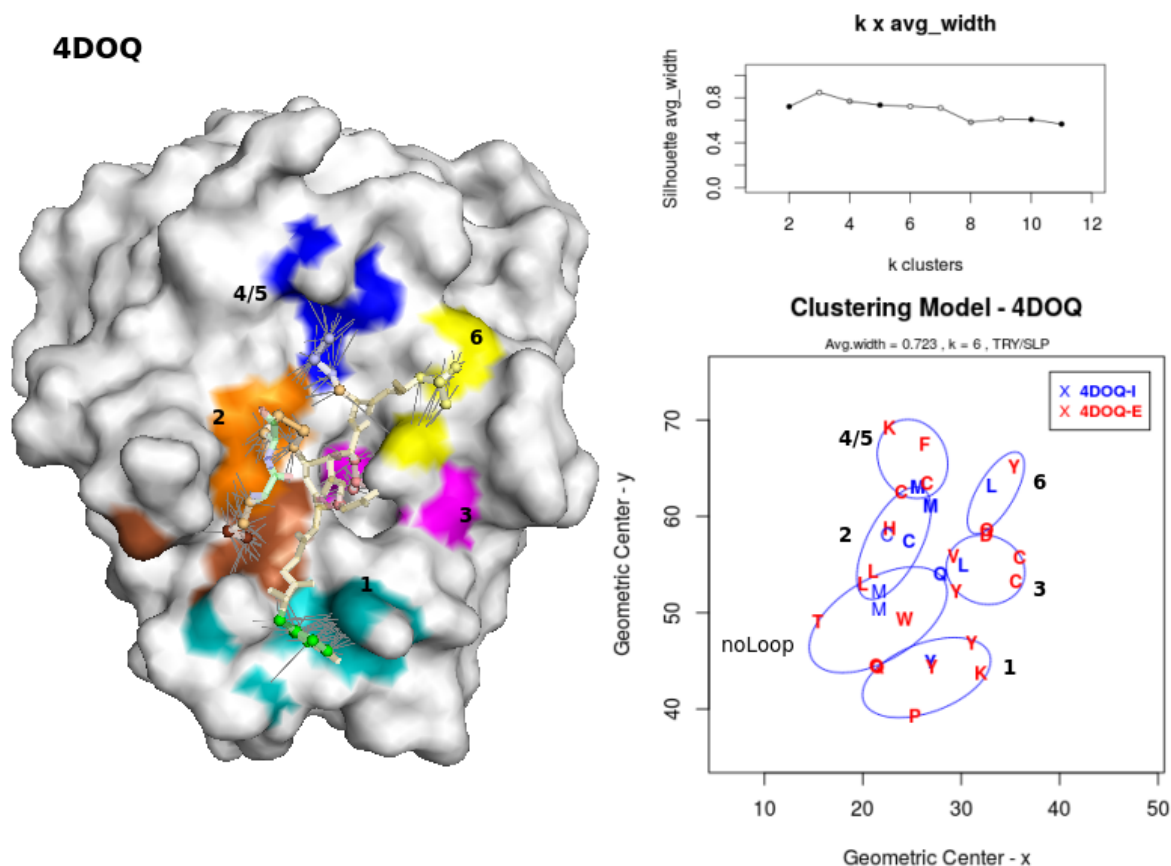


Figura 8.71: 1PPF e 2Z7F: sobreposição de clusters - Inibidores.



8.4.1.22 Complexo 4DOQ

Figura 8.72: 4DOQ (Tripsina e 1/2SLPi): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



Os agrupamentos do complexo 4DOQ (TRY/SLP) foram comparados e sobrepostos com 2Z7F (ELY/SLP), pois ambos os complexos têm o mesmo inibidor 1/2SLPi (Figuras 8.73 e 8.74-A) e também sobrepostos com 1PPF (Figura 8.74-B). Outra comparação realizada foi de 4DOQ com 1FLE (ELY/SKI), porque SLP e SKI são da mesma família I17 (clan IP) com 42% de homologia entre eles [Tsunemi et al. (1996)].

Na comparação com 1PPF, em termos de regiões existentes, os subgrupos são muito similares e a maioria dos resíduos da alça estão bem sobrepostos. A exceção se dá para Y em relação a PA no subcluster 1.

Figura 8.73: 2Z7F e 4DOQ: comparação de clusters.

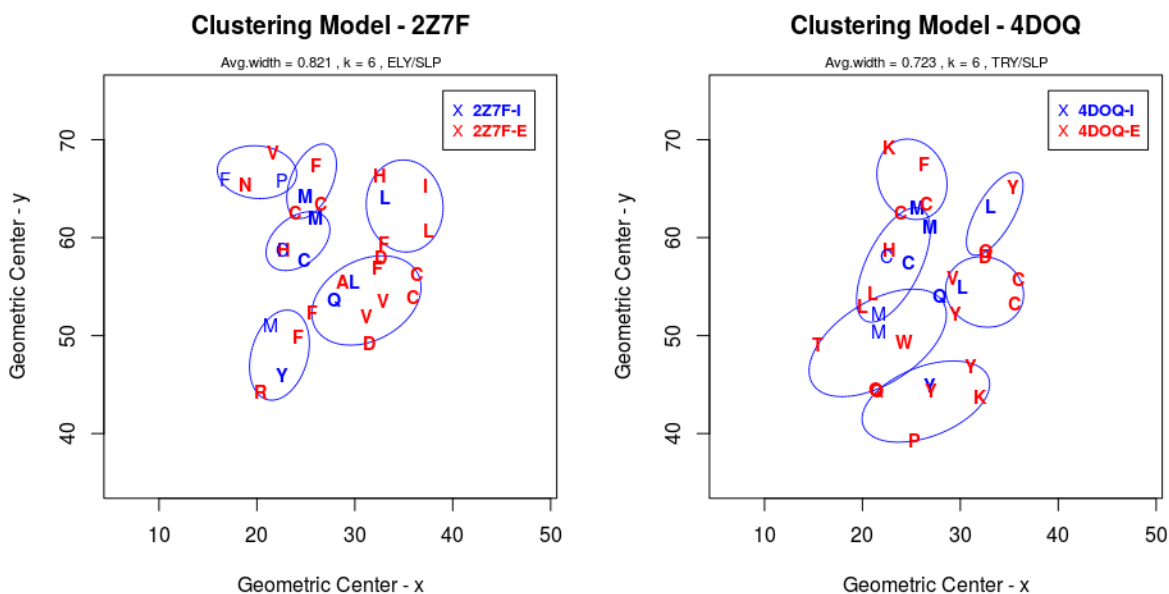


Figura 8.74: Sobreposição de clusters - Inibidores: (A) 2Z7F e 4DOQ, (B) 1PPF e 4DOQ.

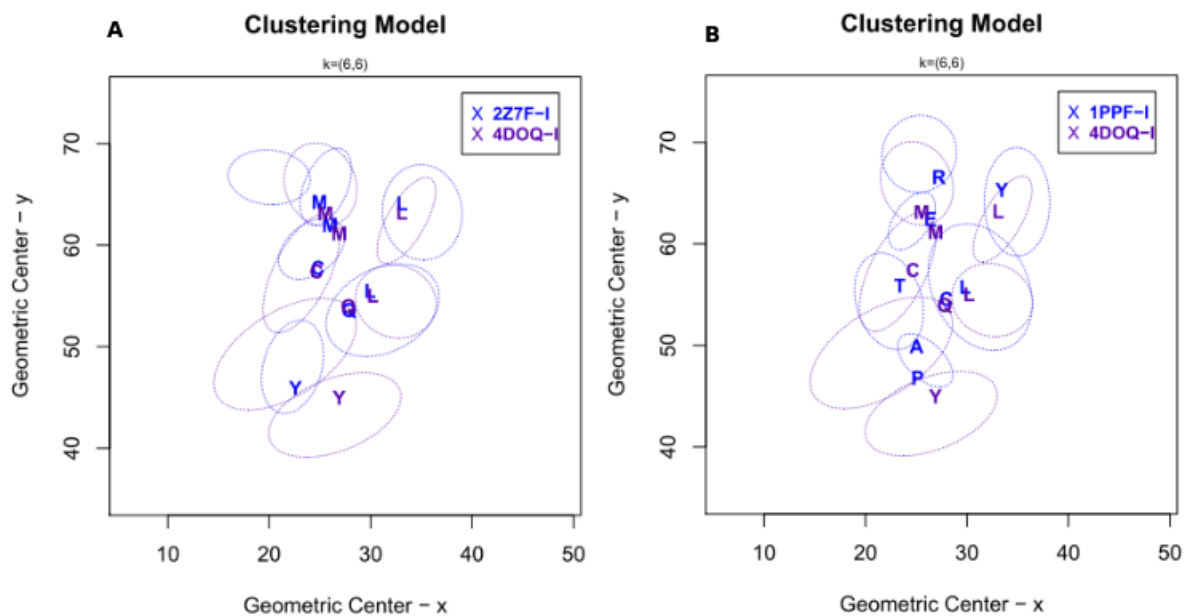


Figura 8.75: 1FLE e 4DOQ: comparação de clusters.

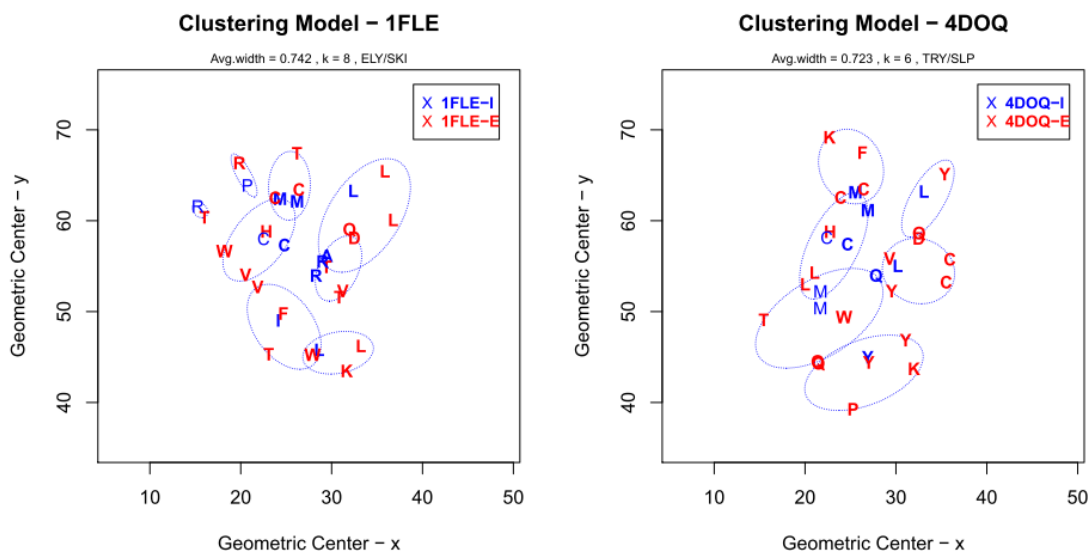
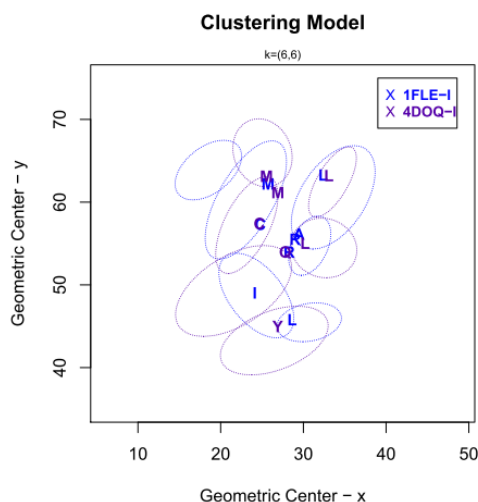


Figura 8.76: 1FLE e 4DOQ: sobreposição de clusters dos inibidores.



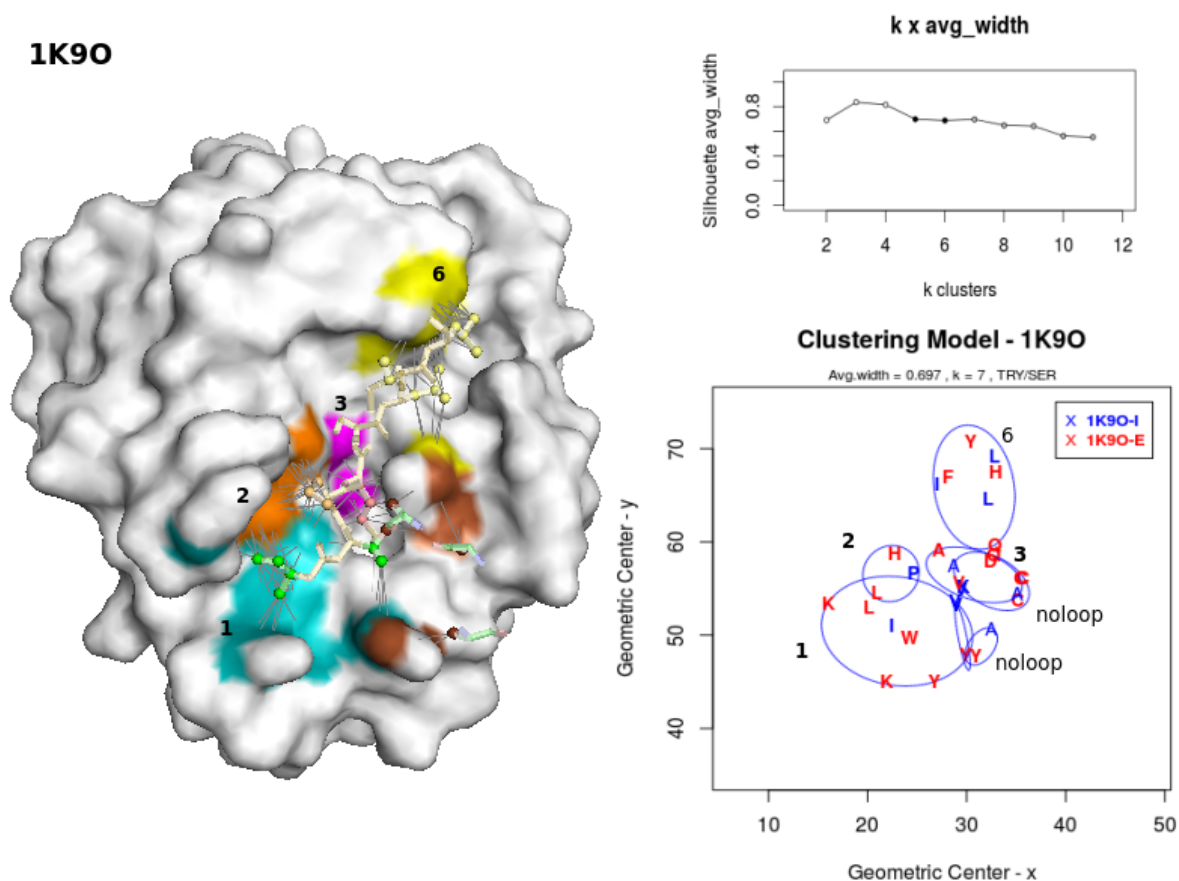
Entre 4DOQ e 1FLE, ocorrem boas sobreposições para praticamente todos os resíduos (Figuras 8.75 e 8.76). O Y da região 1 está mais sobreposto com o L de 1FLE do que com o PA de 1PPF. Apesar das diferenças conformacionais, isso é compreensível, pois SKI e SLP são inibidores da mesma família (I17, clan IP) e OVO é de outro clan (IA). As alças de 4DOQ e 2Z7F são iguais e têm resíduos em comum com 1FLE:

Tabela 8.5: Alça dos inibidores dos complexos 1FLE,4DOQ, 2Z7F

Complexo/inibidor	Alça
1FLE/SKI	LIRCAML
4DOQ/SLP	YGQCLML
2Z7F/SLP	YGQCLML

8.4.1.23 Complexo 1K90

Figura 8.77: 1K90 (Tripsina e ALASERPIN): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O inibidor Alaserpin da família Serpina (I4) explora parcialmente a região hidrofóbica 4/5 (**ER** presente em 1PPF (Figuras 8.78 e 8.79)). Pela sobreposição, percebemos que o subcluster 6 na 1K90 parece localizar-se entre as duas regiões 5 e 6 de 1PPF, caracterizando-se uma região intermediária entre essas regiões do complexo de referência.

Figura 8.78: 1PPF e 1K9O: comparação de clusters.

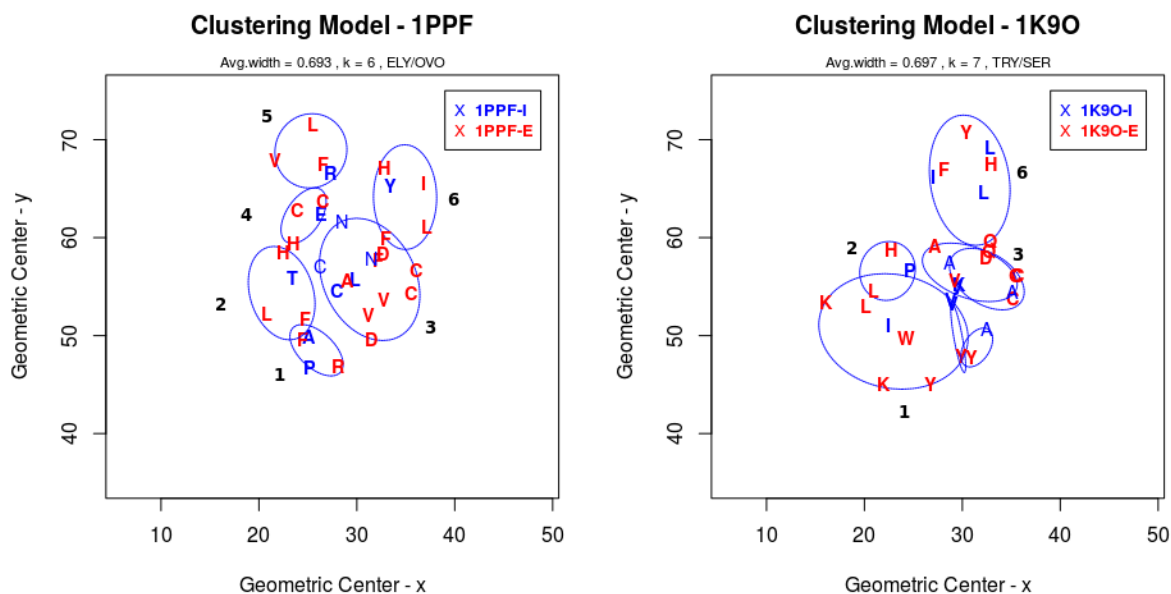
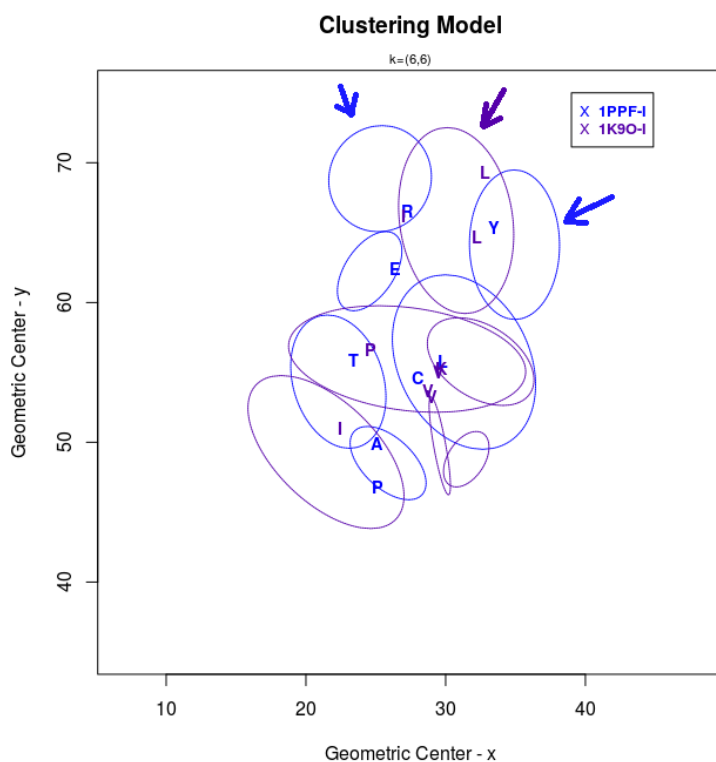


Figura 8.79: 1PPF e 1K9O: sobreposição de clusters contendo apenas resíduos dos inibidores. Destaque para o cluster da 1K9O que se interpõe a dois clusters da 1PPF.



8.4.1.24 Complexo 1OPH

O complexo 1OPH (TRY/SER) tem na sua composição uma tripsina e o inibidor α 1-antitripsina da família das serpinas. Ao compará-lo com a referência 1PPF (ELY/OVO) é evidente na visualização que a região 4/5 não está presente (Figura 8.81). A região 6 de 1OPH parece ocupar uma posição intermediária entre as regiões 5 e 6 de 1PPF, assim como foi notado para 1K9O (TRY/SER).

Figura 8.80: 1OPH (Tripsina e α 1-antitripsina): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

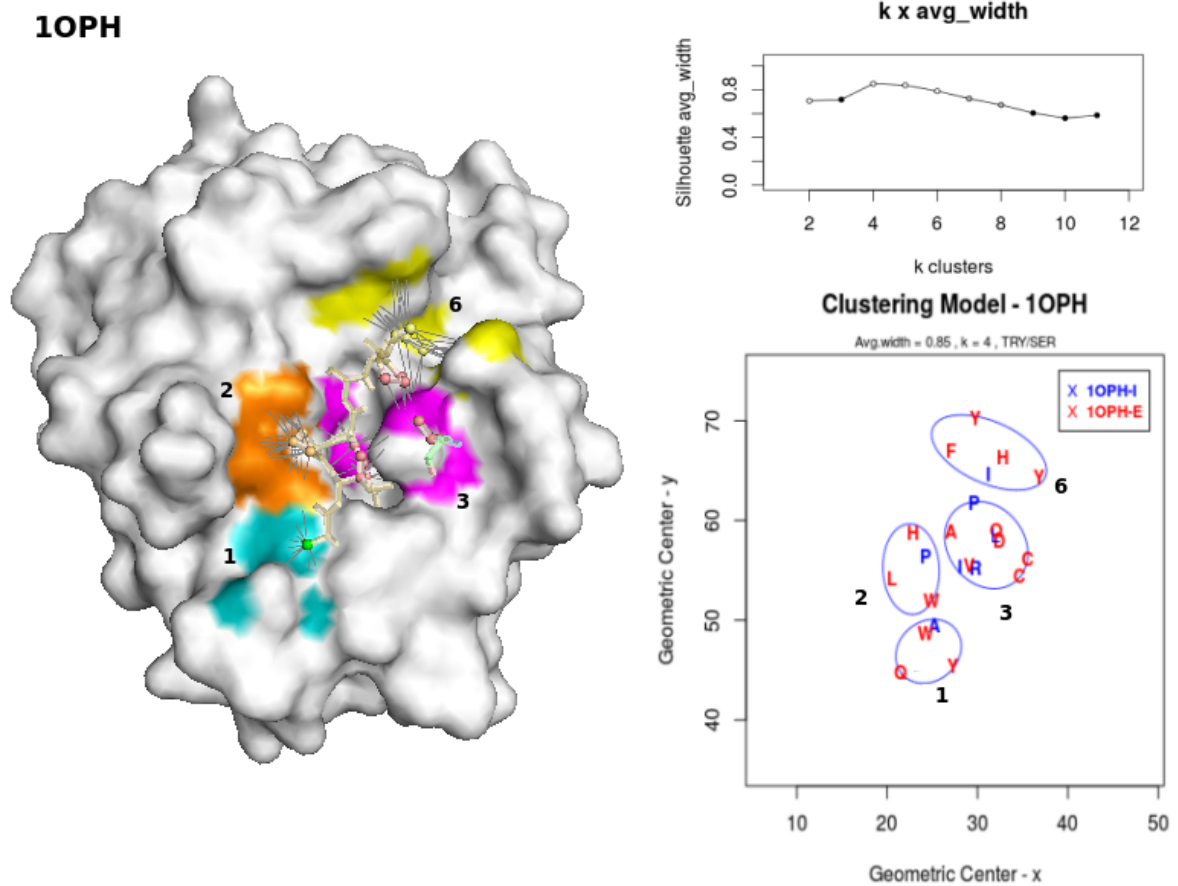


Figura 8.81: 1PPF e 1OPH: comparação de clusters.



1K9O e 1OPH apresentam algumas diferenças em termos de posicionamento dos clusters e de composição (Figura 8.82). Os inibidores mesmo sendo serpinas não têm alças idênticas. Na sobreposição (Figura 8.83) há correspondência entre P-P (região 2), R-K (região 3), e I-L (região 4).

Figura 8.82: 1K9O e 1OPH: comparação de clusters.

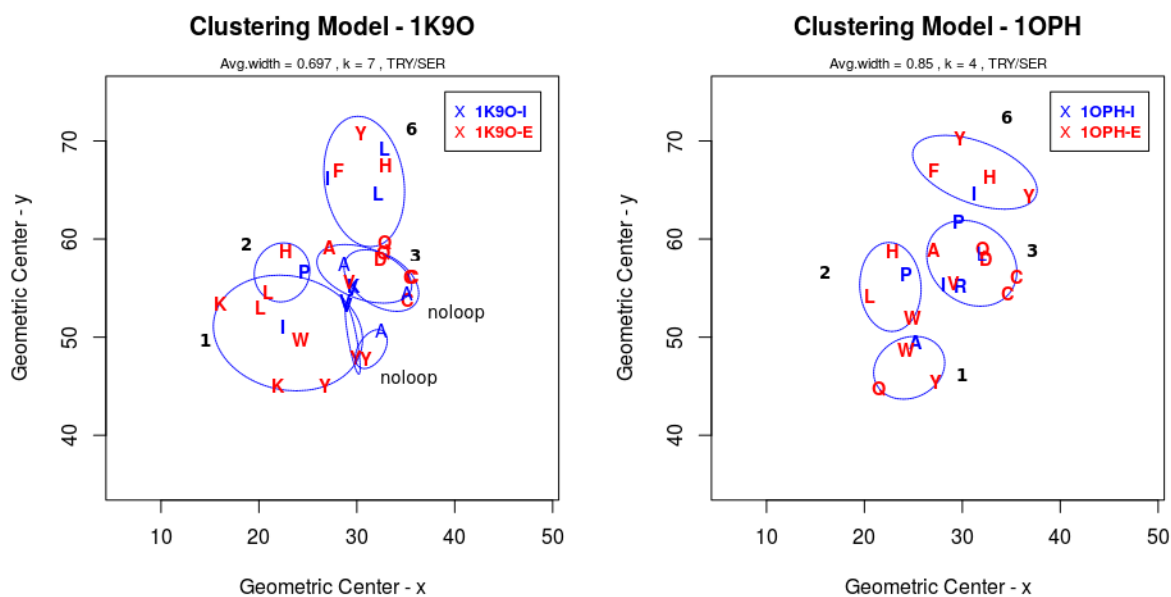
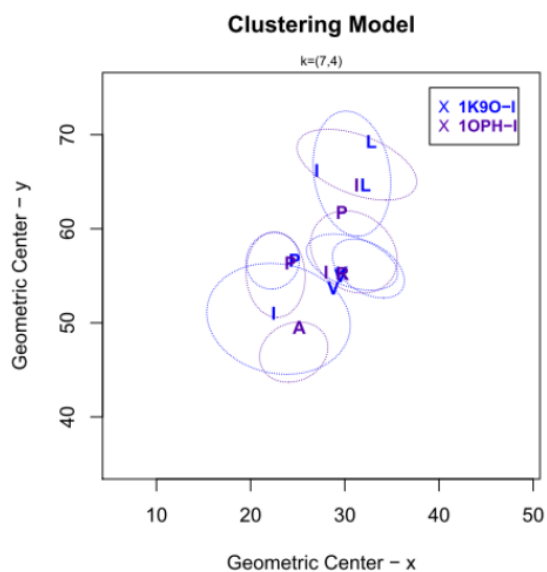
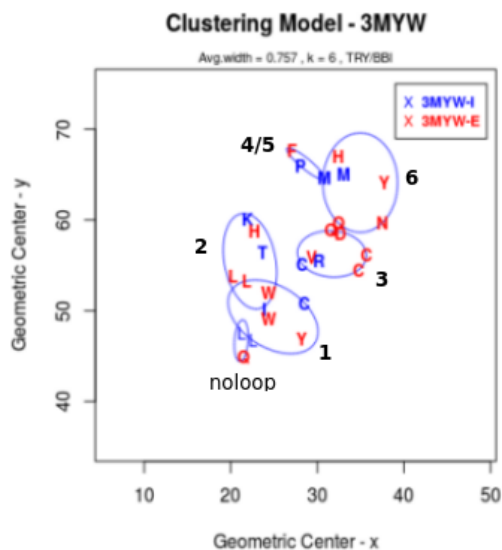
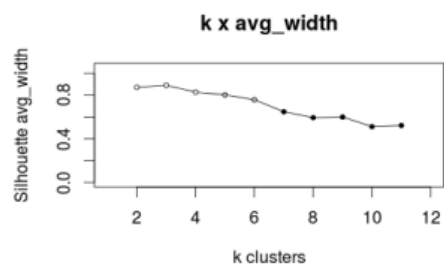
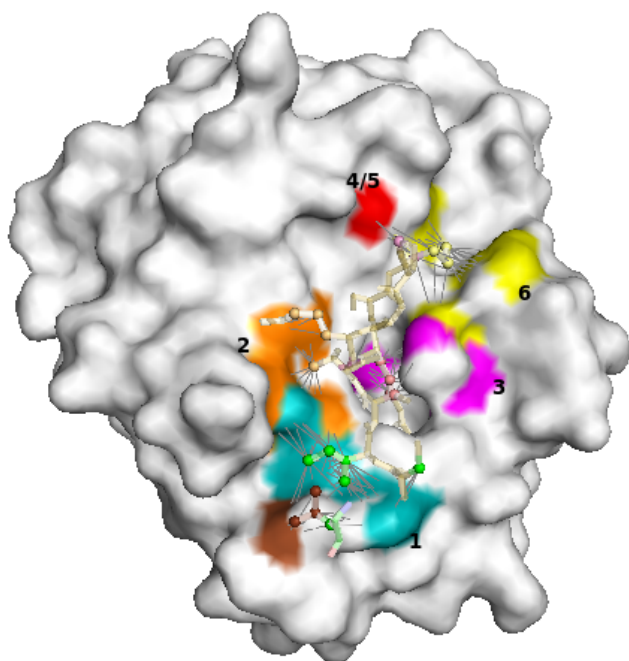


Figura 8.83: 1K9O e 1OPH: sobreposição de clusters.



8.4.1.25 Complexo 3MYW

Figura 8.84: 3MYW (Tripsina e BBI): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

3MYM

O complexo 3MYM (TRY/BBI) apresenta a região 4/5, para $k = 6$, aparentemente formada por um único resíduo do inibidor - **P** e um único da enzima - **F** e com **M** na fronteira entre as regiões 4/5 e 6, sugerindo melhor agrupamento uma fusão das duas (Figura 8.87).

Na sobreposição de 3MYM com 1PPF, além da ausência da região 4 por motivos já mencionados, há as correspondências T-T (região 2), C-C, I-R (região 3), R-P (região 4/5) e (Y-M) (região 6) (Figuras 8.85 e 8.86).

Figura 8.85: 1PPF e 3MYM: comparação de clusters.

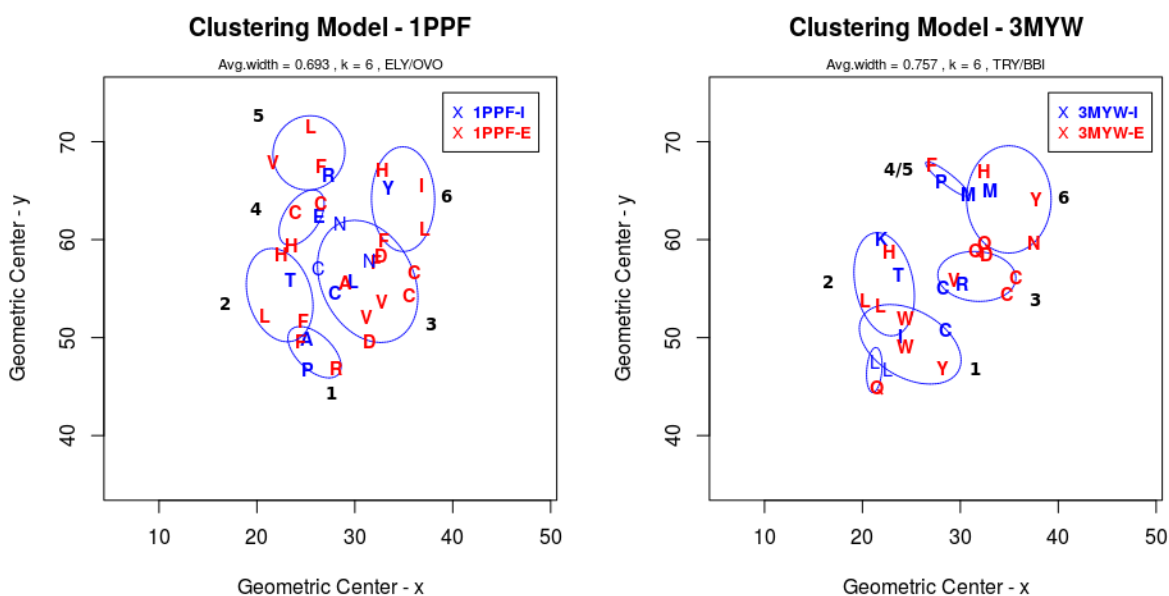
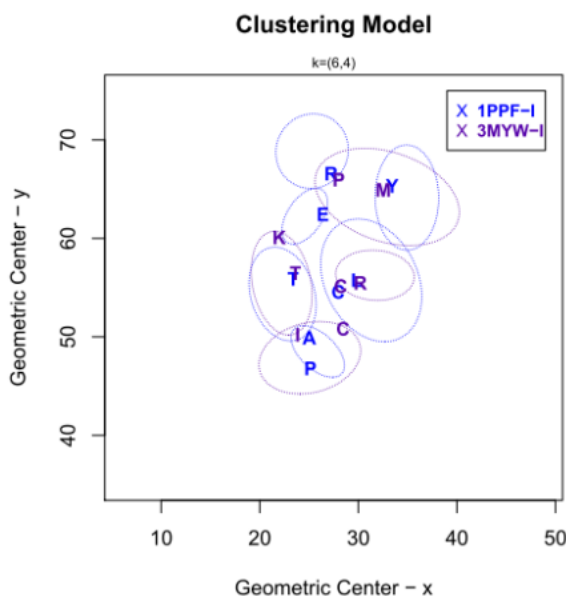


Figura 8.86: 1PPF e 3MYM: sobreposição de clusters - Inibidores.



8.4.1.26 Complexo 3VEQ

Figura 8.87: 3MYW (Tripsina e BBI): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

3VEC

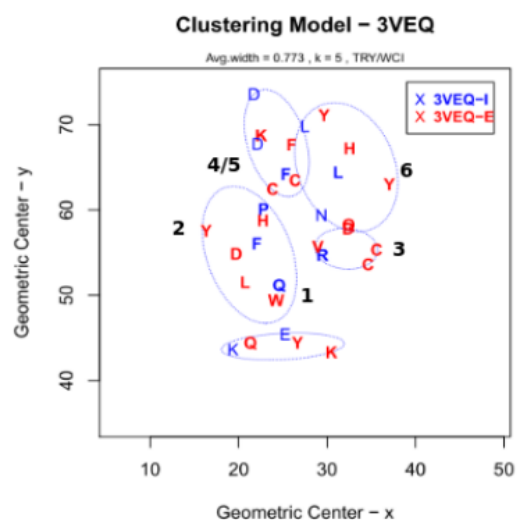
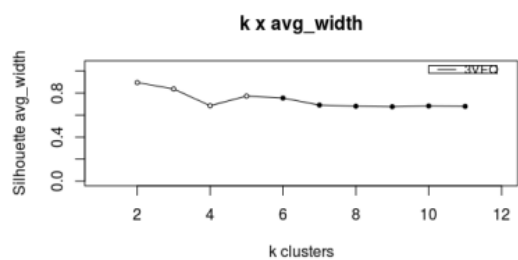
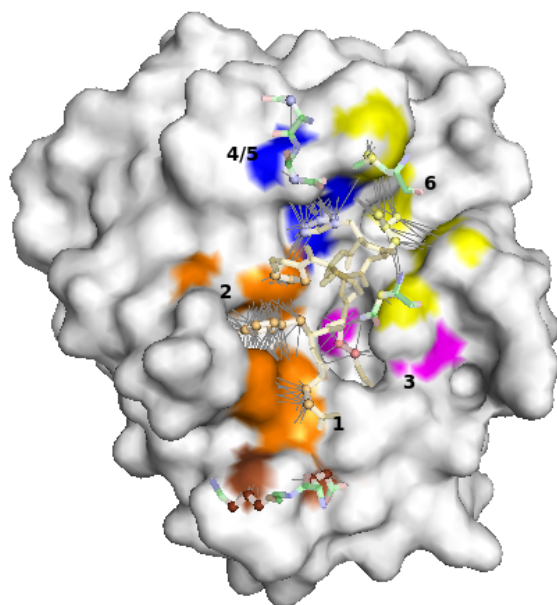
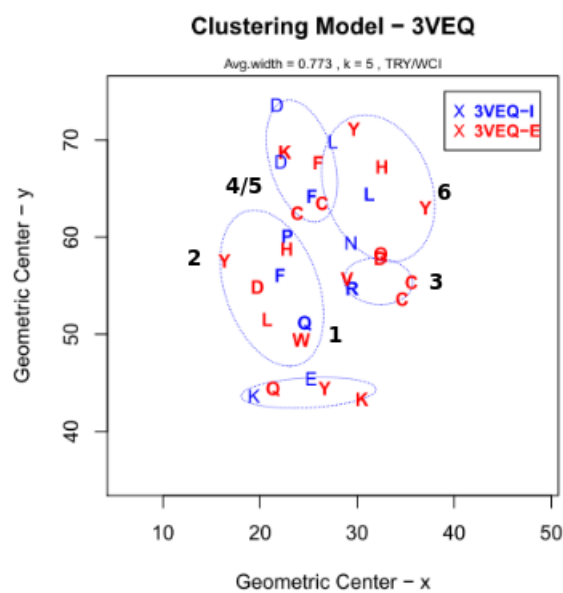
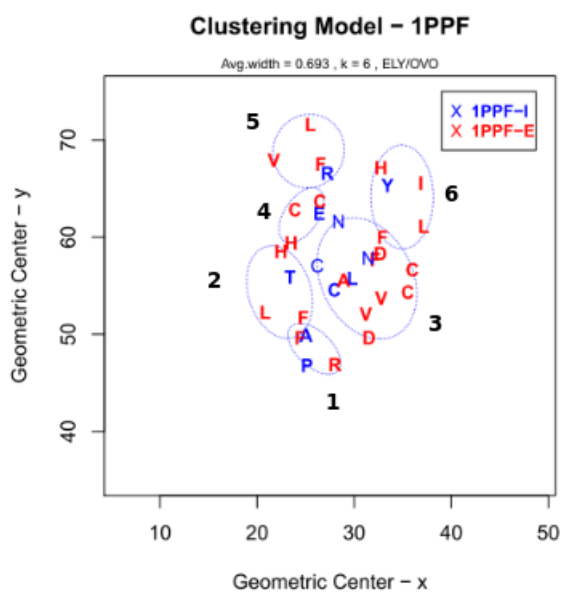


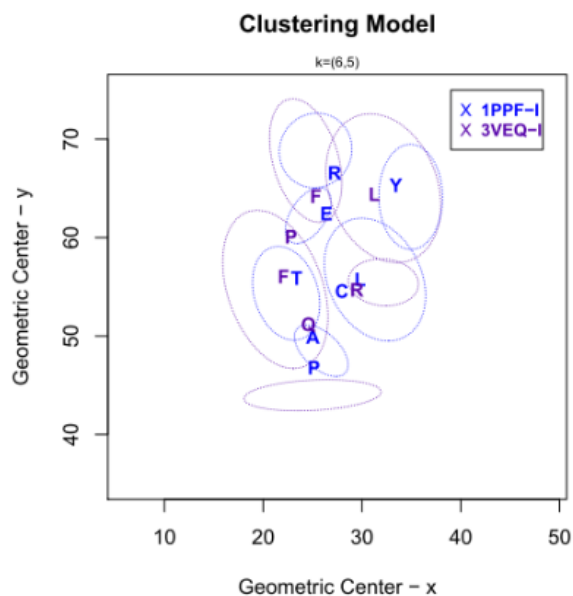
Figura 8.88: 1PPF e 3VEC: comparação de clusters.



O complexo 3VEQ contém o inibidor WCI-3 (*Chymotrypsin inhibitor* da família I3 (Kunitz-P). As regiões 2 e 4/5 são caracterizadas pela presença de resíduos aromáticos (Figura 8.88). Dos complexos analisados, somente 1YC0 apresentou um aromático na região 4/5. Há um subcluster bem evidente formado por resíduos *noloop* [KE].

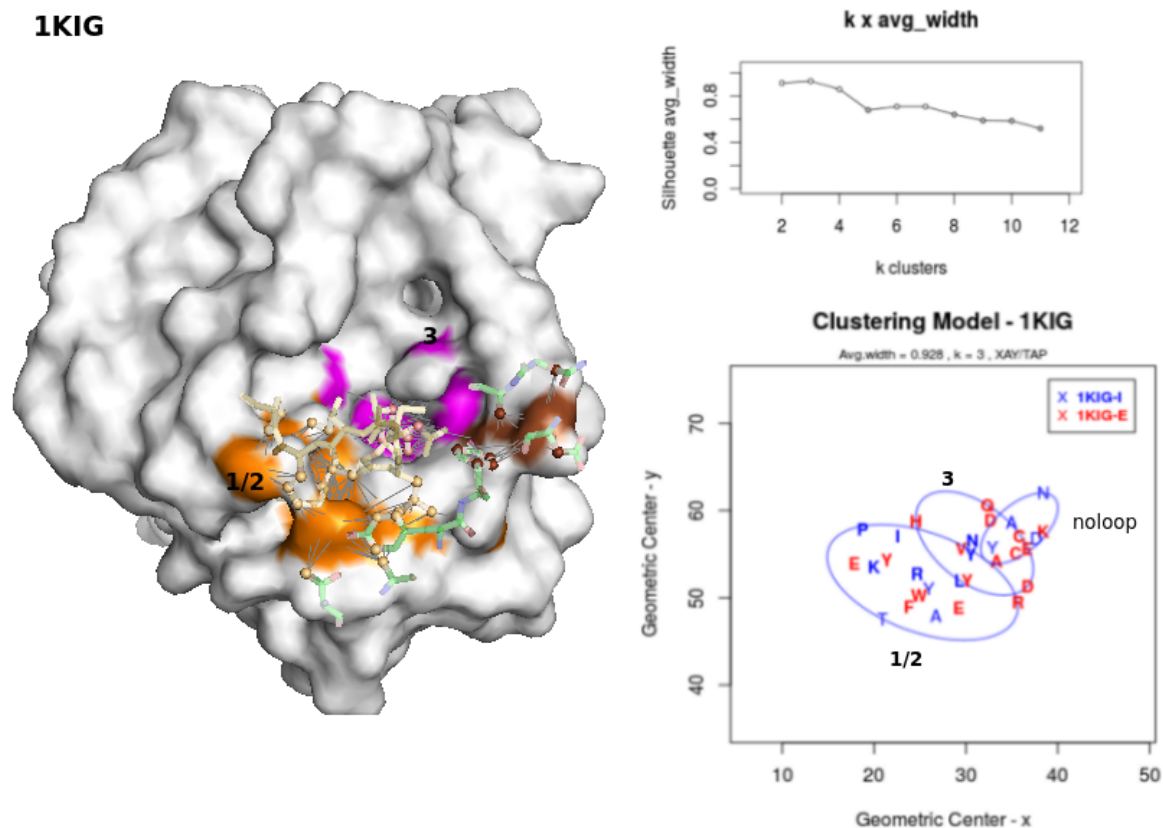
Na sobreposição de 3VEQ com 1PPF, há sobreposição de resíduos, mas eles tendem a ser diferentes. Por exemplo, onde classicamente nos cluster 4/5 se vê resíduos carregados, na 3VEQ percebemos um aromático F (Figura 8.89).

Figura 8.89: 1PPF e 3VEC: sobreposição de clusters.



8.4.1.27 Complexo 1KIG

Figura 8.90: 1KIG (Fator XA e TAP: visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters. Na estrutura tridimensional, os átomos em stiks esverdeados são noloops. O cluster marrom é formado apenas por átomos noloops.



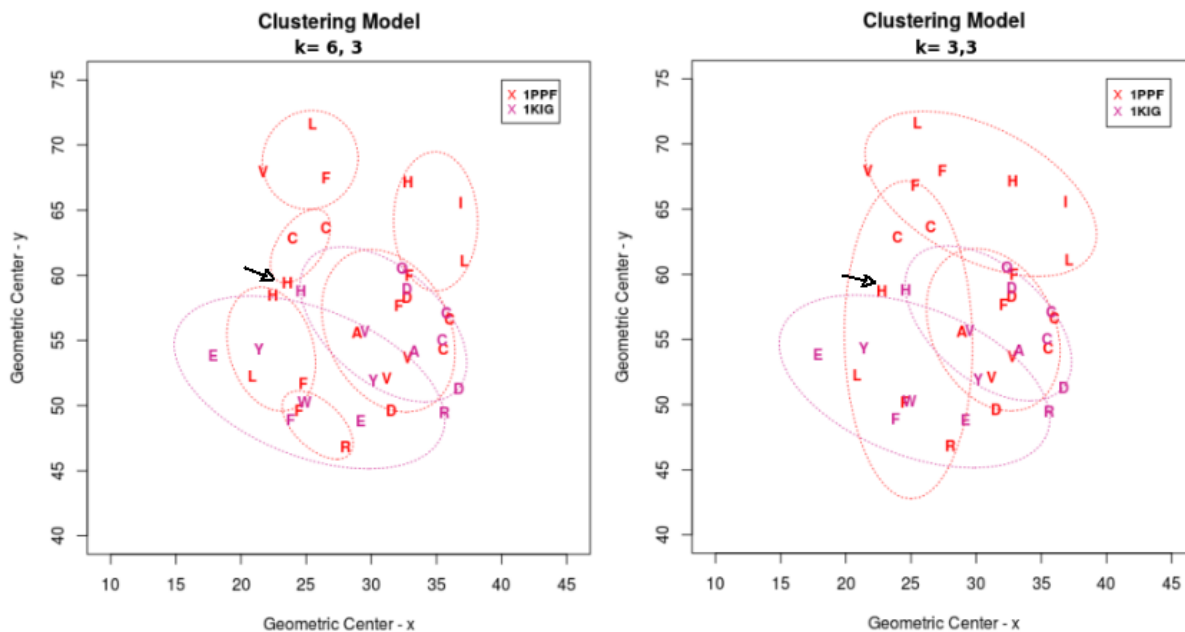
O complexo 1KIG é um exemplo de possível desvio padrão dos complexos analisados, embora mantenha várias características em comum (Figuras 8.90, 8.92 e 8.93). As regiões 1 e 2 constituem uma única região, representada em laranja na estrutura tridimensional da Figura 8.90.

Figura 8.91: 1PPF e 1KIG: comparação de clusters.



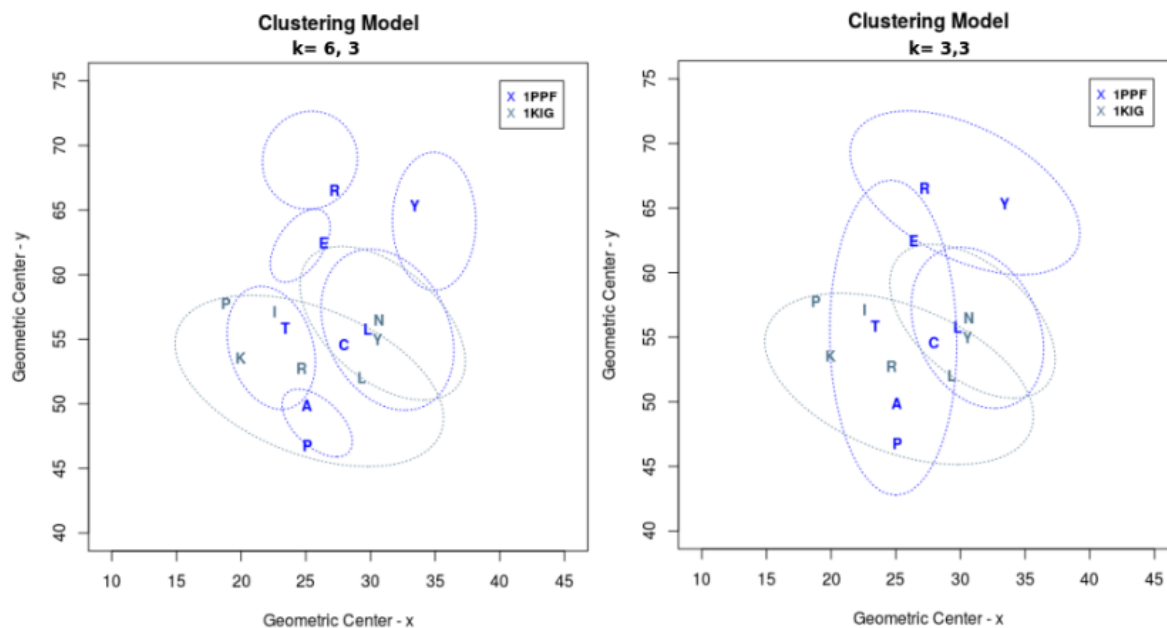
Comparando com a 1PPF, no lado enzima, vemos as Hs da tríade em posições similares (figura 8.92). Vemos também outras sobreposições, o que seria esperado, dado que 1PPF e 1KIG são ambas tipo tripsinas.

Figura 8.92: 1PPF e 1KIG: comparação de clusters lado enzima.



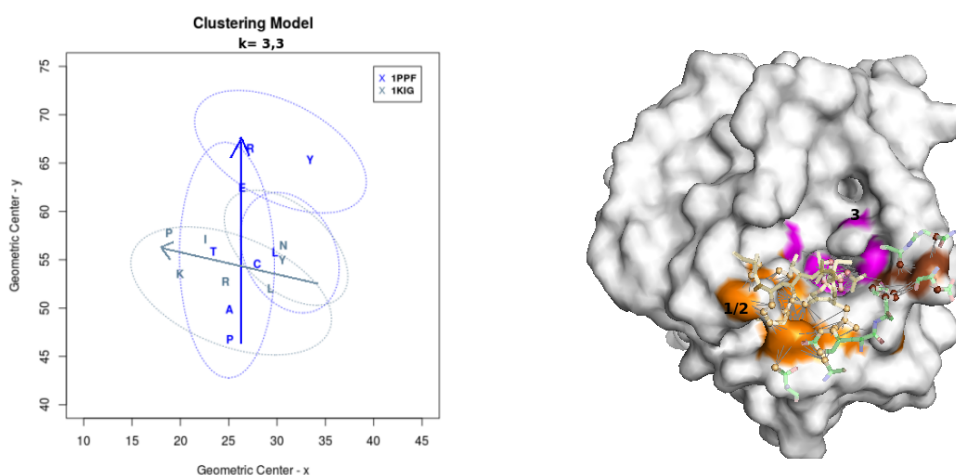
Mas, no lado inibidor, vemos o L no pocket de especificidade da 1PPF também alinhado com Y no pocket da 1KIG (figura 8.93). Também onde há um T na 1PPF, temos um I na 1KIG.

Figura 8.93: 1PPF e 1KIG: comparação de clusters lado inibidor.



Ao que parece, a 1KIG não explorou as potencialidades hidrofóbicas que a 1PPF encontrou nos subclusters 4/5 e 6. Na 1KIG eles estão vazios. Na verdade, parece que a alça da 1KIG segue em sentido quase perpendicular ao da 1PPF. Mais uma vez, parece que há mais de uma solução para o mapeamento das correspondências hidrofóbicas. Um estudo pormenorizado da 1KIG também fará parte de nossos trabalhos futuros.

Figura 8.94: Estrutura 3D de 3 regiões hidrofóbicas de 1KIG e comparação das alças inibitórias de 1PPF E 1KIG. A região em marrom além dos resíduos da enzima é formada somente por resíduos fora da alça inibitória.



8.4.2 Subtilisinas

As subtilisinas Carlsberg (CAN) e Novo (BPN) são serino peptidases de origem bacteriana com alta similaridade estrutural. Também apresentam similaridade conformacional no sítio ativo que não oferece razões óbvias para as diferenças na atividade catalítica das duas [McPhalen and James (1988)]. CAN apresenta duas ligações *cis*, uma entre Tyr167 e Pro168 e outra entre Pro210 e Thr211. Esta última está presente unicamente em CAN, enquanto a primeira também ocorre em BPN. De modo geral, há diferenças estruturais localizadas relacionadas com diferenças na sequência de resíduos de aminoácidos, particularmente na β -hélice. Ambas têm ampla especificidade com preferência para grandes resíduos aromáticos ou alifáticos na posição *P1* do substrato.

Dos 36 complexos, 9 (25 %) são tipo subtilisinas

8.4.2.1 Complexo 1R0R

Os subclusters obtidos para o complexo 1R0R (CAN-OVO) (Figura 8.95) foram comparados com 1PPF (TRY-OVO). Na Figura 8.97 são apresentadas a composição das regiões hidrofóbicas dos complexos 1PPF e 1R0R na representação gráfica 2D do modelo clusters. Já fizemos uma análise detalhada da 1R0R na inibição cruzada no início deste capítulo.

Uma diferença que não podemos deixar de comentar é que em tipo subtilisinas como 1R0R a interface hidrofóbica do lado enzima tende a ser menos complexa e numerosa que nas tipo tripsinas. Percebam, em especial, os contatos hidrofóbicos no bolsão de especificidade (subcluster 3).

Figura 8.95: 1R0R (Subtilisina Carlsberg e OMTKY3): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

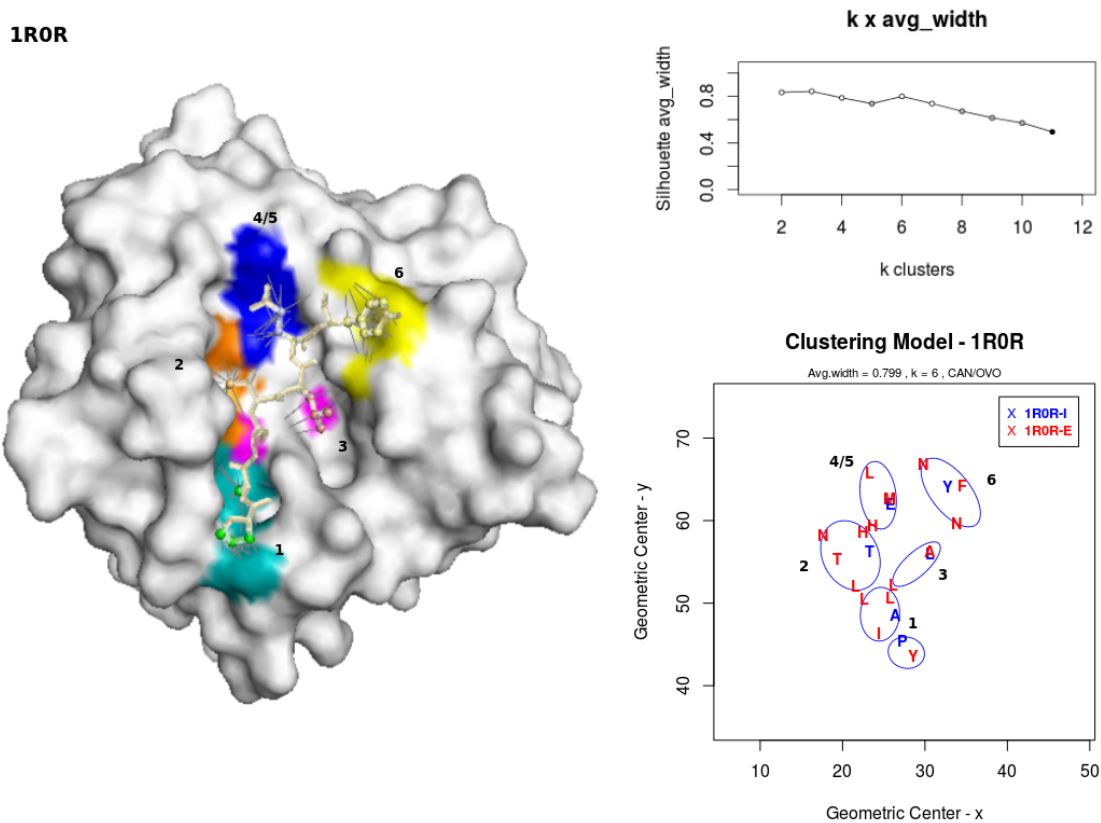


Figura 8.96: Visão 3D de 6 ($k=6$) regiões hidrofóbicas correspondentes (enzima-inibidor) para complexos 1PPF e 1R0R.

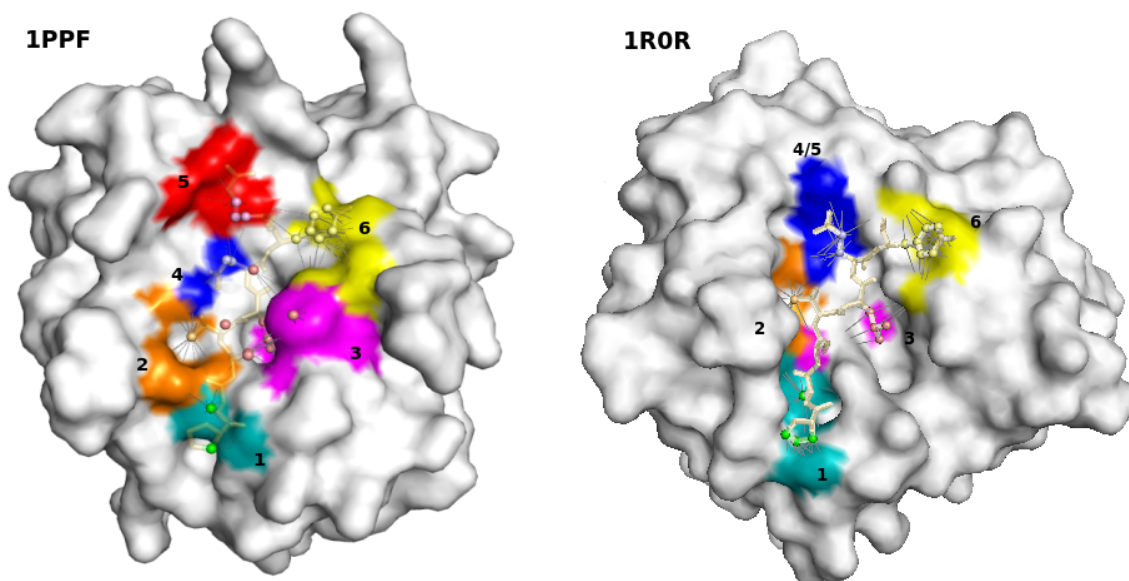
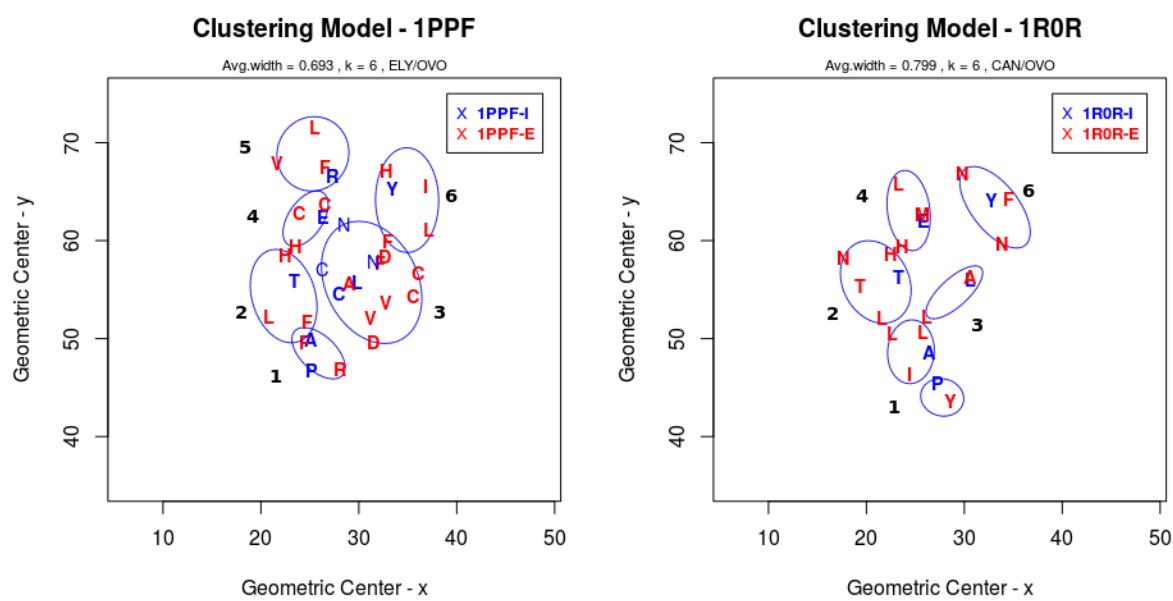
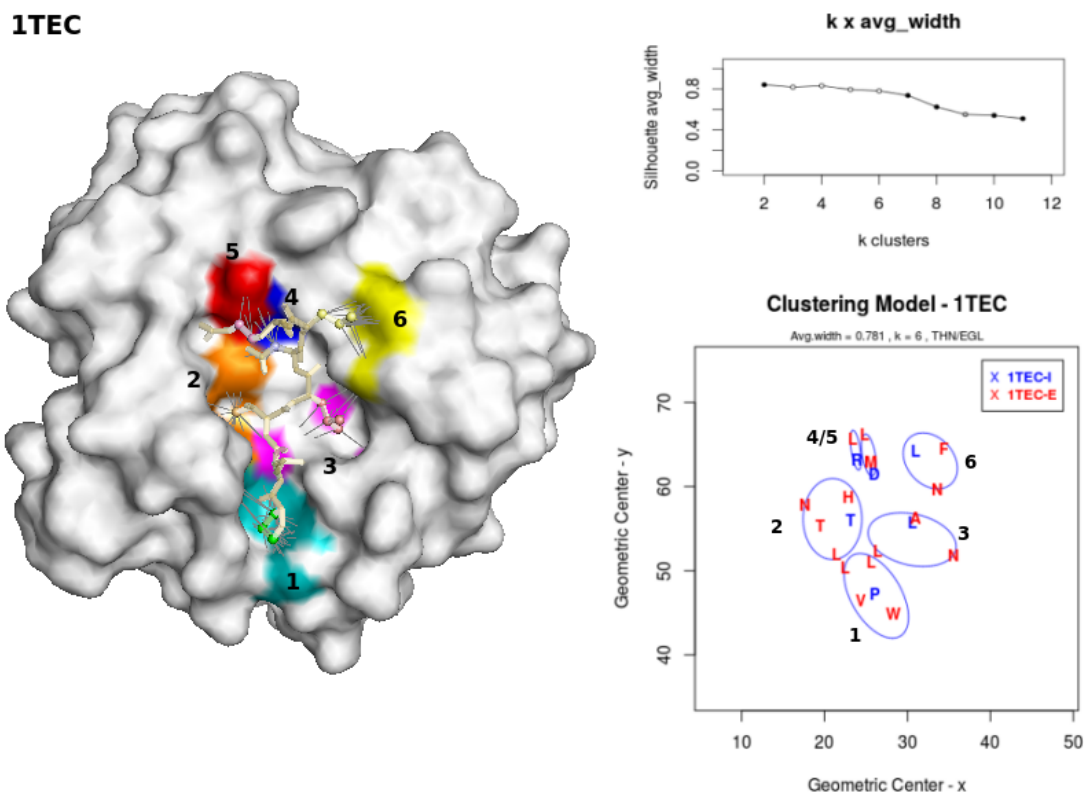


Figura 8.97: 1PPF e 1R0R: comparação de clusters.



8.4.2.2 Complexo 1TEC

Figura 8.98: 1TEC (Termitase com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O complexo 1TEC, assim como 1R0R, segue as 6 regiões observadas nos complexos com enzimas do tipo tripsinas (Figura 8.98). As regiões 4 (azul - resíduos LMD) e 5 (vermelho - resíduos LR), embora pequenas para $k = 6$, sugerindo a possibilidade de mesclar em uma única região, por se encontrarem próximas, podem ser consideradas como bons clusters visto que o coeficiente de silhueta médio é $s_m = 0.794$. Para $k = 5$, esse coeficiente é minimamente maior: $s_m = 0.794$ (Figura 8.99). Nesta mesma figura, verifica-se também ao comparar 1TEC com 1R0R que os dois inibidores diferentes (OVO e EGL) interagem, em termos de regiões hidrofóbicas, com o tipo Subtilisina de modo parecido e seguindo o padrão observado na tipo tripsinas.

Figura 8.99: 1R0R e 1TEC: comparação de clusters

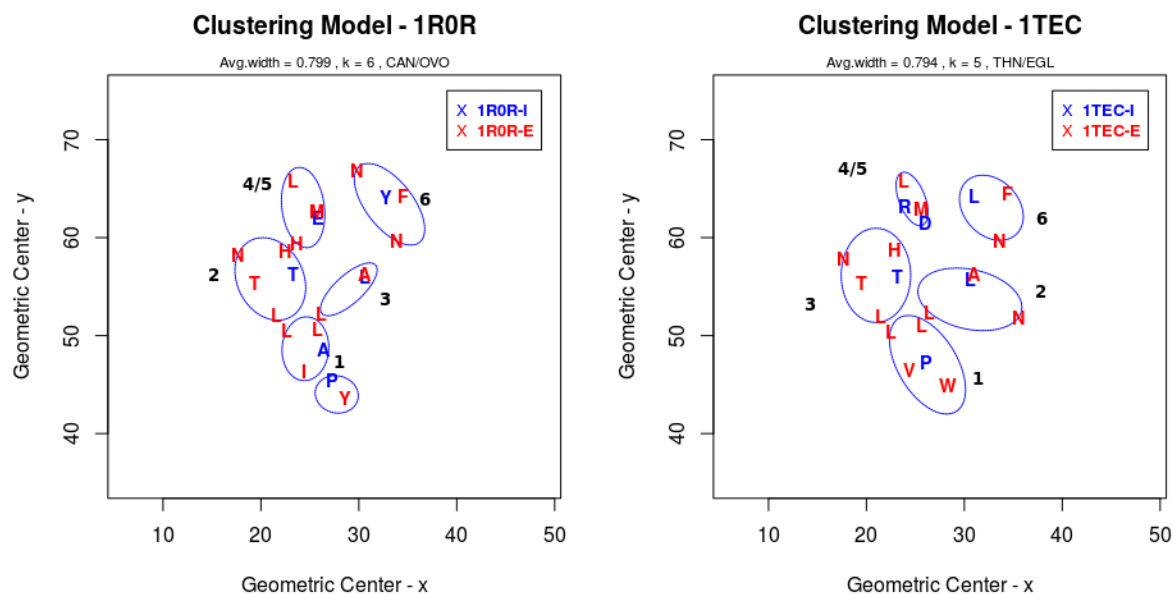
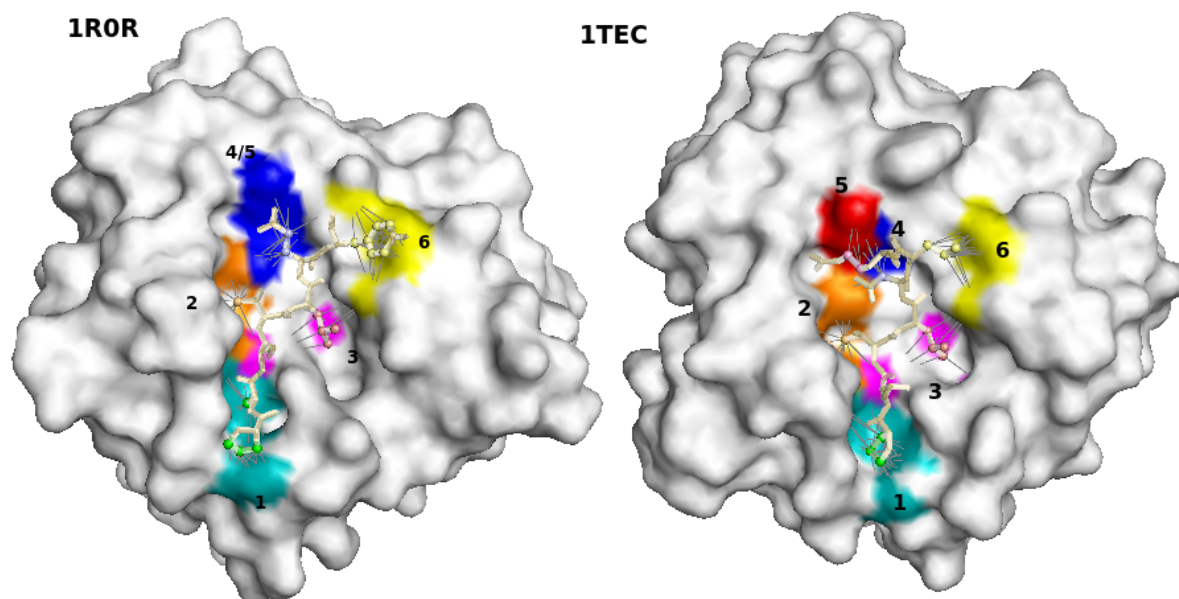
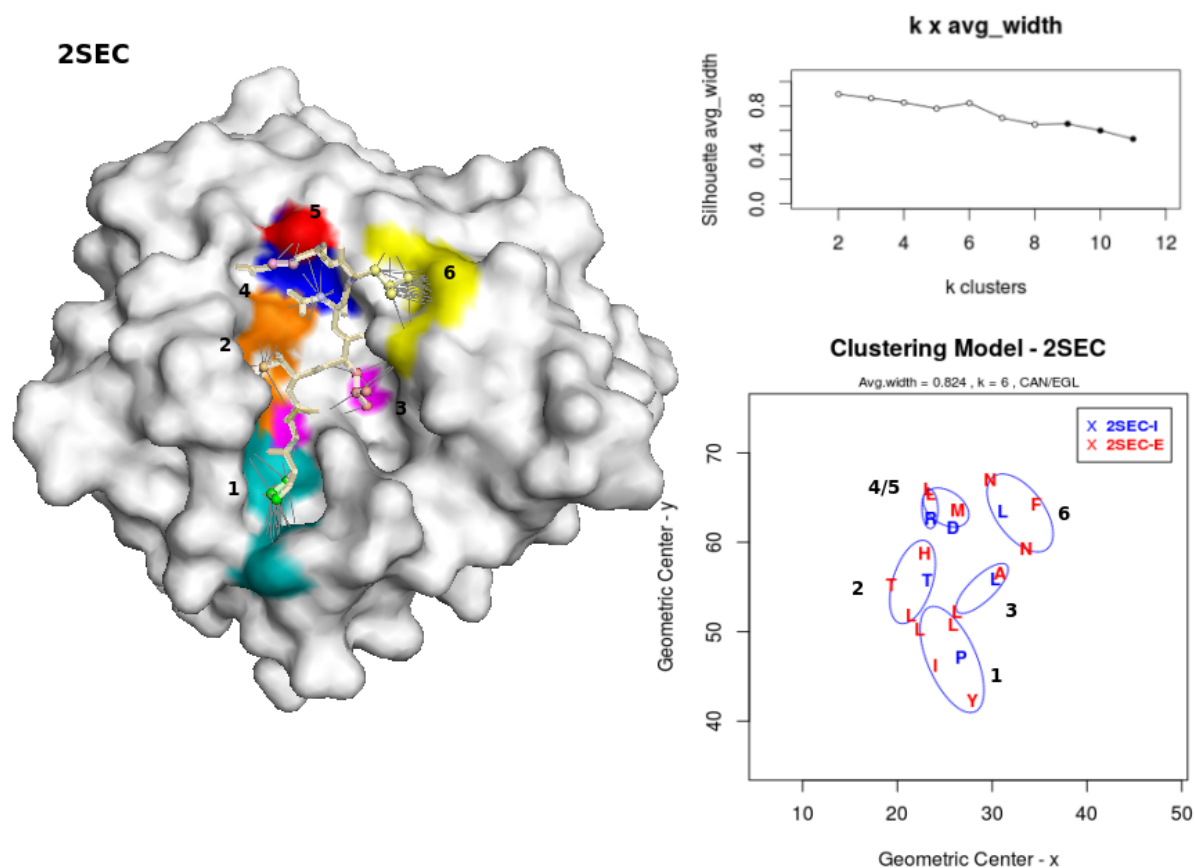


Figura 8.100: 1R0R e 1TEC: comparação de clusters no Pymol



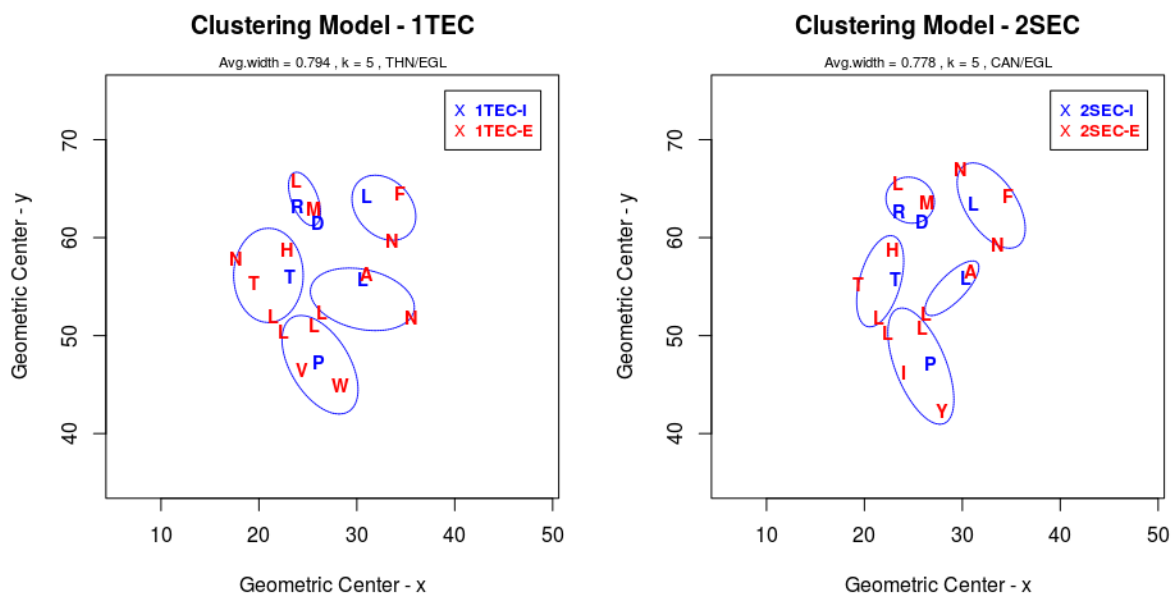
8.4.2.3 Complexo 2SEC

Figura 8.101: 2SEC (Carlsberg com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O complexo 2SEC é formado por CAN/EGL, mesmo inibidor de 1TEC. Os suclusters obtidos para 2SEC assemelham-se muito aos obtidos para 1TEC (Figura 8.102). Na Figura 8.101, são mostradas, para $k = 6$, as seis regiões que também ocorrem para 1PPF (TRY/OVO).

Figura 8.102: 1TEC e 2SEC: comparação de clusters



8.4.2.4 Complexo 1SBN

1SBN composto de BPN e EGL apresenta um conjunto de subclusters, em termos de número, semelhante ao conjunto obtido para 1TEC, mesmo quando se varia k . Porém em termos de composição desses subclusters, há algumas diferenças (Figura 8.104). Para o mesmo inibidor EGL, na região 3, uma **R** de 1SBN tem correspondência com **L** de 1TEC, ambos ocupando a posição $P1$. Na região 4/5, a correspondência é **D** - **RD**. A região 6 de 1SBN contém **LY**, enquanto 1TEC contém somente **L**. Apesar dessas diferenças, a existência das regiões hidrofóbicas correspondentes enzima-inibidor é similar às subtilisinas acima apresentadas (Figura 8.103).

Figura 8.103: 1SBN (Subtilisina BPN com Eglina C): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

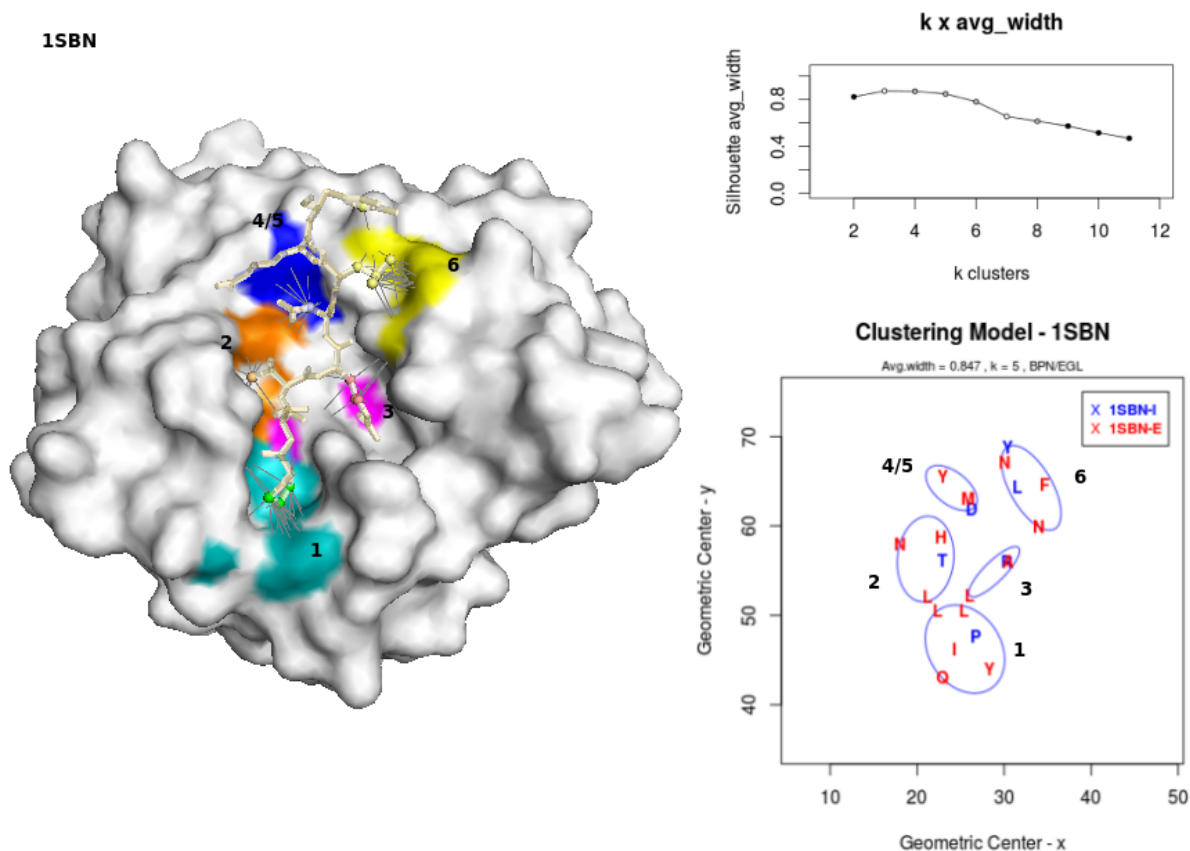
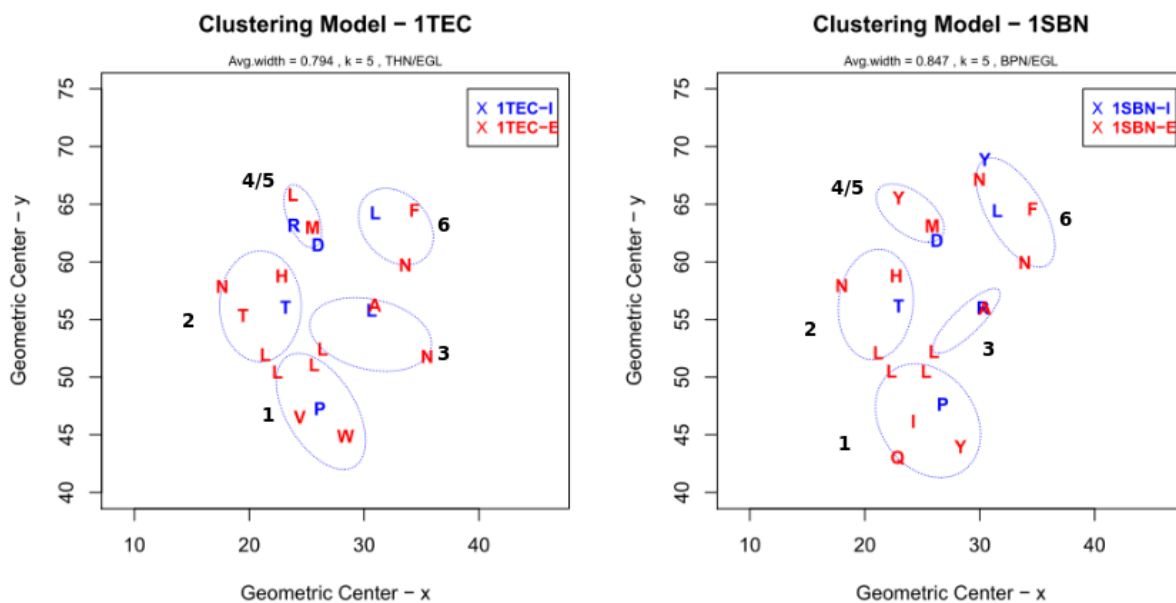


Figura 8.104: 1TEC e 1SBN: comparação de clusters



8.4.2.5 Complexo 1LW6

As regiões hidrofóbicas identificadas em 1LW6 são melhor caracterizadas para $k = 5$, visto que para $k = 6$ a I56 fica na posição intermediária entre a região 1 e 2, podendo pertencer a ambos os subclusters, como pode ser visto à esquerda na Figura 8.106. Quando $k = 5$, a I56 fica bem posicionada no subcluster 1 (à direita da referida figura). As regiões 5 e 4, observadas em 1ROR, ficam condensadas em um único subcluster **ERYM**, identificado na Figura 8.106, à direita, como [4/5] e representado em vermelho na estrutura tridimensional da Figura 8.105.

Da comparação de 1ROR com 1LW6, CAN/OVO e BPN/CI2, vê-se que há muitas similaridades nas regiões hidrofóbicas de ambos os complexos, mesmo apresentando inibidores diferentes (Figura 8.107). Outro aspecto, também observado para os demais complexos de subtilisinas diz respeito à quantidade de resíduos dos clusters.

Figura 8.105: 1LW6 (BPN e Inibidor de Quimotripsina 2 (CI-2)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

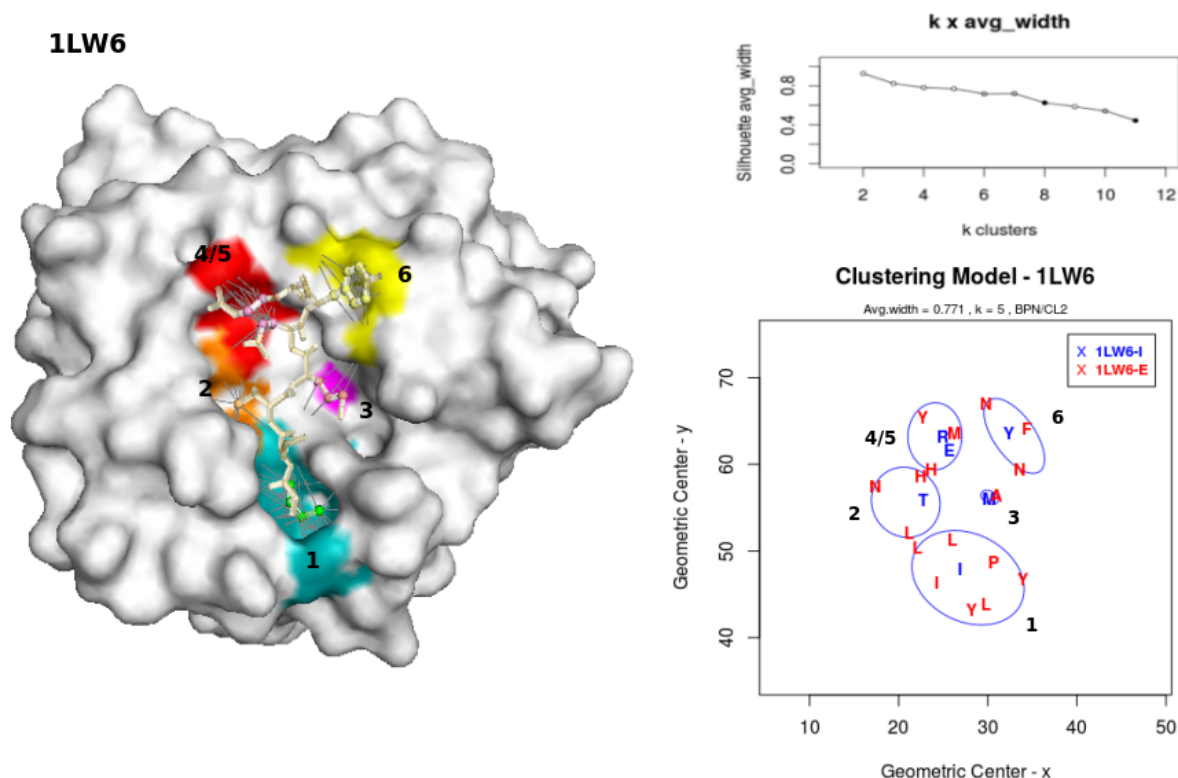


Figura 8.106: 1LW6: clusters para k=6 e k=5.

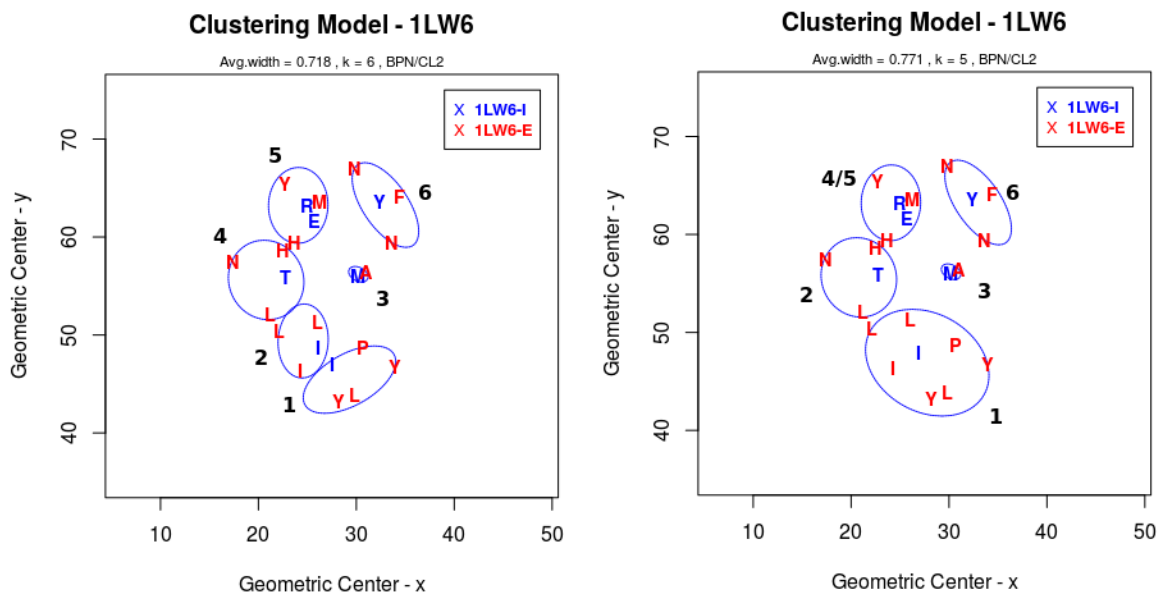
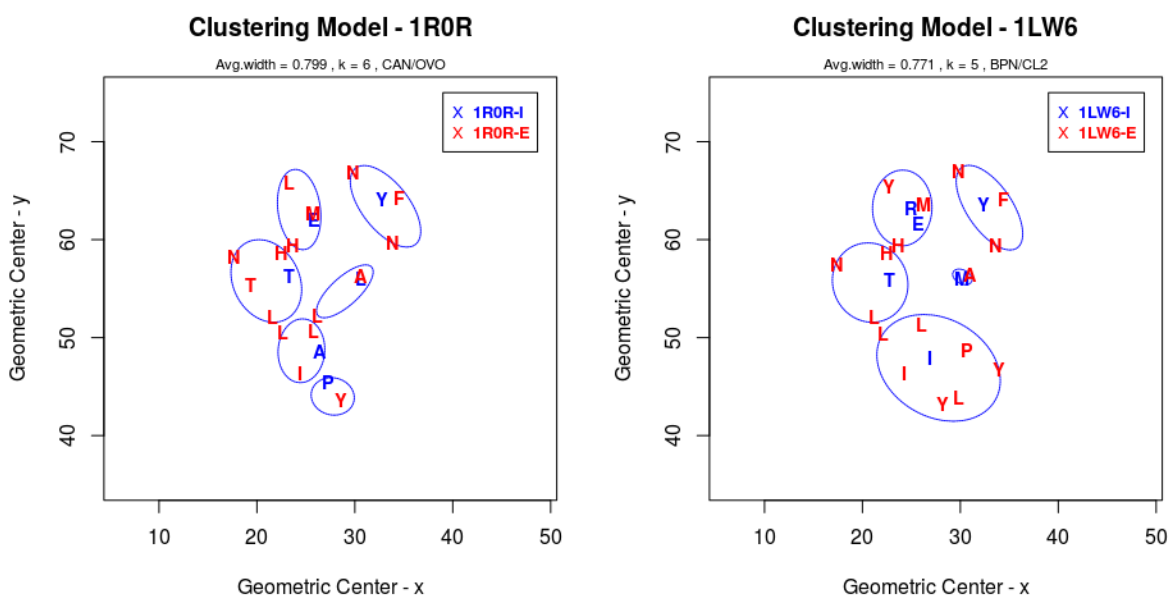


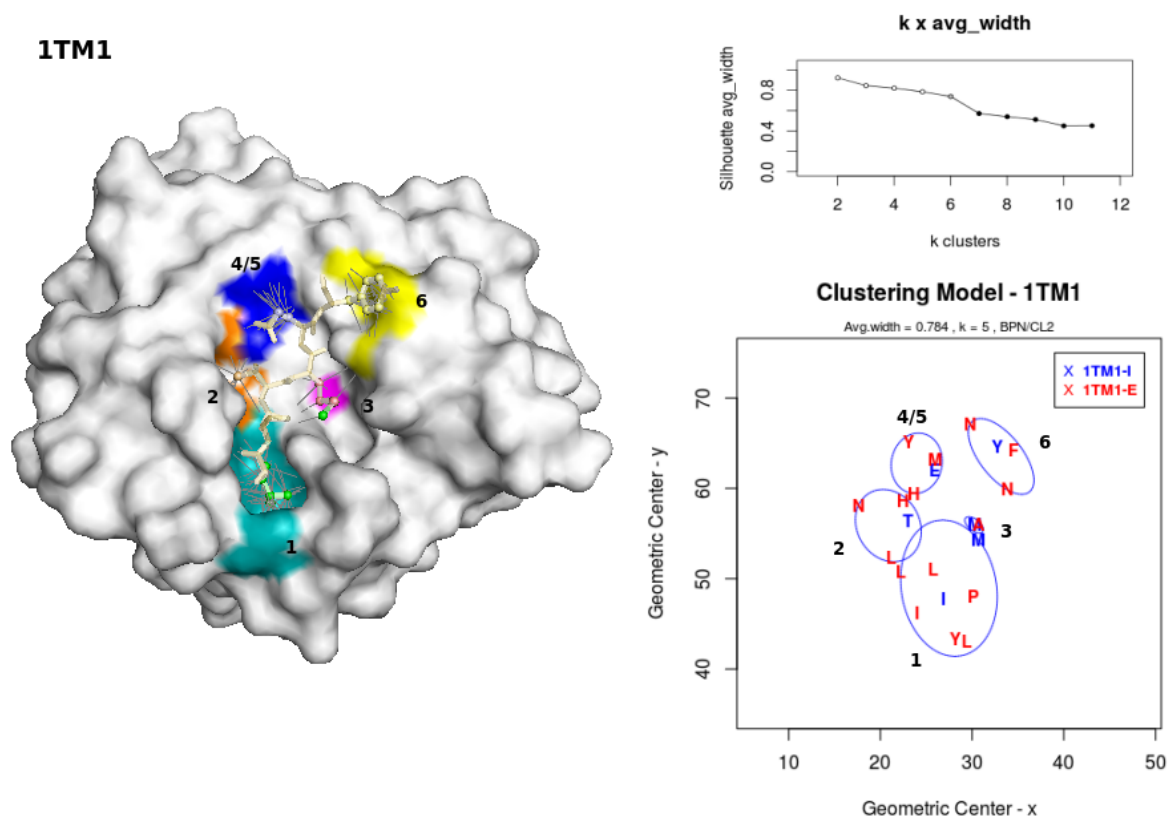
Figura 8.107: 1R0R e 1LW6: comparação de clusters.



Nos complexos com subtilisina, diferentemente daqueles com tipo Tripsinas, a quantidade de resíduos nas regiões tende a ser menor, apontando para um padrão mais simples e “enxuto”. A região 3 é a que mais apresenta essa característica. Para 1LW6, por exemplo, é composta dos resíduos **MA**.

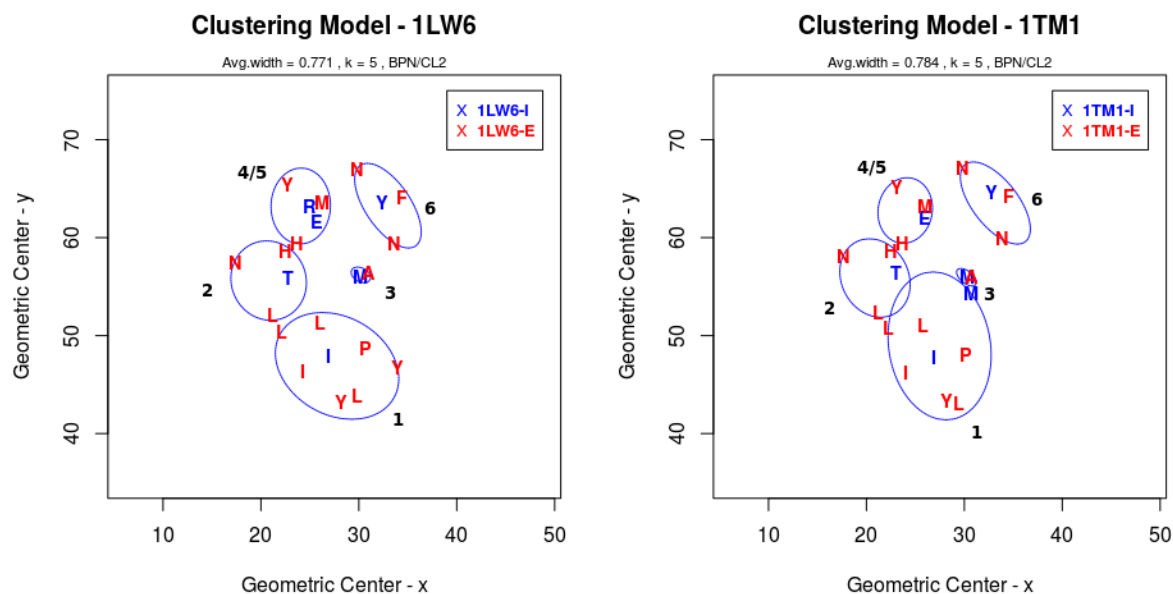
8.4.2.6 Complexo 1TM1

Figura 8.108: 1TM1 (BPN e Inibidor de Quimotripsina 2A (CI-2A)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O inibidor CI-2 de 1LW6 é um mutante do inibidor CI-2A que compõe 1TM1 (Figura 8.108). A Leu20 foi substituída por uma Met [Radisky and Koshland (2002)]. BPN de 1LW6 é uma subtilisina recombinante, cujas substituições em BPN foram Asn155Leu e Met222Ala. Comparando-se os suclusters resultantes para esses dois complexos, observa-se uma forte semelhança (Figura 8.109). Na região 3 de 1LW6, a Met59I tem alguns átomos que fazem contato com resíduos do inibidor da região 2, por isso, a região 3 nesse complexo é exibida mais próxima à região 1 do que em 1TM1.

Figura 8.109: 1LW6 e 1TM1: comparação de clusters



8.4.2.7 Complexo 1OYV

Do inibidor WIP da família I20 (Pot2) foi considerado o segundo domínio, pois este é composto por dois domínios [Barrette-Ng et al. (2003)]. O complexo 1OYV tem um modelo de subclusters mais diferenciado em relação às outras subtilisinas, pois tem 3 subregiões extensas formadas por átomos *noloop* (Y34I, I14I, F100I, D103I). Estas regiões podem ser vistas na Figura 8.110 em marrom. Provavelmente, são essas regiões que estão conferindo à 1OYV seu padrão singular em termos de subtilisinas. As outras regiões, podem ser mapeadas para as regiões observadas em outros complexos contendo subtilisina, como 1R0R (CAN/OVO) e 1TEC (THN/EGL) (Figura 8.111). O inibidor EGL é da família I13 (Pot1). Na região 6 de 1OYV, está presente o resíduo Cys64I (posição P2) que forma pontes dissulfeto com a Cys60I, que são características de membros da família I20.

Figura 8.110: 1OYV (CAN e Wound-induced proteinase inhibitor-II (WIP)): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.

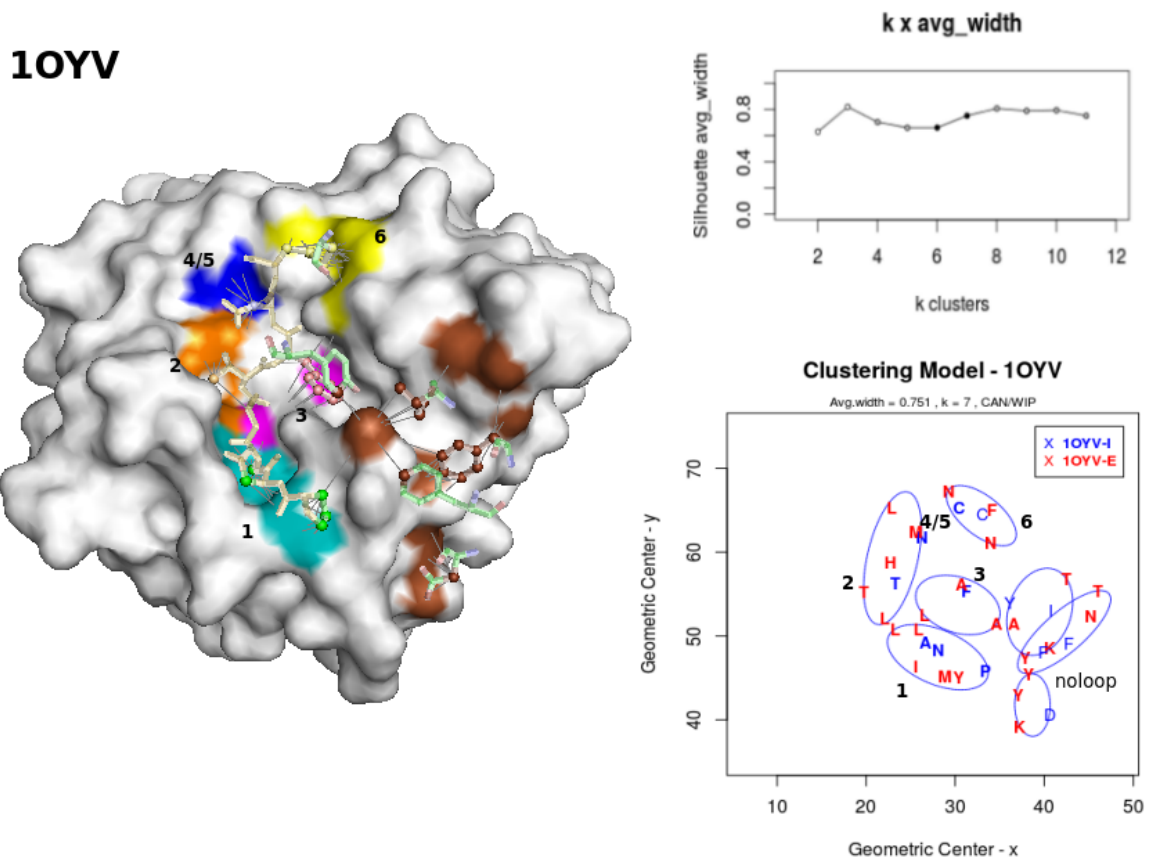
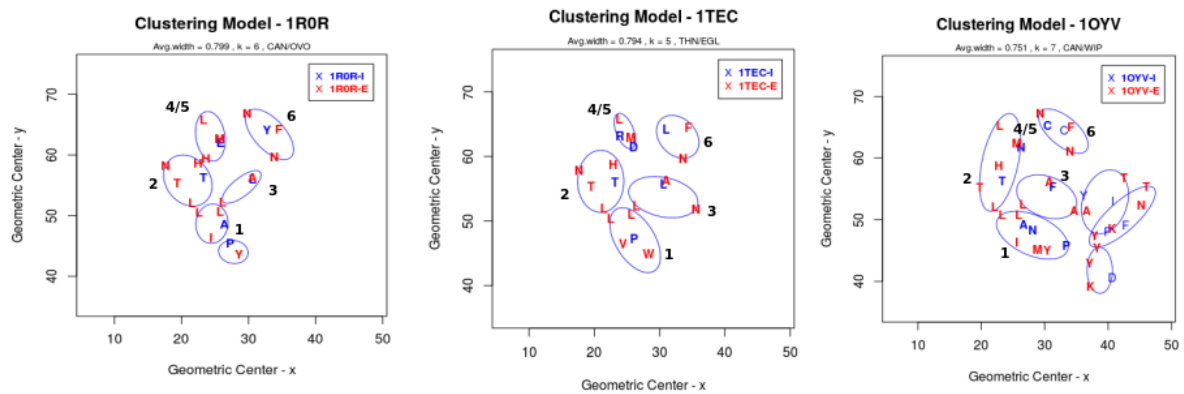
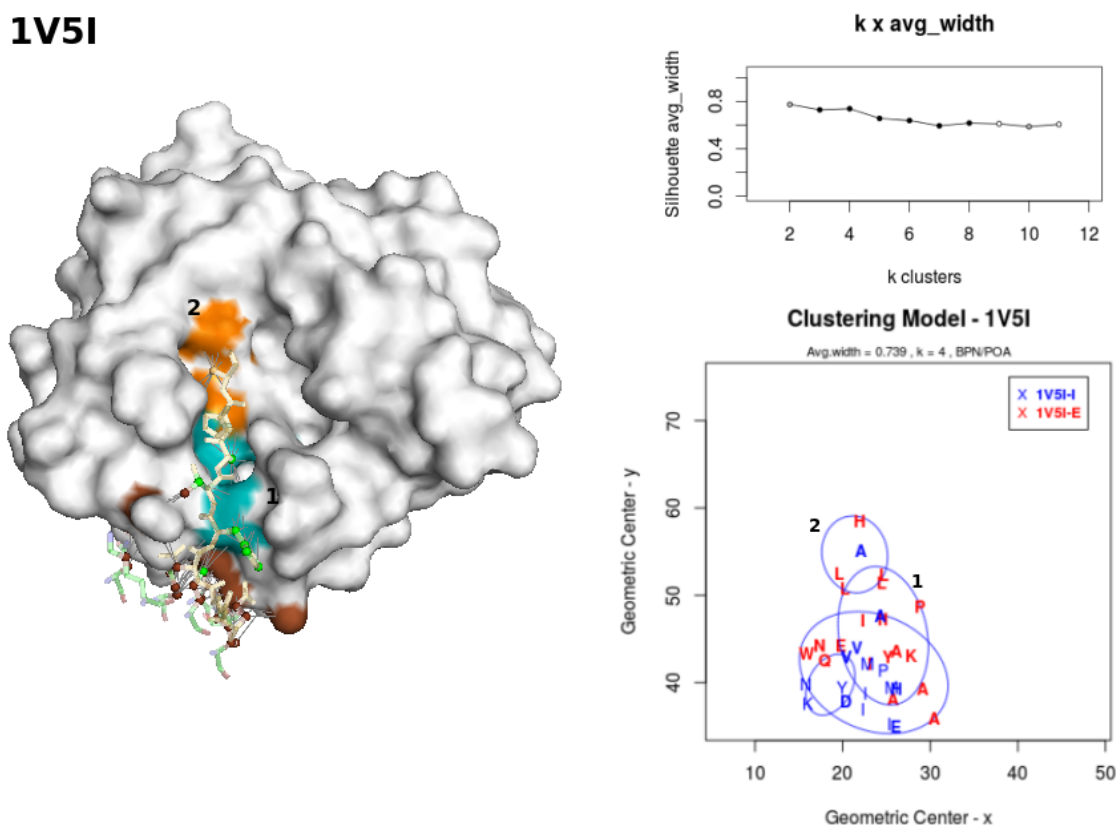


Figura 8.111: 1R0R e 1OYV: comparação de clusters



8.4.2.8 Complexo 1V5I

Figura 8.112: 1V5I (Subtilisina BPN e POIA1): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O inibidor POIA1 (família I9) funciona como uma chaperona intramolecular e como um inibidor de subtilisina [Kojima et al. (2005)]. É homólogo ao propeptídeo de BPN. O sítio reativo de POIA1, composto por resíduos na posição C-terminal, apresenta muita flexibilidade.

Pela estrutura 3D da 1V5I (Figura 8.113), é possível ver que o inibidor ocupa uma vasta região (em marrom) fora do sítio ativo da enzima, que é ocupado apenas parcialmente (laranja e verde). O sítio de especificidade, por exemplo, está praticamente desocupado, ou seja, trata-se de um mecanismo de inibição diferente do clássico, e deve ser função mais de uma "obstrução" do sítio, atrapalhando a "entrada" do substrato. Do ponto de vista da ligação enzima-inibidor observa-se que a não ocupação total do sítio catalítico pode ter sido compensada por interações secundárias fora do sítio ativo.

Abaixo são mostradas sobreposições dos subclusters de 1R0R com 1V5I, para $k = 6$ e $k = 5$, respectivamente (Figuras 8.114, 8.115, 8.116).

Uma certa correspondência espacial dos clusters é observada para as regiões 1 e 2 somente.

Figura 8.113: 1V5I: (A) e (B) visão da estrutura tridimensional com destaque para região hidrofóbica fora do sítio de especificidade. (C) destaque do sítio de especificidade.

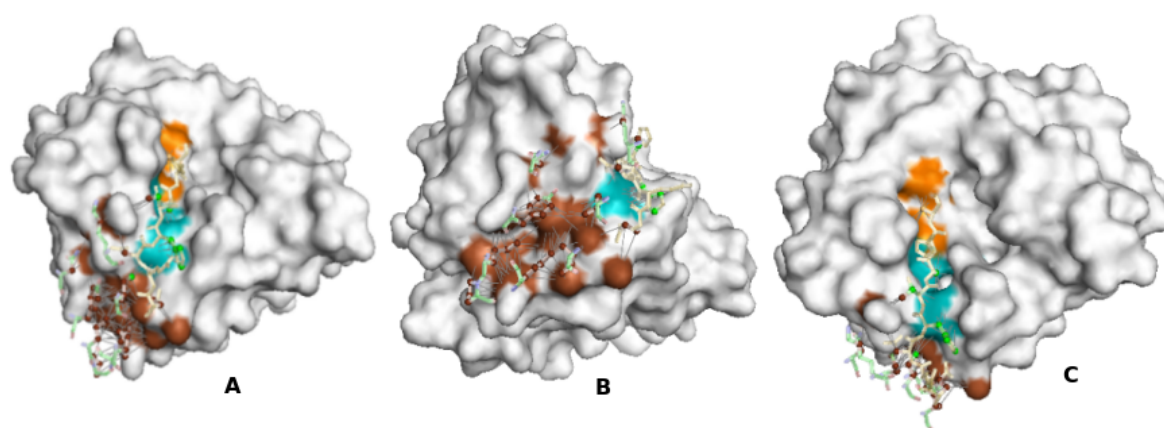


Figura 8.114: 1R0R e 1V5I: comparação de clusters.

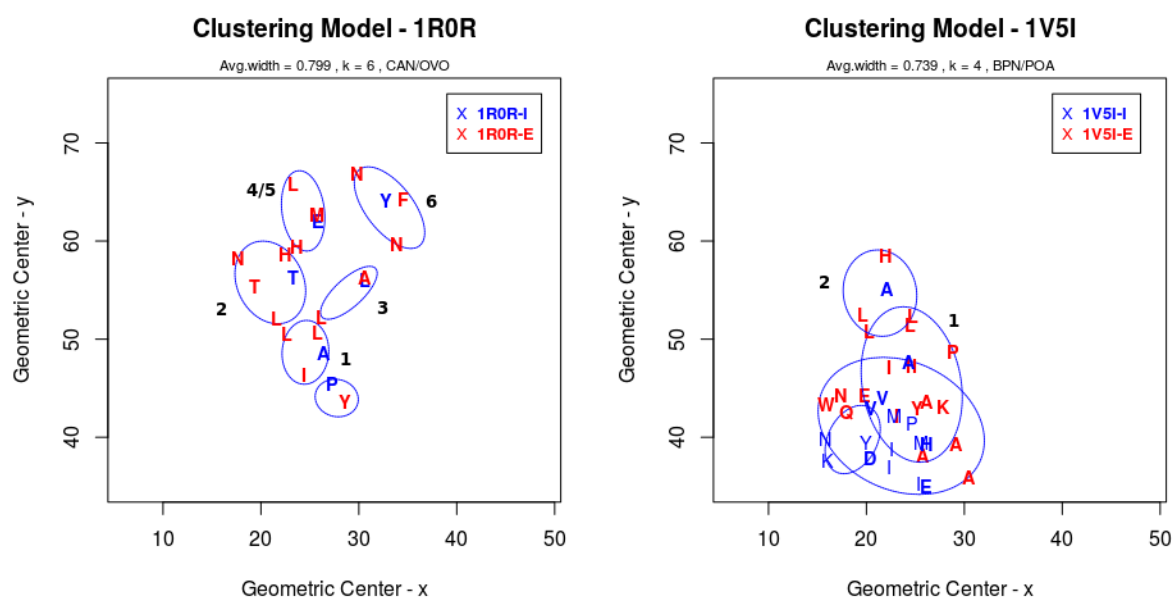
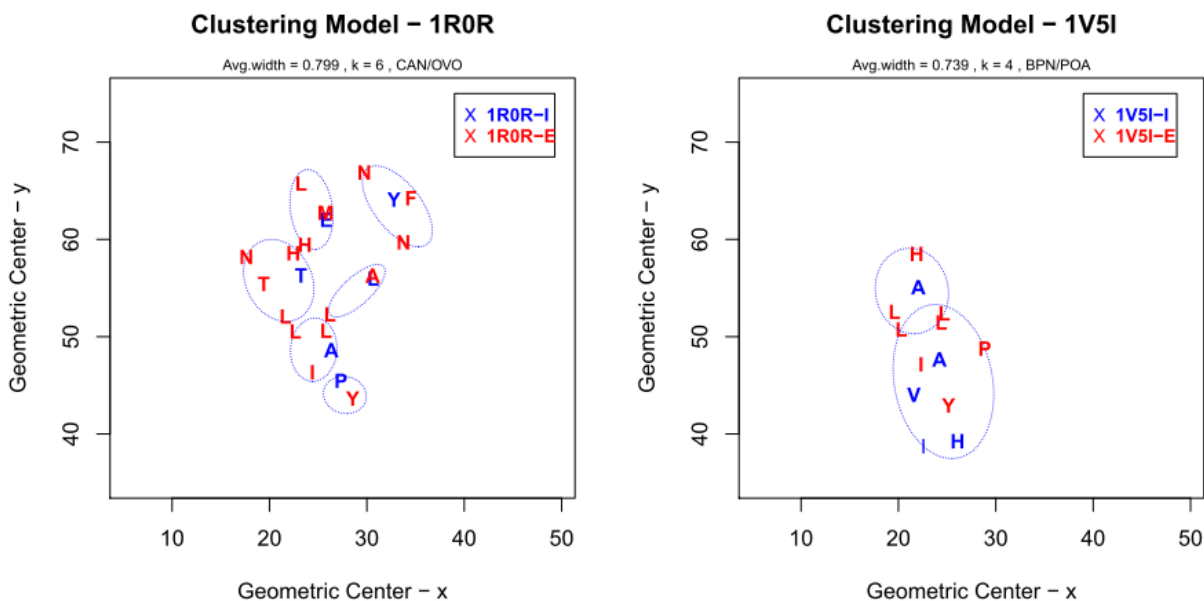
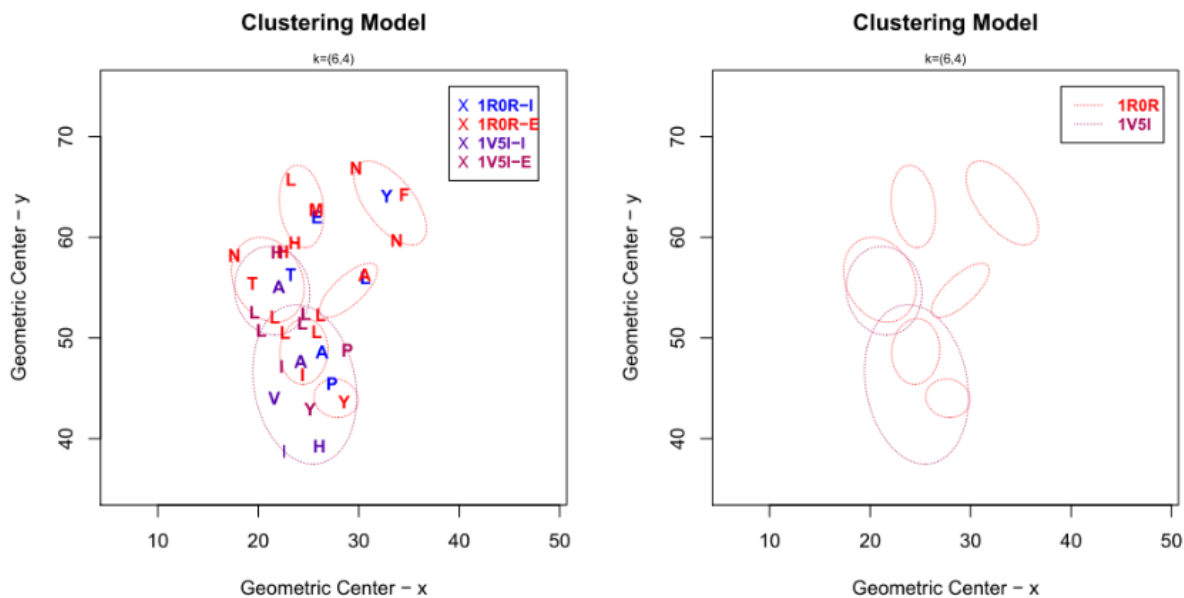


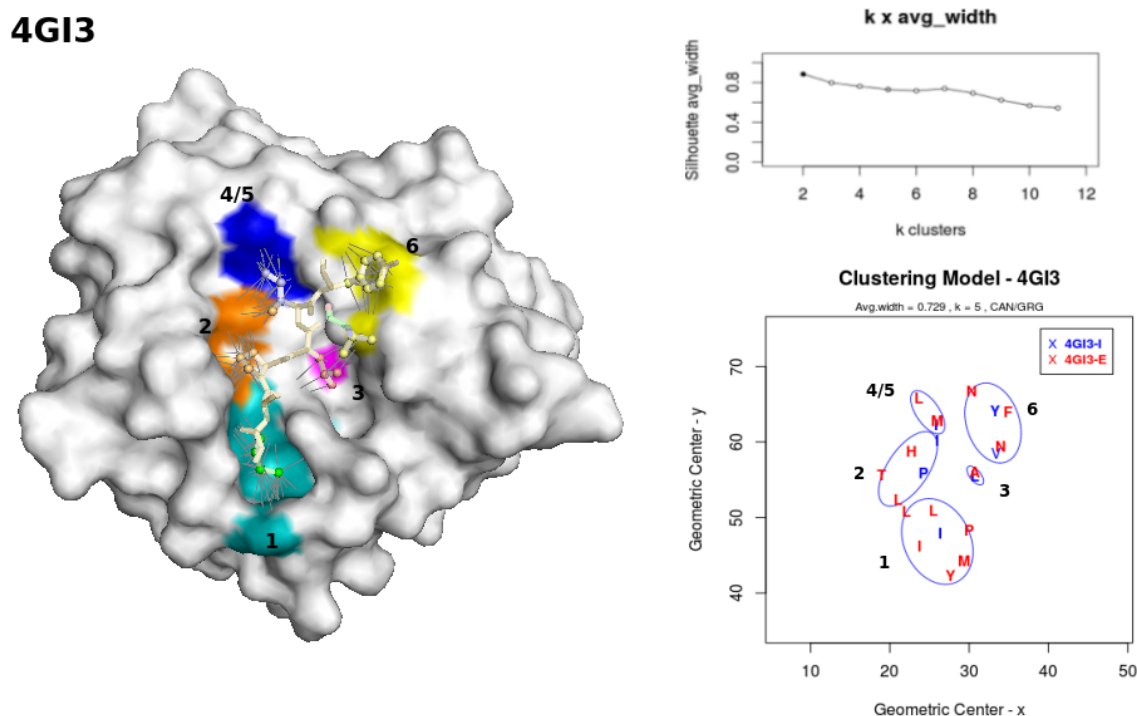
Figura 8.115: 1R0R e 1V5I: sobreposição de clusters.

Figura 8.116: 1R0R e 1V5I: sobreposição simplificada dos clusters de 1V5I, ou seja, os clusters de *noloops* foram excluídos.

Trata-se de mais um caso em que soluções alternativas buscando um melhor mapeamento das correspondências hidrofóbicas foram encontradas pela 1V5I, dessa vez fora da interface "clássica" descrita para 1PPF. A alça, em 1V5I, começa a correr numa extensão anterior aos subclusters 1/2, com agregação de hidrofóbicos, como VIIH.

8.4.2.9 Complexo 4GI3

Figura 8.117: 4GI3 (CAN e Greglina): visão 3D do complexo com regiões hidrofóbicas, coeficiente médio de silhueta, modelo de clusters.



O inibidor Greglina é um forte inibidor de subtilisina e elastase de neutrófilos humana (HLY). Também inibe a α -quimotripsina e a elastase pancreática suína. Greglina representa um membro dos inibidores de Kazal não clássicos: tem uma única região C-terminal adicional (70-83), ligada ao núcleo da molécula por meio de uma ligação de dissulfeto [Derache et al. (2012)].

O complexo 4GI3 apresentou um modelo de subclusters muito similar aos observados para a maioria das subtilisinas. O modelo 4GI3 foi comparado com 1ROR, por possuir mesmo tipo de enzima e com 1F2S por causa dos *loops* de inibição que têm resíduos em comum (ICPRW e ICPLI de 1F2S e 4GI3, respectivamente).

Embora os inibidores OMTKY3 e Greglina tenham especificidades diferentes, de modo que a Greglina prefere subtilisinas (família S1) às peptidases da família S8, onde OMTKY3 atua largamente, o resíduo *L* na posição *P1* é comum aos dois inibidores. A razão da preferência por subtilisinas deve residir em outras regiões da Greglina [Derache et al. (2012)]. Na Figura 8.118), nota-se que as regiões de 1ROR e de 4GI3 são semelhantes, a despeito das diferenças de composição.

Já 1F2S e 4GI3, têm os resíduos dos subgrupos sobrepostos, à exceção de *I* na região 1 (Figuras 8.119 e 8.120).

Figura 8.118: 1R0R e 4GI3: comparação de clusters.

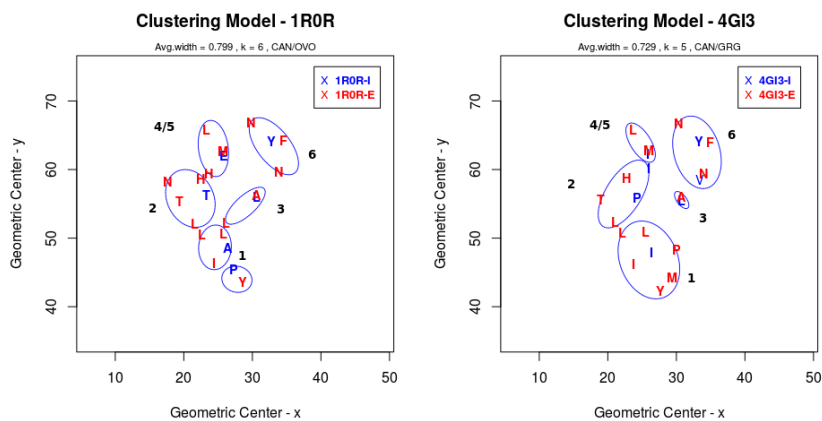


Figura 8.119: 1F2S e 4GI3: comparação de clusters.

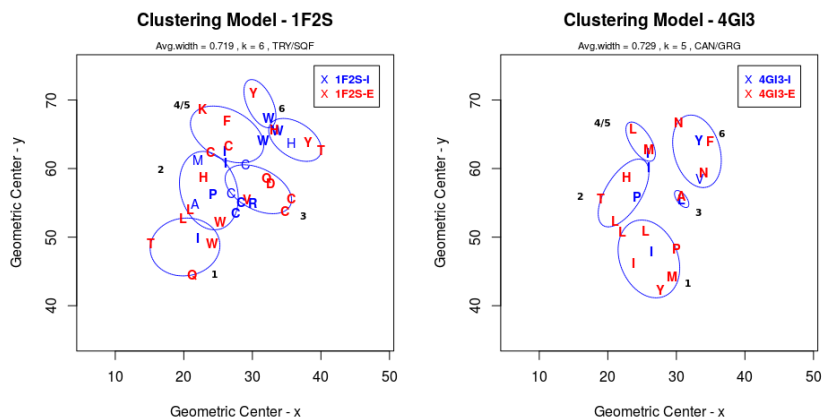
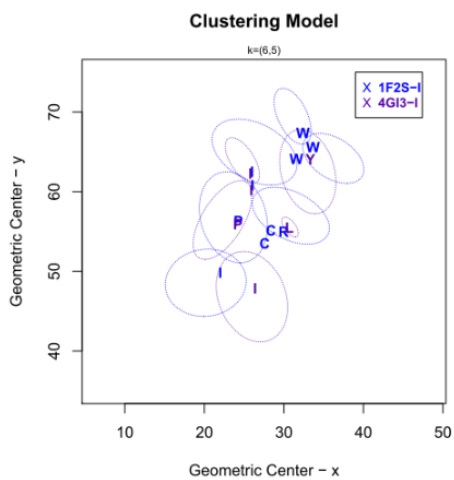


Figura 8.120: 1F2S e 4GI3: sobreposição de clusters somente dos inibidores.



8.5 Clusteres - Modelo Geral

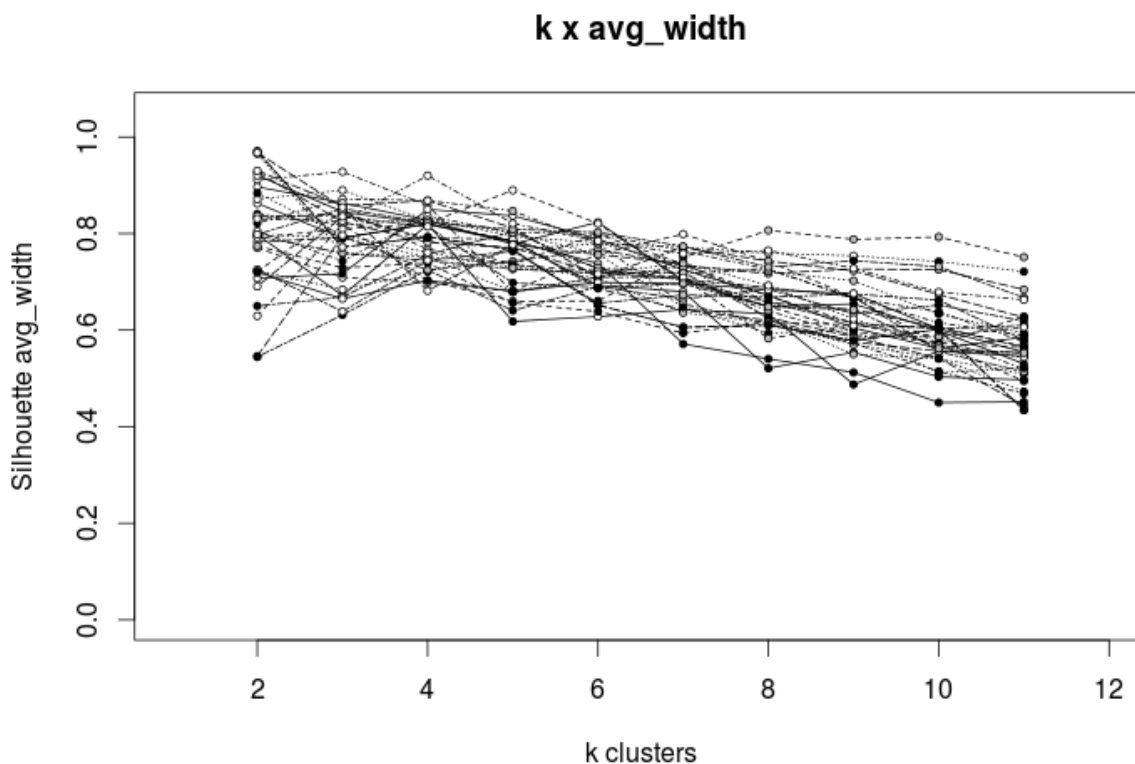
8.5.1 Varredura de Agrupamentos

No gráfico da Figura 8.121, é apresentado uma visão conjunta do Coeficiente médio de Silhueta de cada complexo (s_m) e sua variação em função do número de clusters. Até 6 clusters ($2 \leq k \leq 6$), talvez com um ótimo em $k = 4$, o comportamento de s_m é oscilante e embora se observe pontos pretos (indicando que algum átomo ou conjunto de átomos não foram bem clusterizados), eles estão em menor número. A partir de $k > 6$, a qualidade do processo de clusterização se deteriora (mais pontos pretos são observados). De modo geral, até 6 clusters tem-se um bom compromisso com a qualidade dos clusters formados: dado que $s_m \geq 0.6$, tem-se grupos com uma estrutura razoável (0.51 – 0.70) ou grupos com estrutura muito robusta (0.71 – 1.00). Seja como for, uma análise estatística rigorosa ainda carece de ser feita aqui, mais um item em trabalhos futuros.

Eis porque chamamos nossa técnica de "varredura" de agrupamento espectral. Ante as deficiências ou insuficiências inerentes às métricas de qualidade de clusterização, mesmo tão populares quanto o coeficiente de silhueta, optamos por olhar todo o conjunto de valores da métrica. A percepção global dessa varredura de k nos sugere que a quantidade de subclusters podem variar em nosso conjunto de dados, mas o fazem de forma a não afetar muito essa qualidade aferida pelo coeficiente de silhueta. Pelo menos, até 6 clusters. A partir daí, salvo algumas exceções ditadas principalmente pela presença de regiões *noloops*, a tendência é de deterioração desse parâmetro.

E se temos bons clusters com $2 \leq k \leq 6$, uma conclusão possível é que temos, no fim das contas, um único grande cluster, que se fragmenta de forma não rigorosa em subclusters menores. Se verdade, isso pode estar nos revelando uma curiosa "estrutura" geral subjacente na organização hidrofóbica das interfaces nas peptidases estudadas, algo que discutiremos mais a fundo um pouco adiante.

Figura 8.121: Relação entre o número de clusters e o Coeficiente médio de Silhueta para todos os 36 complexos. Ponto preto: $s_m < 0$, ou seja, algum átomo ou conjunto de átomos não foram clusterizados adequadamente. Ponto cinza: algum átomo ou conjunto de átomos estão num ponto intermediário entre dois clusters. Ponto branco: os átomos têm valores próximos a 1, portanto bem definidos nos clusters onde se encontram.



8.5.2 Alinhamento por Agrupamento

O mapeamento feito com os 36 complexos nos permite fazer um alinhamento de subclusters do lado inibidor, de forma a tentar compor o que seria um candidato a uma alça consenso, para as 6 regiões, tendo como base nosso complexo de referência 1PPF. O alinhamento pode ser visto na figura 8.122. Essa alça consenso poderia ajudar a fundamentar algo sonhado pelo saudoso Prof. Dr. Marcelo Santoro, que orientou a autora dessa tese inicialmente, com um projeto envolvendo a previsão e validação de inibidores mais universais, de amplo espectro de ação.

Infelizmente, não houve tempo hábil para uma análise estatística mais rigorosa desses alinhamentos, que poderia ter sido enriquecido, por exemplo, com um gráfico de logo. Essa análise mais detalhada também será alvo de um trabalho futuro.

Figura 8.122: Modelo Geral dos Clusters - composição lado inibidor. (A) Clusterização para enzimas tipo tripsina com diferentes inibidores. (B) Clusterização tipo subtilisina com diferentes inibidores. `[]` indicam um cluster, `{ }` indicam um componente desconexo, `||` indicam um *noloop*, ou seja, região no inibidor fora da alça inibitória. Resíduos de aminoácidos com letras minúsculas correspondem a resíduos de fronteira, ou seja, que estão numa posição intermediária entre os clusters.

```

(A) CLUSTER NUM:    [ 1 ] [ 2 ] [ 3 ] [ 45 ] [ 6 ]
1F2S (TRY/SQF):      [ I ] [cPi] [cR ] [ I ] [ W ]
1MCT (TRY/SQF):      [ RI ] [ Pi] [CR ] [ Iw ] [ W ]
1PPE (TRY/SQF):      [ V ] [cPi] [cR ] [ I ] [ EL ]
1PPF (ELY/OVO):      [ PA ] [ T ] [CL ] [ ER ] [ Y ]
1HJA (CHY/OVO):      [ PA ] [ T ] [CK ] [ ER ] [ Y ]
3SGB (SGY/OVO):      [ KPA ] [ TC] [ L ] [ ER ] [ Y ]
4SGB (SGY/POT):      [ A ] [ Pc] [cL ] [ N ] [ C ]
1Z7K (TRY/OVO): [IM] [ L ] [ N ] [CK ] [ ANn ] [ LN]
1ACB (CHY/EGL):      [ P ] [ T ] [VL ] [ DR ] [ Rly]
4B2B (TRY/EGL):      [ P ] [ T ] [VK ] [ DR ] [ LY ]
4H4F (CHY/EGL):      [ P ] [ T ] [L ] [ DR ] [ rLY]
1T80 (CHY/BPT):      [   ] [PC ] [TW ] [ AI ] [ IR ]
2F15 (TRY/BPT):      [   ] [PC ] [TK ] [ AI ] [ IR ]
1YC0 (HGY/BPT):      [   ] [RC ] [VR ] [ YF ] { P }
4DG4 (TRY/BPT):      [   ] [PC ] [TK ] [ AI ] [ IR ]
1TAW (TRY/APP):      [   ] [PC ] [TR ] [ AI ] [ M ]
1F18 (GRY/ECO):      [ I ] [ P ] [ D ] [   ] [   ] [PE]
1EZS (TRY/ECO): [P ] [ PV ] [ T ] [ R ] [ M ] [ ACP]
1XX9 (XAY/ECO):      [ V ] [ T ] [ RC] [ M ] { A }
1FLE (ELY/SKI):      [ LI ] [ C ] [ A ] [ M ] [ L ]
4DOQ (TRY/SLP):      [ Y ] [ C ] [QL ] [ M ] [ L ]
2Z7F (ELY/SLP):      [ Y ] [ C ] [QL ] [ M ] [ L ]
1K90 (TRY/SER):      [ Iv ] [ P ] [vK ] [ i ] [ ILL]
1OPH (TRY/SER):      [ A ] [ P ] [IRP] [ i ] [ I ]
3MYW (TRY/BBI):      [ CI ] [ T ] [CR ] [ P ] [ M ]
3VEQ (TRY/WCI):      [ Q ] [ FP] [ R ] [ F ] [ L ]
1KIG (TRY/TAP):      [ R ] [PKLI] [NY ] [   ] [   ]

(B) CLUSTER NUM:    [ 1 ] [ 2 ] [ 3 ] [ 45 ] [ 6 ]
1R0R: (CAN/OVO):      [ PA ] [ T ] [ L ] [ E ] [ Y ]
1TEC: (THN/EGL):      [ P ] [ T ] [ L ] [ DR ] [ L ]
2SEC: (CAN/EGL):      [ P ] [ T ] [ L ] [ DR ] [ L ]
1SBN: (BPN/EGL):      [ P ] [ T ] [ R ] [ D ] [ LY]
1LW6: (BPN/CL2):      [ I ] [ T ] [ M ] [ ER ] [ Y ]
1TMI: (BPN/CL2):      [ I ] [ T ] [ M ] [ E ] [ Y ]
1OYV: (CAN/WIP):      [PNA ] [ T ] [ F ] [ N ] [ Cc ]
1V5I: (BPN/POA):      [HVA ] [ A ] [   ] [   ] [   ]
4GI3: (CAN/GRG):      [ I ] [ P ] [ L ] [ I ] [ Y ]

```

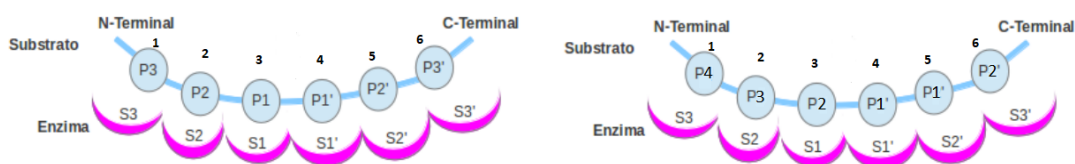
Mas, há algo que já podemos arriscar pré-concluir: todas as 6 regiões, seja no lado enzima, seja no lado inibidor, comportam átomos hidrofóbicos. Mesmo onde em alguns inibidores (como a própria 1PPF) houve uma preponderância de resíduos carregados, no lado enzima eles se assentavam em parte sobre um conjunto de átomos/resíduos hidrofóbicos.

O curioso é que, no clássico artigo de SCHECHTER & BERGER de 1967 [Schechter et al. (1967)], onde surgiram pela primeira vez as nomenclaturas usadas atualmente para descrever os sítios de ligação tanto enzima quanto inibidor, encontramos a seguinte pas-

sagem:

*In order to investigate the size of the active site of papain, we have studied its action on **40 diastereoisomeric peptides of alanine**, ranging in length from **Ala2** to **Ala6**. The variations observed in the rates of hydrolysis lead to the conclusion that papain, an endopeptidase, has a large active site which extends over about 25 Å and can be divided into **7 "subsites"**, each accommodating one amino acid residue of the peptide substrate. The subsites are located on both sides of the catalytic site, 4 on the one side and 3 on the other. (...) The substrate is visualized as fitting into the groove, binding to several subsites of specific geometry. Veja figura 8.123.*

Figura 8.123: Modelo clássico de mapeamento de sítios, definido por SCHECHTER e BERGER de 1967.



Fonte: Adaptada de <https://swift.cmbi.umcn.nl/teach/B2/LINK/NOOT_32.html>. [Schechter et al. (1967)].

Vejam só! Nesse clássico artigo, SCHECHTER & BERGER usaram polianinas de tamanhos variados (entre 2 e 7) para então identificar que seriam necessários 7 subsítios no lado enzima para comportar 6 Alaninas no lado inibidor. Polianinas têm uma carga hidrofóbica razoável, dada a presença de um grupo metil na cadeia lateral. Nosso trabalho identificou ao menos 6 subregiões no lado enzima, todas com potencial hidrofóbico característico. Logo, não é de se estranhar que SCHECHTER & BERGER tenham obtido sucesso na investigação sobre o sítio ativo da papaína usando um peptídeo todo hidrofóbico. E o fato da papaína não ser uma serino-peptidase, mas uma cisteíno-peptidase, sugere que também essas últimas possam ter um conjunto de regiões hidrofóbicas semelhantes às primeiras. Algo que certamente iremos investigar em trabalhos futuros.

SCHECHTER & BERGER achavam também que haveria um mapeamento de: para cada resíduo do inibidor, um sítio na enzima: $P_n \rightarrow S_n$ e $P'_n \rightarrow S'_n$. Nosso estudo das correspondências hidrofóbicas por agrupamento sugere que não é bem assim. Há de fato sítios hidrofóbicos $S_n \dots S_1, S'_1 \dots S'_n$ no lado peptidase, cada um deles podendo ser composto por átomos de diferentes resíduos, mas também átomos pertencentes a resíduos diferentes do lado inibidor podem ocupá-los. A relação resíduo inibidor x sítio enzima não é necessariamente um para um.

Acreditamos que esta talvez possa ser considerada uma das grandes contribuições desta tese: auxiliar numa redefinição e melhor caracterização dos clássicos sítios definidos por SCHECHTER & BERGER nas interfaces peptidases - inibidores proteicos.

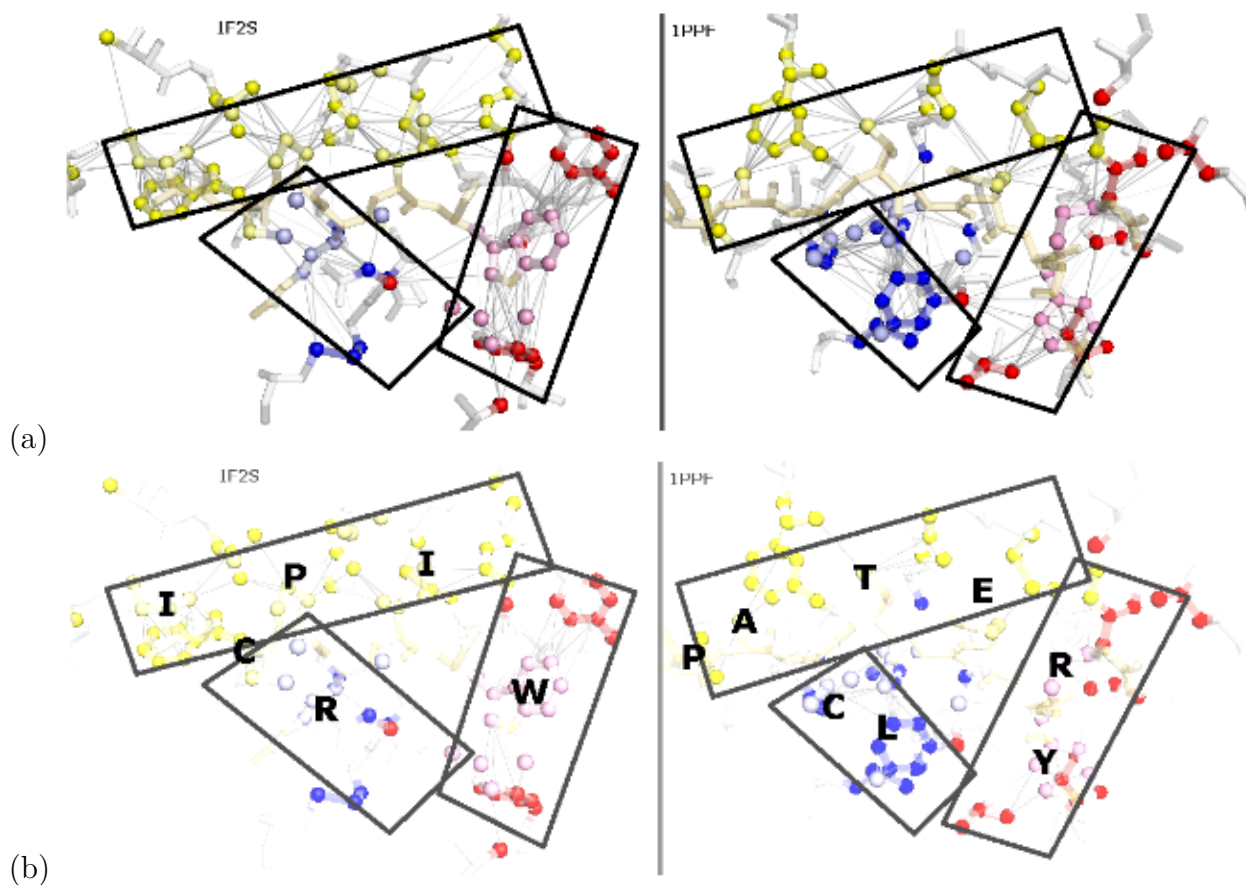
8.5.3 Superestrutura Hidrofóbica Anelar ou Semianelar

Olhando os resultados dos 36 complexos como um todo, observa-se uma clara correspondência/complementaridade entre átomos hidrofóbicos do inibidor com átomos hidrofóbicos da enzima. Acreditamos que isso responde de forma bem mais consistente as indagações feitas por um de nossos revisores do artigo da *Bioinformatics* [Gonçalves-Almeida et al. (2012)], o principal ponto de partida desta tese, no seu momento pós-Santoro.

E talvez tenhamos colhido mais do pretendíamos. O mapeamento das correspondências pode também ter revelado, de forma inesperada, uma superestrutura hidrofóbica inédita até onde pudemos pesquisar. A título de exemplo, vamos comparar dois complexos com tipo tripsinas envolvendo inibidores bem distintos: o inibidor MCTI-II da 1F2S (clan IE) com Ovomucoide da 1PPF (clan IA). Vemos na Figura 8.124 uma superestrutura que sugere um formato ANELAR ou SEMIANELAR.

A alça do inibidor está representada em *sticks*, na cor bege. As cores claras correspondem aos átomos do inibidor e as escuras aos átomos da enzima. Esse “anel” meio deformado envolveria o sítio catalítico, permitindo que substratos diversos possam se acomodar nas peptidases. Em Figura 8.124 (b), esse anel é mostrado em termos dos resíduos da alça do inibidor.

Figura 8.124: Exemplo do formato anelar do modelo geral (1F2S e 1PPF).



Na literatura, encontramos certo suporte para essa ideia. Devemos a [Bogan and Thorn (1998)], um dos artigos pioneiros na descoberta e caracterização das chamadas regiões *hot spots* nas interfaces proteínas-proteínas. Nesse contexto, *hot spots* são regiões com potencial de oferecer maior densidade de interações com seus ligantes. Ou seja, no cálculo do ΔG de *binding*, essas regiões respondem com maior peso. Borgan & Thorn empreenderam um estudo detalhado envolvendo a compilação de mudanças nos parâmetros de afinidade de ligação, de 2325 mutações de resíduos por alanina em um conjunto de cerca de 22 PDBids. Assim, foi possível mapear as contribuições energéticas dos resíduos mutacionados no ΔG de *binding*.

Um dos achados de Borgan & Thorn foi que a maior parte dos *hot spots* eram centrais, cercados por uma vizinhança de contatos pouco importantes do ponto de vista energético. Nas palavras dos autores:

A necessary condition for high-affinity interactions is the exclusion of bulk solvent (...) In many protein-small molecule interactions, this exclusion of solvent is achieved by burying the small molecule ligand in a deep pocket on the protein surface. However, since protein-protein interfaces are often flat, deep pockets are not usually available (...). In these interactions, exclusion of solvent from hot spot residues is achieved by surrounding set of contacts that

are energetically unimportant (analogous to an O-ring in a pipe fitting). This O-ring is required because it occludes bulk solvent, thereby generating suitable effective dielectric and solvation conditions for a hot spot (...).

Essa hipótese de um anel *cold spot* cercando um centro *hot spot* ficou conhecida na literatura como *O-ring Hypothesis*. O termo *O-ring* em inglês diz respeito aos “aneis de vedação” usados nas junções de canos para ajudar a evitar vazamentos nesses pontos de conexão.

Se a estrutura anelar hidrofóbica que estamos identificando na maior parte de nossos complexos tem relação com a *O-ring Hypothesis* é algo que precisa ser melhor investigado. Mas, é razoável admitir que um anel hidrofóbico poderia ter efeitos sobre as águas de solvatação nas interfaces peptidase e inibidor/substrato, de modo a interferir no padrão das flutuações de densidade local do solvente, algo que Chandler e outros autores [Chandler (2005), Willard and Chandler (2008)] vêm demonstrando ser fundamental para o papel da hidrofobicidade na promoção das complexações proteína-proteína.

Um anel desse tipo também ajudaria a dar maior versatilidade ao *binding* de substratos, ampliando o espectro de catálise das peptidases, uma vez que, exceção à glicina, todos os demais aminoácidos comportam um ou mais átomos apolares. A função do bolsão de especificidade seria mais a de introduzir um viés a favor de determinados tipos de cadeia lateral, dado atributos como carga e/ou volume. Todo bom enzimologista sabe o quão promíscuas as peptidases (e seus inibidores e substratos) podem ser.

Capítulo 9

Conclusões

A proposta fundamental desse trabalho foi mapear correspondências hidrofóbicas entre átomos nas interfaces de complexos entre serino peptidases e inibidores proteicos. Esse objetivo foi derivado do trabalho de [Gonçalves-Almeida et al. \(2012\)](#), que propôs a metodologia Hydropace, para identificar padrões conservados de interação hidrofóbica de serino peptidases com inibidores envolvidos na inibição cruzada. Nesta abordagem, foram identificadas regiões do lado da enzima e apontado a existência de *patches* complementares do lado do inibidor.

Inicialmente foi aplicado o Hydropace na base de dados utilizada no fenômeno da inibição cruzada, para delimitar as regiões hidrofóbicas do lado do inibidor. No entanto, a metodologia não se mostrou adequada para estabelecer a relação complementar *patches* hidrofóbicos enzima/inibidor. Com o agrupamento espectral, utilizando-se a matriz Laplaciana normalizada Lw e o algoritmo K -Medoides, identificou-se, cada agrupamentos envolvendo átomos hidrofóbicos da enzima e dos inibidores, evidenciando-se a complementariedade. Para tanto foi montado uma base de dados envolvendo 36 complexos, entre enzimas tipo tripsinas e tipo subtilisinas. Um modelo visual, baseado em projeção 3D delimitado por elipses, foi elaborado para evidenciar esses agrupamentos hidrofóbicos. Scripts pymol foram gerados para visualização da estrutura tridimensional do complexo e respectivas regiões hidrofóbicas.

Por meio da análise de distribuição das áreas de contato polar e apolar em 4 complexos em inibição cruzada envolvendo tipo tripsinas e tipo subtilisinas com inibidores ovomucoide e eglina, ficou evidente que o perfil apolar da superfície de contato é mais discriminante entre esses complexos e que o perfil polar tende à indiferenciação. Também pode ser inferido a partir dessa análise que a exclusão da área superficial parece ser ditada pela enzima e não pelo inibidor. Mas, a análise estendida para os 36 complexos, esse padrão inicial não ficou tão evidente.

Da análise dos clusters formados, foi obtido um modelo que tem bons agrupamentos, não necessariamente ótimos, variando entre 4 e 6 seis regiões hidrofóbicas, para a maioria dos casos da base composta por tipo tripsinas e tipo subtilisinas e inibidores protéicos distintos. Nesse sentido, uma varredura espectral é definida aqui como a possibilidade de se obter de 4 a 6 clusters com boa qualidade que definem a complementariedade

hidrofóbica entre regiões da enzima e do inibidor.

Um dos resultados desta tese foi indicar que todas as 6 regiões hidrofóbicas no lado enzima tem perfil nitidamente hidrofóbico. Isso encontra respaldo nos trabalhos pioneiros de SCHECHTER & BERGER [Schechter et al. (1967)] no mapeamento desses sítios, que fizeram uso de polialaninas, um peptídeo nitidamente hidrofóbico no todo. Mas, SCHECHTER & BERGER previram uma correspondência de um resíduo do inibidor/substrato para cada sítio da enzima. A visão ofertada por esta tese é diferente, dado que átomos de mais de um resíduo, tanto peptidase quanto inibidor, podem compor agrupamentos em que a correspondência não é de um para um.

Outro resultado importante desta tese foi a indicação de que pode haver nas interfaces das peptidases uma superestrutura anelar ou semianelar hidrofóbica. Tal resultado poderia ser condizente com a hipótese *O-ring* defendida por BORGAN & THORN [Bogan and Thorn (1998)] como um anel interações e energias com ligantes mais fracas cercando outras mais fortes ao centro, chamadas *hot spots*. Uma organização anelar hidrofóbica em enzimas poderia ter efeitos sobre as águas de solvatação, de modo a interferir no padrão das flutuações de densidade local do solvente, algo que Chandler e outros autores [Chandler (2005), Willard and Chandler (2008)] vêm demonstrando ser fundamental para as complexações proteína-proteína. Tal anelamento poderia ajudar a explicar também a promiscuidade de certas enzimas com relação a substratos/inibidores que experimentalistas constatarem empiricamente em seus laboratórios.

9.1 Perspectivas

Entende-se que os objetivos pretendidos, elencados no capítulo 2, foram alcançados. Dada os aspectos circunstanciais que esta tese enfrentou, com mudança de orientador, em função da passagem do nosso saudoso Prof. Marcelo Santoro, houve um tempo curto para a readequação de rumos e redefinição de novos objetivos. Mesmo assim, apesar do intenso trabalho feito nos últimos 18 meses, muitas coisas ficaram ainda por ser melhor compreendidas e modeladas.

9.1.1 Padrões de Distribuição de Contatos

A análise prévia de 4 complexos em inibição cruzada revelou um curioso padrões entre os perfis das distribuições APOLARES e POLARES. Mas, ao estender tal plotagem para os 36 complexos, os padrões ficaram ainda mais intrigantes. É necessário que se faça um estudo ainda mais aprofundado dos padrões gerados por essas distribuições, correlacionando-os, principalmente, à parâmetros termodinâmicos e cinético do *binding*.

9.1.2 Estatística das Métricas de Agrupamento

Outro ponto que ficou deficitário diz respeito a um melhor tratamento estatístico nas métricas que aferem qualidade dos agrupamentos. É preciso entender melhor as deficiências e insuficiências do coeficiente de silhueta, principalmente, como os subclusteres *noloop* podem interferir nessa métrica.

9.1.3 Alinhamento dos Agrupamentos

O alinhamentos dos agrupamentos tendo por base o lado inibidor foi um exercício prévio e manual. Além de fazer o mesmo com o lado enzima, é necessário automatizá-los, e pensar em modelos visuais e tratamento estatístico mais adequados, como a visualização em Logo e medidas de entropia informacional.

9.1.4 Visualização dos Modelos

Algumas limitações relacionadas ao modelo visual de apresentação dos clusteres foram identificadas e podem nortear a realização de trabalhos futuros. Nesse modelo, resíduos que se localizam na fronteira entre dois grupos por exemplo, são em alguns casos representados mais de uma vez, o que implica na necessidade de uma inspeção visual da estrutura tridimensional. Além disso, não é possível identificar no modelo o número

dos resíduos. Na comparação entre dois modelos, a identificação dos complexos deve ser alterada no código. Então, pretende-se melhorar o modelo de visualização implementando-se os seguintes recursos:

- visualização dos resíduos conectivos de clusteres e sua identificação no arquivo PDB;
- alinhamento de clusteres a partir dos centros geométricos de resíduos apontados como “sentinelas” ou referência;
- visualização a partir do modelo de clusteres da estrutura tridimensional do complexo e regiões hidrofóbicas, usando, por exemplo, um plugin para ferramentas de visualização;
- entrada de dados de identificação dos complexos por meio do PDB ID.

Com essas melhorias, pretende-se também fazer uma versão Web para a geração automática das regiões hidrofóbicas correspondentes entre enzima e inibidor.

9.1.5 Aspectos Dinâmicos

Para obter resultados que reflitam diretamente a influência da água na hidrofobicidade dos complexos protéicos e que conseqüentemente podem alterar a função das enzimas e as interações enzima/inibidor [Patel et al. (2012)], propõe-se a realização de simulação dinâmica e a partir dos arquivos PDB obtidos a aplicação do agrupamento espectral. Atualmente, as estruturas protéicas do repositório PDB contém pouca ou nenhuma informação sobre a água.

9.1.6 Superestrutura Hidrofóbica Anelar

Explorar o mais fundo possível nossa sugestão da existência de uma superestrutura hidrofóbica anelar e suas possíveis correlações com a hipótese *O-ring* de Borgan & Thorn [Bogan and Thorn (1998)].

9.1.7 Outras Peptidases

Outra meta é ampliar o escopo desse trabalho para outras famílias das serino peptidases e também outras peptidases tais como as cisteíno peptidases.

Referências

- Abboud, R. T., Ford, G. T., and Chapman, K. R. (2005). Emphysema in α 1-antitrypsin deficiency. *Treatments in respiratory medicine*, 4(1):1–8.
- Aggarwal, C. C. and Wang, H. (2010). A survey of clustering algorithms for graph data. In *Managing and mining graph data*, pages 275–301. Springer.
- Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2008). Data growth and its impact on the scop database: new developments. *Nucleic acids research*, 36(suppl 1):D419–D425.
- Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Engineering*, 2(2):101–113.
- 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):557–569.
- Baldwin, R. L. (1986). Temperature dependence of the hydrophobic interaction in protein folding. *Proceedings of the National Academy of Sciences*, 83(21):8069–8072.
- Barrett, A. J., Woessner, J. F., and Rawlings, N. D. (2012). *Handbook of proteolytic enzymes*, volume 1. Elsevier.
- Barrette-Ng, I. H., Ng, K. K.-S., Cherney, M. M., Pearce, G., Ryan, C. A., and James, M. N. (2003). Structural basis of inhibition revealed by a 1: 2 complex of the two-headed tomato inhibitor-ii and subtilisin carlsberg. *Journal of Biological Chemistry*, 278(26):24062–24071.
- Berman, H., Kleywegt, G., Nakamura, H., and Markley, J. (2014). The protein data bank archive as an open data resource. *Journal of Computer-Aided Molecular Design*, 28(10):1009–1014.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000a). The protein data bank. *Nucleic acids research*, 28(1):235–242.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000b). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

- Bickerton, G. R., Higuero, A. P., and Blundell, T. L. (2011). Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the piccolo database. *BMC bioinformatics*, 12(1):313.
- Bode, W. and Huber, R. (1992). Natural protein proteinase inhibitors and their interaction with proteinases. *European Journal of Biochemistry*, 204(2):433–451.
- 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. *The EMBO journal*, 5(10):2453.
- Bogan, A. A. and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9.
- Bondi, A. (1964). van der waals volumes and radii. *The Journal of physical chemistry*, 68(3):441–451.
- Brandes, U., Gaertler, M., and Wagner, D. (2007). Engineering graph clustering: Models and experimental evaluation. *ACM Journal of Experimental Algorithmics*, 12(1.1):1–26.
- Brewer, M. L. (2007). Development of a spectral clustering method for the analysis of molecular data sets. *Journal of chemical information and modeling*, 47(5):1727–1733.
- Campbell, T. (2007). Bioquímica clínica de aves. *THRALL, MA et al. Hematologia e bioquímica clínica veterinária. São Paulo: Roca*, pages 415–435.
- Chandler, D. (2005). Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–647.
- 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(5520):705–708.
- Chung, F. R. (1997). Spectral graph theory (cbms regional conference series in mathematics, no. 92).
- Colonna-Cesari, F. and Sander, C. (1990). Excluded volume approximation to protein-solvent interaction. the solvent contact model. *Biophysical journal*, 57(5):1103.
- Connolly, M. L. (1983). Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558.

- da Silveira, C. H., Pires, D. E., Minardi, R. C., Ribeiro, C., Veloso, C. J., Lopes, J. C., Meira, W., Neshich, G., Ramos, C. H., Habesch, R., et al. (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 74(3):727–743.
- Derache, C., Epinette, C., Roussel, A., Gabant, G., Cadene, M., Korkmaz, B., Gauthier, F., and Kellenberger, C. (2012). Crystal structure of greglin, a novel non-classical kazal inhibitor, in complex with subtilisin. *FEBS Journal*, 279(24):4466–4478.
- Dickerson, R. E. and Geis, I. (1983). *Hemoglobin: structure, function, evolution, and pathology*, volume 1983. Benjamin-Cummings Publishing Company.
- Dill, K. A. (1999). Polymer principles and protein folding. *Protein Science*, 8(06):1166–1180.
- Eisenberg, D. and McLachlan, A. D. (1986). Solvation energy in protein folding and binding.
- Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., and Scharf, M. (1995). The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, 16(3):273–284.
- Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584.
- Farady, C. J., Egea, P. F., Schneider, E. L., Darragh, M. R., and Craik, C. S. (2008). Structure of an fab–protease complex reveals a highly specific non-canonical mechanism of inhibition. *Journal of molecular biology*, 380(2):351–360.
- Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., and Mistry (2014). Pfam: the protein families database. *Nucleic acids research*, pages D222–D230.
- Flake, G. W., Tarjan, R. E., and Tsioutsoulis, K. (2004). Graph clustering and minimum cut trees. *Internet Mathematics*, 1(4):385–408.
- Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014). Scope: Structural classification of proteins–extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309.
- Frigerio, F., Coda, A., Pugliese, L., Lionetti, C., Menegatti, E., Amiconi, G., Schnebli, H. P., Ascenzi, P., and Bolognesi, M. (1992). Crystal and molecular structure of the

- bovine α -chymotrypsin-eglin c complex at 2.0 Å resolution. *Journal of molecular biology*, 225(1):107–123.
- Gao, M. and Skolnick, J. (2012). The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. *Proceedings of the National Academy of Sciences*, 109(10):3784–3789.
- Gibbs, J. W. (1873). *A method of geometrical representation of the thermodynamic properties of substances by means of surfaces*. Connecticut Academy of Arts and Sciences.
- Gillmor, S. A., Takeuchi, T., Yang, S. Q., Craik, C. S., and Fletterick, R. J. (2000). Compromise and accommodation in ecotin, a dimeric macromolecular inhibitor of serine proteases. *Journal of molecular biology*, 299(4):993–1003.
- Goncalves, W. R., Goncalves-Almeida, V. M., Arruda, A. L., Meira Jr, W., da Silveira, C. H., Pires, D. E., and de Melo-Minardi, R. C. (2015). Pdbest: A user-friendly platform for manipulating and enhancing protein structures. *Bioinformatics*, 31(17):2894–2896.
- Gonçalves-Almeida, V. (2011). *Hydropace: uma metodologia para análise de inibição cruzada em serino proteases através de centroides de regiões hidrofóbicas*. PhD thesis, ICB/UFMG.
- Gonçalves-Almeida, V., Pires, D. E., de Melo-Minardi, R. C., da Silveira, C. H., Meira, W., and Santoro, M. M. (2012). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342–349.
- Greenblatt, H. M., Ryan, C. A., and James, M. N. (1989). Structure of the complex of streptomyces griseus proteinase b and polypeptide chymotrypsin inhibitor-1 from russet burbank potato tubers at 2.1 Å resolution. *Journal of molecular biology*, 205(1):201–228.
- Hagen, L. and Kahng, A. B. (1992). New spectral methods for ratio cut partitioning and clustering. *Computer-aided design of integrated circuits and systems, iee transactions on*, 11(9):1074–1085.
- Harrell, E. M. (2015). *A Short History of Operator Theory*.
- Heifetz, A., Barker, O., Verquin, G., Wimmer, N., Meutermans, W., Pal, S., Law, R. J., and Whittaker, M. (2013). Fighting obesity with a sugar-based library: discovery of novel mch-1r antagonists by a new computational–vast approach for exploration of gpcr binding sites. *Journal of chemical information and modeling*, 53(5):1084–1099.

- Horn, J. R., Ramaswamy, S., and Murphy, K. P. (2003). Structure and energetics of protein–protein interactions: the role of conformational heterogeneity in omtky3 binding to serine proteases. *Journal of molecular biology*, 331(2):497–508.
- Hubbard, S. J. and Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2(1).
- Huber, R. and Bode, W. (1978). Structural basis of the activation and action of trypsin. *Accounts of Chemical Research*, 11(3):114–122.
- Ibrahim, B. S. and Pattabhi, V. (2004). Crystal structure of trypsin–turkey egg white inhibitor complex. *Biochemical and biophysical research communications*, 313(1):8–16.
- Inoue, K., Li, W., and Kurata, H. (2010). Diffusion model based spectral clustering for protein-protein interaction networks.
- Jain, A., Whitesides, G. M., Alexander, R. S., and Christianson, D. W. (1994). Identification of two hydrophobic patches in the active-site cavity of human carbonic anhydrase ii by solution-phase and solid-state studies and their use in the development of tight-binding inhibitors. *Journal of medicinal chemistry*, 37(13):2100–2105.
- Jamadagni, S. N., Godawat, R., and Garde, S. (2011). Hydrophobicity of proteins and interfaces: Insights from density fluctuations. *Annual review of chemical and biomolecular engineering*, 2:147–171.
- Janin, J., Chothia, C., Shabb, J., Ng, L., and Corbin, J. (1990). The structure of protein-protein recognition sites. *J. biol. Chem*, 265.
- Jin, L., Pandey, P., Babine, R. E., Gorga, J. C., Seidl, K. J., Gelfand, E., Weaver, D. T., Abdel-Meguid, S. S., and Strickler, J. E. (2005). Crystal structures of the fxa catalytic domain in complex with ecotin mutants reveal substrate-like interactions. *Journal of Biological Chemistry*, 280(6):4704–4712.
- Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20.
- Jones, S. and Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *Journal of molecular biology*, 272(1):121–132.
- Kannan, N. and Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *Journal of molecular biology*, 292(2):441–464.
- Kannan, R., Vempala, S., and Vetta, A. (2004). On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515.

- Karbalaei-Heidari, H. R., Amoozegar, M. A., Hajighasemi, M., Ziaee, A.-A., and Ventosa, A. (2009). Production, optimization and purification of a novel extracellular protease from the moderately halophilic bacterium *Halobacillus karajensis*. *Journal of industrial microbiology & biotechnology*, 36(1):21–27.
- Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.
- Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Prot. Chem.*, 14:1–63.
- Keskin, O., Ma, B., and Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of molecular biology*, 345(5):1281–1294.
- Kojima, S., Iwahara, A., and Yanai, H. (2005). Inhibitor-assisted refolding of protease: A protease inhibitor as an intramolecular chaperone. *FEBS letters*, 579(20):4430–4436.
- Korn, A. P. and Burnett, R. M. (1991). Distribution and complementarity of hydrophathy in mutisunit proteins. *Proteins: structure, function, and bioinformatics*, 9(1):37–55.
- Krowarsch, D., Cierpicki, T., Jelen, F., and Otlewski, J. (2003). Canonical protein inhibitors of serine proteases. *Cellular and Molecular Life Sciences CMLS*, 60(11):2427–2444.
- Laskowski, R. A., Luscombe, N. M., Swindells, M. B., and Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein science: a publication of the Protein Society*, 5(12):2438.
- Laskowski Jr, M. and Kato, I. (1980). Protein inhibitors of proteinases. *Annual review of biochemistry*, 49(1):593–626.
- Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4.
- Lehninger, A., Nelson, D., and Cox, M. (2006). *Princípios de Bioquímica. 4a*. Sarvier Editora de Livros Médicos Ltda.
- Li, J., Zhang, C., Xu, X., Wang, J., Yu, H., Lai, R., and Gong, W. (2007). Trypsin inhibitory loop is an excellent lead structure to design serine protease inhibitors and antimicrobial peptides. *The FASEB Journal*, 21(10):2466–2473.
- Liao, C.-S., Lu, K., Baym, M., Singh, R., and Berger, B. (2009). Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258.

- Lijnzaad, P., Berendsen, H. J., and Argos, P. (1996). A method for detecting hydrophobic patches on protein surfaces. *Proteins Structure Function and Bioinformatics*, (26):192–203.
- Lopez-Otin, C. and Bond, J. S. (2008). Proteases: multifunctional enzymes in life and disease. *Journal of Biological Chemistry*, 283(45):30433–30437.
- Loss, L. A., Sadanandam, A., Durinck, S., Nautiyal, S., Flaucher, D., Carlton, V. E., Moorhead, M., Lu, Y., Gray, J. W., Faham, M., et al. (2010). Prediction of epigenetically regulated genes in breast cancer cell lines. *BMC bioinformatics*, 11(1):305.
- MacQueen, J. (1965). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Math., Stat., and Prob*, pages 281–297.
- McPhalen, C. A. and James, M. N. (1988). Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin carlsberg and ci-2-subtilisin novo. *Biochemistry*, 27(17):6582–6598.
- Merops (2015). *Classification: three orthogonal approaches*. <https://www.ebi.ac.uk/merops/about/classification.shtml>.
- Mitsuya, H., Yarchoan, R., and Broder, S. (1990). Molecular targets for aids therapy. *Science*, 249(4976):1533–1544.
- Mittag, T. and Forman-Kay, J. D. (2007). Atomic-level characterization of disordered protein ensembles. *Current opinion in structural biology*, 17(1):3–14.
- Murphy, K. P. and Freire, E. (1992). Thermodynamics of structural stability and cooperative folding behavior in proteins. *Advances in Protein Chemistry*, 43:313–361.
- Nepusz, T., Sasidharan, R., and Paccanaro, A. (2010). Scps: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. *BMC bioinformatics*, 11(1):120.
- Neres, J., Brewer, M. L., Ratier, L., Botti, H., Buschiazzo, A., Edwards, P. N., Mortenson, P. N., Charlton, M. H., Alzari, P. M., Frasch, A. C., et al. (2009). Discovery of novel inhibitors of trypanosoma cruzi-trans-sialidase from in silico screening. *Bioorganic & medicinal chemistry letters*, 19(3):589–596.
- Neurath, H. (1989). Proteolytic processing and physiological regulation. *Trends in biochemical sciences*, 14(7):268–271.
- Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

- Olsen, K., Otte, J., and Skibsted, L. H. (2000). Steady-state kinetics and thermodynamics of the hydrolysis of β -lactoglobulin by trypsin. *Journal of agricultural and food chemistry*, 48(8):3086–3089.
- Paccanaro, A., Casbon, J. A., and Saqi, M. A. (2006). Spectral clustering of protein sequences. *Nucleic acids research*, 34(5):1571–1580.
- Page, M. and Di Cera, E. (2008). Serine peptidases: classification, structure and function. *Cellular and Molecular Life Sciences*, 65(7):1220–1236.
- Patel, A. J., Varilly, P., Jamadagni, S. N., Hagan, M. F., Chandler, D., and Garde, S. (2012). Sitting at the edge: How biomolecules use hydrophobicity to tune their interactions and function. *The Journal of Physical Chemistry B*, 116(8):2498–2503.
- Perkins, A. D. and Langston, M. A. (2009). Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC bioinformatics*, 10(Suppl 11):S4.
- Pettit, F. K., Bare, E., Tsai, A., and Bowie, J. U. (2007). Hotpatch: a statistical approach to finding biologically relevant features on protein surfaces. *Journal of molecular biology*, 369(3):863–879.
- Phuc, D. and Phung, N. T. K. (2010). Visualization of the similar protein structures using som neural network and graph spectra. In *Intelligent Information and Database Systems*, pages 258–267. Springer.
- Pires, D., Silveira, C., Santoro, M., and Meira Jr, W. (2007). Pdbest-pdb enhanced structures toolkit. In *Proceedings of the 3rd International Conference of Brazil Association for Bioinformatics*, page 39.
- Polgar, L. (2005). The catalytic triad of serine peptidases. *Cellular and molecular life sciences*, 62(19):2161–2172.
- Ponstingl, H., Henrick, K., and Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins: Structure, Function, and Bioinformatics*, 41(1):47–57.
- Puente, X., Snchez, L., Gutierrez-Fernandez, A., Velasco, G., and Lopez-Otin, C. (2005). A genomic view of the complexity of mammalian proteolytic systems. *Biochemical Society Transactions*, 33:331–334.
- Qin, G. and Gao, L. (2010). Spectral clustering for detecting protein complexes in protein–protein interaction (ppi) networks. *Mathematical and Computer Modelling*, 52(11):2066–2074.

- Radisky, E. S. and Koshland, D. E. (2002). A clogged gutter mechanism for protease inhibitors. *Proceedings of the National Academy of Sciences*, 99(16):10316–10321.
- Rao, M. B., Tanksale, A. M., Ghatge, M. S., and Deshpande, V. V. (1998). Molecular and biotechnological aspects of microbial proteases. *Microbiology and molecular biology reviews*, 62(3):597–635.
- Rawlings, N. and Salvesen, G. (2012). *Handbook of proteolytic enzymes*. Academic Press.
- Rawlings, N., Tolle, D., and Barrett, A. (2004). Evolutionary families of peptidase inhibitors. *Biochem. J*, 378:705–716.
- Rawlings, N. D. and Barrett, A. J. (1993). Evolutionary families of peptidases. *Biochem. J*, 290:205–218.
- Rawlings, N. D., Barrett, A. J., and Bateman, A. (2011). Asparagine peptide lyases a seventh catalytic type of proteolytic enzymes. *Journal of Biological Chemistry*, 286(44):38321–38328.
- Rawlings, N. D., Waller, M., Barrett, A. J., and Bateman, A. (2014a). Merops: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, 42:D503–D509.
- Rawlings, N. D., Waller, M., Barrett, A. J., and Bateman, A. (2014b). Merops: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, pages D503–D509.
- Read, R. J., Adams, P. D., Arendall, W. B., Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lütke, T., Otwinowski, Z., et al. (2011). A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412.
- Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of molecular biology*, 82(1):1–14.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Salleh, A., Rahman, N., and Basri, M. (2006). *New lipases and proteases*. Nova Science Pub Incorporated.
- Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.
- Schechter, I., Berger, A., et al. (1967). On the size of the active site in proteases. i. papain. *Biochemical and biophysical research communications*, 27(2):157.

- Shaffer, C. A. (2013). *Data Structures and Algorithm Analysis*. Dover Publications.
- Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905.
- Shia, S., Stamos, J., Kirchhofer, D., Fan, B., Wu, J., Corpuz, R. T., Santell, L., Lazarus, R. A., and Eigenbrot, C. (2005). Conformational lability in serine protease active sites: structures of hepatocyte growth factor activator (hgfa) alone and with the inhibitory domain from hgfa inhibitor-1b. *Journal of molecular biology*, 346(5):1335–1349.
- Shrake, A. and Rupley, J. (1973). Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology*, 79(2):351–371.
- Skillicorn, D. (2007). *Understanding complex datasets: data mining with matrix decompositions*. CRC press.
- 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.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Tsunemi, M., Matsuura, Y., Sakakibara, S., and Katsube, Y. (1996). Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 Å resolution. *Biochemistry*, 35(36):11570–11576.
- Van Dongen, S. M. (2001). Graph clustering by flow simulation.
- Verli, H. et al. (2014). *Bioinformática da Biologia à flexibilidade molecular*, volume 1. E-Book.
- Vermelho, A. B., Melo, A., Sá, M., Santos, A., D’Avila-Levy, C. M., Couri, S., and Bon, E. P. (2008). Enzimas proteolíticas: Aplicações biotecnológicas. *Enzimas em biotecnologia-Produção, aplicações e mercado*, pages 273–287.
- Voet, D. and Voet, J. G. (2013). *Bioquímica*. Artmed Editora.
- Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.
- Waugh, S. M., Harris, J. L., Fletterick, R., and Craik, C. S. (2000). The structure of the pro-apoptotic protease granzyme b reveals the molecular determinants of its specificity. *Nature Structural & Molecular Biology*, 7(9):762–765.

- Webb, E. et al. (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Number Ed. 6. Academic Press.
- Willard, A. P. and Chandler, D. (2008). The role of solvent fluctuations in hydrophobic assembly. *The Journal of Physical Chemistry B*, 112(19):6187–6192.
- Woods, S., Farrall, A., Procko, C., and Whitelaw, M. L. (2008). The bhlh/per-arnt-sim transcription factor sim2 regulates muscle transcript myomesin2 via a novel, non-canonical e-box sequence. *Nucleic acids research*, 36(11):3716–3727.
- Young, L., Jernigan, R., and Covell, D. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein science: a publication of the Protein Society*, 3(5):717.
- Yu, Z., Li, L., You, J., Wong, H.-S., and Han, G. (2012). Sc³: Triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(6):1751–1765.

Apêndice A

Método Silveira-Romanelli

Vamos considerar duas esferas, uma com centro na origem,

$$x^2 + y^2 + z^2 = R_1^2 \quad (\text{A.1})$$

e outra com centro sobre o eixo z dada por

$$x^2 + y^2 + (z - d)^2 = R_2^2 \quad (\text{A.2})$$

onde $d > 0$ e $R_1 > R_2$. A ideia é calcular a área de cada uma das calotas internas gerada pela interseção das duas esferas utilizando a integral dupla para superfícies genéricas:

$$A = \iint_D \sqrt{f_x(x, y)^2 + f_y(x, y)^2 + 1} dA. \quad (\text{A.3})$$

Para isso precisamos encontrar a região D e as funções que descrevem as superfícies.

Por [A.1](#) temos

$$x^2 + y^2 = R_1^2 - z^2.$$

Agora, substituindo em [A.2](#), obtemos

$$R_1^2 - z^2 + (z - d)^2 = R_2^2.$$

Isolando z na equação acima chegamos em

$$z = \frac{d^2 + R_1^2 - R_2^2}{2d}. \quad (\text{A.4})$$

substituindo o valor z na equação [A.1](#) obtemos a equação da circunferência que limita a região de integração D :

$$x^2 + y^2 = R_1^2 - \frac{d^2 + R_1^2 - R_2^2}{2d}. \quad (\text{A.5})$$

Isolando z na equação [A.1](#) obtemos a função, $z = f(x, y)$,

$$z = \sqrt{R_1^2 - x^2 - y^2}.$$

e suas derivadas parciais

$$\frac{\partial f}{\partial x} = -\frac{1}{2} \frac{1}{(R_1^2 - x^2 - y^2)^{\frac{1}{2}}} 2x$$

e

$$\frac{\partial f}{\partial y} = -\frac{1}{2} \frac{1}{(R_1^2 - x^2 - y^2)^{\frac{1}{2}}} 2y.$$

Substituindo na fórmula A.3 temos

$$\begin{aligned} A(S_1) &= \iint_D \sqrt{\frac{x^2}{R_1 - x^2 - y^2} + \frac{y^2}{R_1 - x^2 - y^2} + 1} dx dy \\ &= \iint_D \sqrt{\frac{x^2 + y^2 + R_1^2 - x^2 - y^2}{R_1 - x^2 - y^2}} dx dy \\ &= \iint_D \sqrt{\frac{R_1^2}{R_1 - x^2 - y^2}} dx dy \end{aligned}$$

Colocando, para simplificar a notação, $M = \frac{d^2 + R_1^2 - R_2^2}{2d}$ reescrevemos a integral acima em coordenadas polares e obtemos

$$\begin{aligned} A(S_1) &= \int_0^{2\pi} \int_0^{\sqrt{R_1^2 - M^2}} \sqrt{\frac{R_1^2}{R_1 - r^2}} r dr d\theta \\ &= \int_0^{2\pi} \int_0^{\sqrt{R_1^2 - M^2}} \frac{R_1}{\sqrt{R_1 - r^2}} r dr d\theta. \end{aligned}$$

Fazendo a substituição $u = R_1^2 - r^2$, a integral acima se torna

$$A(S_1) = -\frac{R_1}{2} \int_0^{2\pi} \int_{R_1^2}^{M^2} \frac{1}{\sqrt{u}} du d\theta$$

e, fazendo as contas, temos

$$A(S_1) = 2\pi[R_1^2 - R_1 M]. \quad (\text{A.6})$$

Da mesma forma, para a equação A.2, a área da superfície S_2 é dada por

$$A(S_2) = 2\pi[R_2^2 - R_2 M]. \quad (\text{A.7})$$

Juntando as equações A.6 com A.7, depois de algum algebrismo, obtem-se:

$$\begin{aligned} A_c(R_1, R_2, d) &= A(S_1) + A(S_2) \\ &= 2\pi(R_1^2 + R_2^2) - \pi(R_1 + R_2)d \left[1 + \left(\frac{R_1 - R_2}{d} \right)^2 \right]. \end{aligned}$$