



JOSÉ AUGUSTO SALIM

**Aplicação de técnicas de reconhecimento de padrões
usando os descritores estruturais de proteínas da
base de dados do software STING para
discriminação do sítio catalítico de enzimas**

CAMPINAS
2015



UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade Engenharia Elétrica e Computação

JOSÉ AUGUSTO SALIM

Aplicação de técnicas de reconhecimento de padrões usando os descritores estruturais de proteínas da base de dados do software STING para discriminação do sítio catalítico de enzimas

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Supervisor/*Orientador*: FERNANDO JOSÉ VON ZUBEN
Co-supervisor/*Co-orientador*: GORAN NESHICH

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DO ALUNO JOSÉ AUGUSTO SALIM, CO-ORIENTADA PELO PROF. DR. FERNANDO JOSÉ VON ZUBEN E PELO PROF. DR. GORAN NESHICH.

CAMPINAS

2015

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Elizangela Aparecida dos Santos Souza - CRB 8/8098

Sa33a Salim, José Augusto, 1986-
Aplicação de técnicas de reconhecimento de padrões usando os descritores estruturais de proteínas da base de dados do software STING para discriminação do sítio catalítico de enzimas / José Augusto Salim. – Campinas, SP : [s.n.], 2015.

Orientador: Fernando José Von Zuben.
Coorientador: Goran Neshich.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Aprendizado de máquina. 2. Catálise enzimática. 3. Reconhecimento de padrões. 4. Enzimas. I. Von Zuben, Fernando José, 1968-. II. Neshich, Goran. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Pattern recognition using structural protein descriptors from STING database to discriminate the active site of enzymes

Palavras-chave em inglês:

Machine learning

Catalysis

Pattern recognition

Enzymes

Área de concentração: Engenharia de Computação

Titulação: Mestre em Engenharia Elétrica

Banca examinadora:

Fernando José Von Zuben [Orientador]

Guilherme Palermo Coelho

Ricardo Aparicio

Data de defesa: 24-02-2015

Programa de Pós-Graduação: Engenharia Elétrica

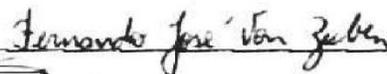
COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: José Augusto Salm

Data da Defesa: 24 de fevereiro de 2015

Título da Tese: "Aplicação de Técnicas de Reconhecimento de Padrões Usando os Descritores Estruturais de Proteínas da Base de Dados do Software STING para Discriminação de Sítio Catalítico de Enzimas"

Prof. Dr. Fernando José Von Zuben (Presidente)



Prof. Dr. Ricardo Aparício



Prof. Dr. Guilherme Palermo Coelho



RESUMO

As enzimas têm sua função determinada essencialmente por alguns resíduos específicos, denominados resíduos de aminoácidos catalíticos. A função de uma determinada proteína é mantida por milhares de anos de pressão seletiva que ocasionam a preservação de uma estrutura composta por padrões físicos, químicos e estruturais necessários para mantê-la. É frequente observar que enzimas quaisquer presentes em organismos distantemente relacionados exerçam exatamente a mesma função biológica e possuam o mesmo conjunto de resíduos de aminoácidos catalíticos, apesar de possuírem sequências proteicas muito dissimilares. Estes padrões que se conservaram por anos de evolução para manter a função das enzimas têm sido bastante estudados na literatura. Assim, o presente trabalho buscou identificar, dentre os descritores estruturais de proteínas (disponíveis na base de dados da plataforma Blue Star STING) aqueles de maior relevância para discriminar os resíduos de aminoácidos catalíticos dos não catalíticos, por meio do nanoambiente no qual estes se inserem. Buscou-se por modelos classificadores capazes de favorecerem uma interpretação de suas escolhas através de regras na forma SE-ENTÃO, compostas por descritores e seus respectivos valores. Regras foram extraídas para conjuntos de enzimas responsáveis pela catálise da mesma reação enzimática (mesma sub-subclasse EC), de forma a caracterizar o nanoambiente comum aos seus resíduos de aminoácidos catalíticos. Primeiramente, foram considerados apenas descritores estruturais de proteínas, ou seja, excluem-se descritores de conservação de estrutura primária. Esta opção foi feita com base no fato de a conservação de um determinado resíduo em uma determinada posição ser uma consequência (e não causa) de sua crucial função para a atividade de uma enzima. Buscou-se, portanto, compreender a fundo o “por que” de um resíduo ser conservado, utilizando uma “linguagem” puramente estrutural.

As doze mais representativas sub-subclasses EC foram escolhidas e regras foram extraídas de forma a caracterizar os resíduos de aminoácido catalíticos de seus membros. Os resultados obtidos variam de acordo com o número de amostras catalíticas disponíveis, sendo as classes com maior número de amostras as que resultaram em regras com maior capacidade de generalização. Ainda que a caracterização dos resíduos de aminoácidos catalíticos possa ser feita apenas com os dados disponíveis, a predição de novas amostras introduz diversos desafios discutidos neste trabalho. Diferentes técnicas de amostragem e seleção de atributos foram estudadas e o impacto de tais técnicas no treinamento é também discutido.

Novos descritores estruturais de proteína foram adicionados ao Blue Star STING, assim como foi feito o desenvolvimento de uma biblioteca de programação para facilitar e agilizar a extensão do conjunto de descritores do Blue Star STING.

ABSTRACT

The function of enzymes are determined by specific residues, called catalytic amino acids residues. The protein function is maintained for thousands of years of selective pressure which preserves in its structure many physical-chemical and structural patterns. Frequently, enzymes from distinct organisms exert exactly the same biological function due to similar catalytic amino acid residues, even with low sequence similarities.

The majority of catalytic amino acid residues prediction methods use sequence conservation features to provide classification. Seeking to understand these conserved patterns in enzyme structures, that even after years of evolution perform the same biological function, the present work searches to identify which protein structural descriptors (available in Blue Star STING platform) are capable of discriminating the amino acid catalytic residues from non-catalytic residues by means of their nanoenvironments properties. Therefore, we studied the use of classification methods available in the literature and STING structural protein descriptors to predict amino acid catalytic residues with no dependency of homologous enzymes. Considering methods capable of extracting IF-THEN rules composed of descriptors and their respective values, sets of rules were built to characterize the amino acid catalytic residues of enzymes catalyzing the same chemical reaction (same EC sub-subclass). Furthermore, it was considered only structural protein descriptors, i.e. no sequence conservation descriptor were considered. The conservation of certain amino acid in a given position is a consequence (not cause) of its crucial function for the enzyme activity. Therefore, the main purpose was to understand in depth the reason why a residue is preserved, employing a purely structural language.

Twelve most representative EC sub-subclasses were considered and rules were extracted to characterize the amino acid catalytic residues of their members. The results vary as the number of available structures for each sub-subclass increases. Once it is possible to characterize the amino acid residues of a set of enzymes catalyzing the same chemical reaction, the prediction of amino acid residues in new enzymes faces several challenges discussed in this work, been the major problem the lack of data for amino acid catalytic residues in available databases. Many different techniques as sampling and feature selection methods are employed to alleviate the imbalance of data, and their impact on training are discussed.

As a result, we also incorporate new structural protein descriptors to Blue Star STING and developed a new programming library to allow faster and easier extension of the Blue Star STING descriptors set.

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO	1
1.1 ENZIMAS E RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS	1
1.2 DESAFIOS E AVANÇOS NA IDENTIFICAÇÃO DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS EM ENZIMAS	4
1.3 NOMENCLATURA PARA CLASSIFICAÇÃO HIERÁRQUICA DE ENZIMAS – NÚMERO EC	5
1.4 DESCRITORES ESTRUTURAIS DE PROTEÍNAS E IDENTIFICAÇÃO <i>IN SILICO</i> DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS	6
CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA	9
2.1 PREDIÇÃO DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS A PARTIR DE SEQUÊNCIAS OU DE ESTRUTURAS PROTEICAS.....	9
2.2 MÉTODOS BASEADOS EM PROPRIEDADES ESTRUTURAIS DAS PROTEÍNAS	11
2.3 MÉTODOS BASEADOS EM PROPRIEDADES EXTRAÍDAS DAS SEQUÊNCIAS PROTEICAS	15
2.4 SEPARAÇÃO DE ENZIMAS SEGUNDO REAÇÕES CATALISADAS (NÚMERO EC) E DESAFIOS À PREDIÇÃO DOS CSR'S	17
CAPÍTULO 3 - CONJUNTO DE DADOS	21
3.1 BASE DE DADOS DE RESÍDUOS CATALÍTICOS – CATALYTIC SITE ATLAS	21
3.2 BASE DE DADOS DE NÚMEROS EC - PDBSPROTEC	22
3.3 BASE DE DADOS CATALÍTICA SEPARADA POR NÚMEROS EC.....	23
3.4 BLUE STAR STING, STING_DB E STING_RDB	28
3.5 BIBLIOTECA DE DESCRITORES ESTRUTURAIS DE PROTEÍNA – <i>STING DESCRIPTOR LIBRARY (SDL)</i>	30
3.6 BASE DE DADOS RELACIONAL DE DESCRITORES ESTRUTURAIS E FÍSICO-QUÍMICOS DE PROTEÍNAS: STING_RDB.....	33
CAPÍTULO 4 - DESCRITORES DE PROTEÍNAS	35
4.1 DESCRITORES FÍSICO-QUÍMICOS E ESTRUTURAIS DE PROTEÍNAS	35
4.1.1 <i>Potencial Eletrostático</i>	35
4.1.2 <i>Hidrofobicidade</i>	36
4.1.3 <i>Contatos e Densidade de Energia de Contatos</i>	37
4.1.4 <i>Acessibilidade</i>	38
4.1.5 <i>Ordem de Cross Link</i>	41
4.1.6 <i>Ordem de Cross Presence</i>	42
4.1.7 <i>Contatos não Usados</i>	42
4.1.8 <i>Densidade</i>	44
4.1.9 <i>Esponjicidade</i>	45
4.1.10 <i>Distâncias</i>	45
4.1.11 <i>Cavidades e Pockets</i>	45
4.1.12 <i>Curvatura</i>	46
4.1.13 <i>Patches Hidrofóbicos</i>	47
4.1.14 <i>Ligand Pocket e Water Contact Residues – LPR e WCR</i>	47
4.1.15 <i>Rotâmeros</i>	47
4.1.16 <i>Contatos proteína-ligante</i>	48
4.1.17 <i>Estrutura secundária</i>	49
4.1.18 <i>Energia de solvatação</i>	49
4.1.19 <i>Descritores baseados em grafos</i>	50
4.2 DESCRITORES DE CONSERVAÇÃO DE ESTRUTURA PRIMÁRIA	54
4.2.1 <i>Entropia e entropia relativa</i>	54
4.2.2 <i>Densidade de entropia 3D</i>	56
4.2.3 <i>Pressão evolutiva</i>	56
4.3 DESCRITORES DE VIZINHANÇA DOS AMINOÁCIDOS	57
4.3.1 <i>Descritores Ponderados pela Acessibilidade Relativa e Distâncias espaciais</i>	57
4.3.2 <i>Descritores Ponderados pela “força de contato”</i>	58
4.3.3 <i>Descritores de vizinhança sequenciais (Janelas deslizantes)</i>	59
4.3.4 <i>Descritores de vizinhança a partir de grafos de contatos</i>	59

CAPÍTULO 5 - APRENDIZADO DE MÁQUINA.....	61
5.1 MODELOS CLASSIFICADORES.....	62
5.2 DESBALANCEAMENTO ENTRE CLASSES	63
5.3 ÁRVORES DE DECISÃO	74
5.4 INDUÇÃO DE REGRAS	77
5.5 SELEÇÃO DE ATRIBUTOS.....	79
CAPÍTULO 6 - METODOLOGIA A SER ADOTADA COM BASE EM FERRAMENTAS DE APRENDIZADO DE MÁQUINA	81
6.1 CARACTERIZAÇÃO DE RESÍDUOS CATALÍTICOS	82
6.2 PREDIÇÃO DE RESÍDUOS CATALÍTICOS E SUBAMOSTRAGEM ALEATÓRIA.....	83
6.3 SELEÇÃO DE ATRIBUTOS E REDUÇÃO DE DIMENSIONALIDADE	84
6.4 SOBREAMOSTRAGEM DA CLASSE MINORITÁRIA	84
6.5 COMITÊS DE CLASSIFICADORES – ENSEMBLES	85
6.6 INCORPORANDO DESCRITORES DE CONSERVAÇÃO	86
6.7 MODELOS CLASSIFICADORES GERAIS	86
CAPÍTULO 7 - RESULTADOS E DISCUSSÕES	89
7.1 SELEÇÃO DE RESÍDUOS CATALÍTICOS UTILIZANDO ^{JAVA} PROTEIN DOSSIER (¹ PD).....	89
7.2 EXTENSÃO DO ¹ PD PARA A BUSCA DE REGRAS	98
7.3 ANÁLISE DE DADOS DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS	99
7.4 EXTRAÇÃO DE REGRAS PARA A CARACTERIZAÇÃO DE RESÍDUOS CATALÍTICOS	111
7.5 PREDIÇÃO DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS	117
7.6 IMPACTO DA SUBAMOSTRAGEM NO TREINAMENTO DOS CLASSIFICADORES.....	120
7.7 ENSEMBLES DE CLASSIFICADORES VIA <i>BOOSTING</i> E <i>BAGGING</i>	126
7.8 REDUÇÃO DE DIMENSIONALIDADE E SOBREAMOSTRAGEM VIA <i>SMOTE</i>	128
7.9 DESCRITORES DE CONSERVAÇÃO NA PREDIÇÃO DE RESÍDUOS CATALÍTICOS	130
7.10 DESAFIOS À PREDIÇÃO DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS	132
7.11 EXPANSÃO DO NÚMERO DE ENZIMAS VIA ALINHAMENTOS ESTRUTURAIS	137
7.12 CLASSIFICADOR GENÉRICO E COMPARAÇÃO COM OUTROS MÉTODOS NA LITERATURA.....	147
CAPÍTULO 8 - CONCLUSÃO	155
CAPÍTULO 9 - REFERÊNCIAS.....	159
CAPÍTULO 10 - APÊNDICES	171
APÊNDICE A - MODELO RELACIONAL STING_RDB	171
APÊNDICE B - MODELO UML DE CLASSES SIMPLIFICADO - STING DESCRIPTOR LIBRARY (SDL).....	172
APÊNDICE C - FUNÇÕES DE DISTRIBUIÇÃO ACUMULADO EMPÍRICA PARA OS DESCRITORES DO BLUE STAR STING	172
APÊNDICE D - LISTA DE REGRAS ENCONTRADAS UTILIZANDO-SE JPD PARA DIVERSAS FAMÍLIAS ENZIMÁTICAS DURANTES OS ANOS 2008 E 2009	181
APÊNDICE E - REGRAS PARA CARACTERIZAÇÃO DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS PARA TODAS AS SUB-SUBCLASSES ESTUDADAS.	187
APÊNDICE F - TESTES ESTATÍSTICOS PARA COMPARAÇÃO ENTRE CLASSIFICADORES	191
APÊNDICE G - DESEMPENHOS DOS CLASSIFICADORES APÓS PRÉ-PROCESSAMENTO (SUBAMOSTRAGEM E SOBREAMOSTRAGEM ALEATÓRIAS)	204
APÊNDICE H - DESEMPENHO ENSEMBLE DE CLASSIFICADORES (<i>BOOSTING</i> E <i>BAGGING</i>).....	209
APÊNDICE I - RELAÇÃO DOS DESCRITORES EM CADA SUB-SUBCLASSE EC APÓS REDUÇÃO DE DIMENSIONALIDADE	214

DEDICATÓRIA

Dedico este trabalho à minha família e aos amigos que fazem da minha simplória existência uma jornada repleta de amor e felicidade.

AGRADECIMENTOS

Agradeço,

ao Prof. Dr. Goran Neshich, pela dedicação e confiança em mim depositadas, essenciais para a realização deste trabalho.

Ao Prof. Dr. Fernando José Von Zuben pelos ensinamentos e orientação neste importante passo de engrandecimento pessoal e profissional.

A todos os membros do Grupo de Biologia Computacional da Embrapa Informática Agropecuária, pelas preciosas informações e contribuições diretas e indiretas para este trabalho, em especial Ivan Mazzoni, Inácio Yano, José Jardine, Fabio Moraes, Luiz Borro e Fabian Romero, meus agradecimentos.

Agradeço toda a minha família, em especial à Maria Cristina Salim, minha irmã, pela ajuda e apoio nas tarefas diárias que a vida em família nos exige e ensina.

Agradeço Alexandra Guidelli pela companhia e confiança em todas as ocasiões.

E principalmente aos meus pais por me apoiarem e contribuírem com esforços diversos, colocando meus sonhos sempre à frente dos deles.

Agradeço às instituições de fomento: FAPESP e CAPES pelo apoio financeiro para realização deste trabalho.

LISTA DE ILUSTRAÇÕES

FIGURA 1-1: DIFERENTES REGIÕES E RESÍDUOS PARA A ELASTASE DE LEUCÓCITO HUMANO (PDB: 1PPF CADEIA E)	3
FIGURA 1-2: CLASSIFICAÇÃO DAS ENZIMAS FOSFORRIBOSIL PIROFOSFATO SINTETASE (EC.2.7.6.1) SEGUNDO NOMENCLATURA EC. ESSAS ENZIMAS SÃO RESPONSÁVEIS PELA CATÁLISE DE REAÇÕES DE TRANSFERÊNCIA DE GRUPOS DE PIROFOSFATO ENTRE MOLÉCULAS DE ATP E RIBOSE 5-FOSFATO. A FUNÇÃO DA ENZIMA (REAÇÃO GENÉRICA CATALISADA) É DADA PELOS TRÊS PRIMEIROS DÍGITOS EC (SUB-SUBCLASSE), ENQUANTO QUE O QUARTO DÍGITO JUNTAMENTE COM OS OUTROS TRÊS DEFINEM A ESPECIFICIDADE DA ENZIMA (SUBSTRATO DE LIGAÇÃO).....	6
FIGURA 2-1: PROPENSIDADES DOS RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS OBTIDAS A PARTIR DO BANCO DE DADOS EMPREGADO NESTE TRABALHO. AMINOÁCIDOS QUE APRESENTAM VALORES MAIORES DO QUE ZERO PODEM SER CONSIDERADOS MAIS PROPENSOS PARA SEREM CATALÍTICOS DO QUE AQUELES COM VALORES INFERIORES OU IGUAIS A ZERO.....	11
FIGURA 3-1: ANOTAÇÃO DE RESÍDUO DE AMINOÁCIDO CATALÍTICO NO CSA DE ENZIMA MUTANTE (PDB 1DKI CADEIA A).....	22
FIGURA 3-2: GRÁFICO DO NÚMERO DE CADEIAS NÃO REDUNDANTES CONTIDAS NO CSA PARA DIFERENTES NÍVEIS DE SIMILARIDADE SEQUENCIAL...	24
FIGURA 3-3: DISTRIBUIÇÃO DAS ESTRUTURAS NO PDB POR CLASSE ENZIMÁTICA (PRIMEIRO NÍVEL EC).....	25
FIGURA 3-4: DIAGRAMA DO PROCESSO DE FILTRAGEM E ESCOLHA DAS CADEIAS UTILIZADAS NESTE ESTUDO.....	27
FIGURA 3-5: ESQUEMA DIDÁTICO DA COMPOSIÇÃO DE UM CRISTAL E APLICAÇÃO DE OPERAÇÕES DE TRANSLAÇÃO E ROTAÇÃO A PARTIR DA UMA UNIDADE ASSIMÉTRICA PARA OBTENÇÃO DA CÉLULA UNITÁRIA (UNIDADE REPETITIVA DO CRISTAL). FIGURA BASEADA NAQUELA DISPONÍVEL NOS RECURSOS EDUCACIONAIS NO WEBSITE DO PDB (HTTP://WWW.RCSB.ORG/PDB/101/STATIC101.DO?P=EDUCATION_DISCUSSION/LOOKING-AT-STRUCTURES/BIOASSEMBLY_TUTORIAL.HTML).....	32
FIGURA 4-1: ÁREA ACESSÍVEL AO SOLVENTE DE ACORDO COM A SUPERFÍCIE DE VAN DER WAALS.....	40
FIGURA 4-2: ESFERA Sonda centrada no carbono alfa do resíduo GLY-25 (magenta), que estabelece contatos com os resíduos THR-31, THR-49 e THR-50 (linhas tracejadas). Considerando segmentos de 5 aminoácidos o resíduo GLY-25 possui ordem de cross link igual a 2, pois os resíduos THR-49 e THR-50 estão a uma distância na sequência menor que o tamanho do segmento (5 aminoácidos). Assim, são contabilizados somente os dois contatos em amarelo.....	42
FIGURA 4-3: COMPLEXO FORMADO PELA PROTEÍNA ELASTASE DE LEUCÓCITOS HUMANOS E O TERCEIRO DOMÍNIO DO INIBIDOR DE OVOMUCÓIDE DE MELEAGRIS GALLOPAVO (CÓDIGO PDB: 1PPF).....	46
FIGURA 4-4: ÂNGULO DIEDRAIS DA CADEIA PRINCIPAL E LATERAL DO AMINOÁCIDO ARGININA (ÚNICO QUE SE ESTENDE ATÉ X ₅).....	48
FIGURA 4-5: (A) CENTRALIDADE POR PROXIMIDADE (CLOSENESS) PARA ENZIMA TRIPSINA DE SALMO SALAR (PDB: 1A0J CADEIA A), COM TRIÁDE CATALÍTICA DESTACADA EM CONTOURO PRETO (HIS57, ASP102, E SER195). RESÍDUOS DE AMINOÁCIDO CATALÍTICOS E OUTROS RESÍDUOS PRÓXIMOS AO CENTRO GEOMÉTRICO DA CADEIA APRESENTAM VALORES MAIS ELEVADOS DE CLOSENESS. (B) CENTRALIDADE DE PROXIMIDADE LOCAL PARA ENZIMA TRIPSINA DE SALMO SALAR (PDB: 1A0J CADEIA A). A CENTRALIDADE LOCAL CONSIDERA A VIZINHANÇA DE CADA RESÍDUO DE AMINOÁCIDO, SENDO INDIFERENTE À SUA LOCALIZAÇÃO EM RELAÇÃO AO CENTRO DE MASSA DA PROTEÍNA.....	52
FIGURA 5-1: MATRIZ DE CONFUSÃO PARA CLASSIFICAÇÃO BINÁRIA (DUAS CLASSES).....	63
FIGURA 5-2: ESQUEMA DIDÁTICO DA APLICAÇÃO DO SMOTE, OCASIONANDO A INTRODUÇÃO DE AMOSTRAS SINTÉTICAS DA CLASSE POSITIVA PRÓXIMAS A AMOSTRAS NEGATIVAS.....	67
FIGURA 5-3: DIVISÃO DOS DADOS SINTÉTICOS PARA DOIS DIFERENTES NÍVEIS DE COMPLEXIDADE UTILIZADOS EM (JAPKOWICZ, 2003).....	68
FIGURA 5-4: DESAFIOS NA CLASSIFICAÇÃO DE DADOS DESBALANCEADOS.....	69
FIGURA 5-5: PROPOSTA EVOLUTIVA PARA A SELEÇÃO DE CONJUNTOS DE TREINAMENTO BALANCEADOS. MODIFICAÇÃO DA ABORDAGEM PROPOSTA EM CANO ET AL. (2007). AE: ALGORITMO EVOLUTIVO; TS: TRAINING SET (CONJUNTO DE TREINAMENTO).....	71
FIGURA 5-6: COMPARAÇÃO ILUSTRATIVA DE ÁRVORES DE DECISÃO UNI E MULTIVARIADAS.....	75
FIGURA 7-1: REGRA ENCONTRA UTILIZANDO O ¹ PD PARA CINCO ESTRUTURAS DA ENZIMA ESTREPTOGRISINA B DE STREPTOMYCES GRISÉUS (EC: 3.4.21.81).....	92
FIGURA 7-2: REGRA ENCONTRADA UTILIZANDO O ¹ PD (QUADRO EM CINZA) PARA ENZIMAS POLIGALACTURONASE (EC: 3.2.1.15) DE FUSARIUM VERTICILLIOIDES (PDB: 1HG8:A), CHONDROSTEREUM PURPUREUM (PDB: 1K5C:A) E PECTOBACTERIUM CAROTOVORUM SUBSP. CAROTOVORUM (PDB: 1BHE:A).....	94
FIGURA 7-3: REGRA ENCONTRADA UTILIZANDO O ¹ PD PARA CINCO METALOPROTEASES. A REGRA SELECIONA A DÍADE CATALÍTICA COMPOSTA PELOS ÁCIDO GLUTÂMICO E METIONINA PARA AS ENZIMAS COM CÓDIGOS NO PDB: 1A85A (EC 3.4.24.34), 2AIGP (EC 3.4.24.46), 1ND1A (EC 3.4.24), 2DW2A (EC 3.4.24) E 1R54A (EC 3.4.24).....	96
FIGURA 7-4: EXTENSÃO DO ¹ PD PARA ENCONTRAR REGRAS PARA RESÍDUOS DE AMINOÁCIDO SELECIONADOS PELO USUÁRIO.....	101
FIGURA 7-5: FDA'S EMPÍRICAS PARA O DESCRITOR DE VOLUME DE POCKET DAS CLASSES DE RESÍDUOS CATALÍTICOS E NÃO CATALÍTICOS. VOLUME = 0 INDICA QUE O RESÍDUO NÃO SE ENCONTRA EM NENHUM POCKET. PELAS FUNÇÕES, PERCEBE-SE QUE MENOS DE 25% DOS CSR'S POSSUEM VOLUME DE POCKET = 0, ENQUANTO QUE POUCO MAIS DE 66% DOS NÃO CATALÍTICOS POSSUEM VOLUME DE POCKET = 0, INDICANDO A PREFERÊNCIA DE LOCALIZAÇÃO DOS CSR'S NESSAS REGIÕES DA SUPERFÍCIE DAS MOLÉCULAS.....	101
FIGURA 7-6: FDA'S DOS DESCRITORES OBTIDOS A PARTIR DE GRAFOS APRESENTAM SIGNIFICANTES DIFERENÇAS ENTRE AS DISTRIBUIÇÕES PARA OS RESÍDUOS CATALÍTICOS E DEMAIS RESÍDUOS.....	102
FIGURA 7-7: FDA'S PARA DESCRITORES DENSIDADE, CONSIDERANDO ESFERAS SONDAS DE RAIOS: 3, 4, 5, 6, 7, 10, 20 E 30Å, CENTRADAS NOS ÁTOMOS CA (A) E LHA (B), COM AS RESPECTIVAS DISTÂNCIAS MÁXIMAS ENTRE AS DUAS CURVAS (ESTATÍSTICA DE KOLMOGOROV-SMIRNOV).....	

NOTA-SE UMA MAIOR DIFERENÇA ENTRE AS DISTRIBUIÇÕES PARA RAIOS DA ESFERA SONDA IGUAL A 20Å (DENSIDADE: KS=0.3818 EM (A) E KS=0.3987 EM (B)).	104
FIGURA 7-8: FDA'S PARA DESCRITORES ESPONJICIDADE, CONSIDERANDO ESFERAS SONDAS DE RAIOS: 3, 4, 5, 6, 7, 10, 20 E 30Å, CENTRADAS NOS ÁTOMOS CA (A) E LHA (B), COM AS RESPECTIVAS DISTÂNCIAS MÁXIMAS ENTRE AS DUAS CURVAS (ESTATÍSTICA DE KOLMOGOROV-SMIRNOV). NOTA-SE UMA MAIOR DIFERENÇA ENTRE AS DISTRIBUIÇÕES PARA RAIOS DA ESFERA SONDA IGUAL A 20Å: KS=0.3866 EM (A) E KS=0.4138, EM (B).	105
FIGURA 7-9: NÚMERO DE CADEIAS POR FAIXA DE VOLUME. CADEIAS APRESENTAM MAIOR CONCENTRAÇÃO PARA VOLUME DE 20,000Å ³ ATÉ 60,000Å ³ , SENDO A MÉDIA DOS VOLUMES DAS CADEIAS IGUAL A 48,000Å ³ .	106
FIGURA 7-10: FDA'S PARA DESCRITORES DE POTENCIAL ELETROSTÁTICO COM OS RESPECTIVOS RESULTADOS DO TESTE DE KOLMOGOROV-SMIRNOV.	107
FIGURA 7-11: POTENCIAL ELETROSTÁTICO NO LHA CALCULADO CONSIDERANDO VIZINHANÇA DEFINIDA PELA REPRESENTAÇÃO DA PROTEÍNA COMO GRAFO NÃO DIRECIONADO (GN). METADE DOS CSR'S APRESENTAM VALORES DE EP INFERIOR A -2,150Kt/J/MOL, ENQUANTO APENAS 12.7% DOS NÃO-CSR APRESENTARAM TAIS VALORES.	108
FIGURA 7-12: HISTOGRAMAS PARA DESCRITOR DE CROSS PRESENCE ORDER, CONSIDERANDO VALORES DE RAIOS E SEGMENTOS PRÉ-CALCULADOS PELO BLUE STAR STING; RAIOS (R) = {3.5, 5.0 8.5}Å E SEGMENTOS (L) = {15, 20, 30}AA'S.	110
FIGURA 7-13: DESCRITORES DE CROSS PRESENCE ORDER CALCULADOS UTILIZANDO-SE ESFERAS DE RAIOS 10, 20 E 30Å E SEGMENTOS DE TAMANHOS 30, 40 E 50 AMINOÁCIDOS.	111
FIGURA 7-14: SOLUÇÕES NÃO-DOMINADAS ENCONTRADAS PELO MOEA-RIPPER.	115
FIGURA 7-15: SOLUÇÕES NÃO-DOMINADAS ENCONTRADAS PELO MOEA-RIPPER.	116
FIGURA 7-16: RELAÇÃO ENTRE O NÚMERO DE RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS E DESEMPENHOS DOS CLASSIFICADORES TREINADOS PARA AS DIFERENTES SUB-SUBCLASSES EC.	119
FIGURA 7-17: COMPARAÇÃO ENTRE O DESEMPENHO DOS CLASSIFICADORES TREINADOS COM SUBAMOSTRAGEM ALEATÓRIA (RUS), SOBREAMOSTRAGEM ALEATÓRIA (ROS) E ATRIBUIÇÃO DE PESOS (RWOS) PARA DIFERENTES PROPORÇÕES ENTRE AS CLASSES.	125
FIGURA 7-18: COMPARAÇÃO DE DESEMPENHO DOS CLASSIFICADORES TREINADOS SEM PRÉ-PROCESSAMENTO COM AQUELES QUE OBTIVERAM MÁXIMO DESEMPENHO UTILIZANDO SUBAMOSTRAGEM E SOBREAMOSTRAGEM, PARA CADA SUB-SUBCLASSE.	126
FIGURA 7-19: COMPARAÇÃO DE DESEMPENHO DOS CLASSIFICADORES TREINADOS UTILIZANDO RUSBOOST PARA DIFERENTES TAXAS DE SUBAMOSTRAGEM E O CLASSIFICADOR SEM PRÉ-PROCESSAMENTO (ORIGINAL).	127
FIGURA 7-20: COMPARAÇÃO DE DESEMPENHO DOS CLASSIFICADORES TREINADOS APLICANDO SMOTE APÓS REDUÇÃO DE DIMENSIONALIDADE E O CLASSIFICADOR SEM PRÉ-PROCESSAMENTO.	129
FIGURA 7-21: DESEMPENHO DOS MELHORES CLASSIFICADORES TREINADOS PARA QUATRO CONFIGURAÇÕES DIFERENTES: SEM PRÉ-PROCESSAMENTO, RUS, FS+SMOTE E FS+SMOTE COM CONSERVAÇÃO PARA CADA SUB-SUBCLASSE.	132
FIGURA 7-22: NÚMERO DE VEZES EM QUE OS ATRIBUTOS APARECERAM NAS POSIÇÕES DE 1 A 10 CONSIDERANDO AS 11 EXECUÇÕES PARA AS DUAS SUB-SUBCLASSES: GLICOSIDASES (EC.3.2.1) E METILTRANSFERASES (EC.2.1.1).	135
FIGURA 7-23: GRÁFICO TIPO BOXPLOT PARA GANHO DE INFORMAÇÕES DOS 10 MELHORES ATRIBUTOS, CONSIDERANDO AS 11 EXECUÇÕES DO ALGORITMO RIPPER PARA AS SUB-SUBCLASSES DAS GLICOSIDASES (EC.3.2.1) E METILTRANSFERASES (EC.2.1.1).	136
FIGURA 7-24: COMPARAÇÃO ENTRE O NÚMERO DE CADEIAS ENZIMÁTICAS PRESENTES NO BANCO DE DADOS UTILIZADO ANTES E APÓS A TRANSFERÊNCIA DE ANOTAÇÃO COM E SEM REDUNDÂNCIAS SEQUENCIAIS.	139
FIGURA 7-25: ALINHAMENTO ESTRUTURAL ENTRE SUBTILISINA DE BACILLUS LICHENIFORMIS (PDB: 1SBC, MAGENTA) E ELASTASE DE HOMO SAPIENS (PDB: 1PPF, CIANO), DEMONSTRA QUE AS ESTRUTURAS SÃO BASTANTE DIFERENTES, PORÉM CATALISAM A MESMA REAÇÃO QUÍMICA.	140
FIGURA 7-26: REGRA SELECIONANDO CSR'S DE ENZIMAS EVOLUTIVAMENTE CONVERGENTES: SUBTILISINA DE BACILLUS LICHENIFORMIS (PDB: 1SBC) E ELASTASE DE HOMO SAPIENS (PDB: 1PPF CADEIA E). APESAR DE GRANDES DIFERENÇAS SEQUENCIAIS E ESTRUTURAIS, AS PROPRIEDADES DO NANOAMBIENTE DOS CSR'S SÃO PRESERVADAS E AMBAS ESTRUTURAS, SENDO POSSÍVEL QUE A MESMA REGRA (QUADRO EM CINZA) SELECIONE OS CSR'S DE AMBAS AS ENZIMAS (DESTAQUE EM VERMELHO).	141
FIGURA 7-27: PROPENSIDADES DOS 20 TIPOS DE AMINOÁCIDOS EM SEREM CATALÍTICOS PARA AS 12 SUB-SUBCLASSES CONSIDERADAS ANTES E APÓS A TRANSFERÊNCIA DE ANOTAÇÃO.	145
FIGURA 7-28: DESEMPENHO DOS MELHORES CLASSIFICADORES TREINADOS APÓS A TRANSFERÊNCIA DE ANOTAÇÃO E O NÚMERO DE CADEIAS EM CADA SUB-SUBCLASSE ANTES E APÓS AS TRANSFERÊNCIAS.	146
FIGURA 7-29: COMPARAÇÃO DE DESEMPENHO DOS CLASSIFICADORES TREINADOS ANTES E APÓS A TRANSFERÊNCIA DE ANOTAÇÕES.	146
FIGURA 7-30: CURVAS PR (A) E (B) E ROC (C) E (D) PARA OS CLASSIFICADORES STING-CSR E STING-CSR-CONSERV NOS CONJUNTOS DE DADOS UTILIZADOS PARA COMPARAÇÃO.	149
FIGURA 7-31: CURVAS ROC DOS MÉTODOS EXIA, EXIA+PSSM, POOL(T), POOL(T+G), POOL(T+C), POOL(T+G+C), STING-CSR E STING-CSR-CONSERV.	153
FIGURA 7-32: CURVA ROC E PR PARA OS CLASSIFICADORES STING-CSR E STING-CSR-CONSERV APLICADOS AO CONJUNTO DE DADOS UB78, CONTENDO ENZIMAS SEM O SUBSTRATO (UNBONDED).	153

LISTA DE TABELAS

TABELA 3-1: SUB-SUBCLASSES EC SELECIONADAS NESTE ESTUDO PARA EXTRAÇÃO DE REGRAS E CARACTERIZAÇÃO DOS RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS.	28
TABELA 4-1: SUBCONJUNTO DE DESCRITORES ESTRUTURAIS DE PROTEÍNAS PRESENTES NO STING_RDB E UTILIZADOS NESTE TRABALHO.....	36
TABELA 4-2: ESCALA DE HIDROFOBICIDADE DEFINIDA POR RADZICKA & WOLFENDEN (1988) E KYTE & DOOLITTLE (1982) PARA CADA TIPO DE AMINOÁCIDO, COM SUAS CARACTERÍSTICAS FÍSICO-QUÍMICAS.	37
TABELA 4-3: DIFERENTES TIPOS DE CONTATOS INTERNOS ARMAZENADOS NO STING_RDB E SEUS RESPECTIVOS VALORES DE DISTÂNCIAS MÍNIMAS E MÁXIMAS PARA ESTABELECIMENTO DE CONTATO E ENERGIA DE INTERAÇÃO.	39
TABELA 4-4: ACESSIBILIDADES MÁXIMAS POR TIPO DE AMINOÁCIDO.	41
TABELA 4-5: NÚMERO MÁXIMO DE CONTATOS ESTABELECIDOS POR CADA TIPO DE AMINOÁCIDO.....	43
TABELA 4-6: TIPOS DE CONTATOS ENTRE ÁTOMOS DE DIFERENTES CLASSES PARA O CÁLCULO DO DESCRITOR DE CONTATOS PROTEÍNA-LIGANTE.	49
TABELA 5-1: MÉTRICAS DE AVALIAÇÃO MAIS APROPRIADAS PARA AVALIAR CLASSIFICADORES EM PROBLEMAS COM DESBALANCEAMENTO ENTRE CLASSES.....	65
TABELA 7-1: ALINHAMENTOS MÚLTIPLOS SEQUENCIAL E ESTRUTURAL PARA AS ENZIMAS DA FIGURA 7-1.....	93
TABELA 7-2: ALINHAMENTOS MÚLTIPLOS SEQUENCIAL E ESTRUTURAL PARA AS ENZIMAS DA FIGURA 7-2.....	95
TABELA 7-3: ALINHAMENTOS MÚLTIPLOS SEQUENCIAL E ESTRUTURAL PARA AS ENZIMAS DA FIGURA 7-3.....	97
TABELA 7-4: REGRA DESCREVENDO O NANOAMBIENTE DOS RESÍDUOS CATALÍTICOS PARA A HIV-1 INTEGRASE.	98
TABELA 7-5: SENSIBILIDADE (SENS.), PRECISÃO (PREC.) E NÚMERO DE REGRAS PARA TODAS AS SUB-SUBCLASSES EC ESTUDADAS ENCONTRADAS PELO RIPPER-UNPRUNUED (SEM FASE DE PODA) COM CF=2 E CF=6.	113
TABELA 7-6: REGRAS ENCONTRADAS PELO RIPPER SEM FASE DE PODA, UTILIZANDO TODO O CONJUNTO DE AMOSTRAS DE DUAS SUB-SUBCLASSES: EC's 2.1.1 E 3.2.1.	114
TABELA 7-7: PROPORÇÃO DOS RESÍDUOS DE AMINOÁCIDOS CATALÍTICOS COBERTOS PELAS REGRAS ENCONTRADAS PARA AS SUB-SUBCLASSES (LINHAS) QUANDO APLICADAS A OUTRAS SUB-SUBCLASSES(COLUNAS).	117
TABELA 7-8: VALORES MÉDIOS PARA SENSIBILIDADE (SENS.), PRECISÃO (PREC.) E F-MEASURE COM RESPECTIVOS DESVIOS PADRÕES (EM PARÊNTESES) DAS VALIDAÇÕES CRUZADAS UTILIZANDO RIPPER E C4.5 PARA TODAS AS SUB-SUBCLASSES EC.	119
TABELA 7-9: DESEMPENHO DOS CLASSIFICADORES TREINADOS COM RIPPER APÓS SUBAMOSTRAGEM DA CLASSE NEGATIVA. VALORES DE SENSIBILIDADE (S), PRECISÃO (P) E F-MEASURE (F) MÉDIOS, COM RESPECTIVOS DESVIOS PADRÕES, DURANTE VALIDAÇÃO CRUZADA. EM DESTAQUE, PROPORÇÕES QUE LEVARAM AOS MAIORES F-MEASURES POR SUB-SUBCLASSE.....	121
TABELA 7-10: DESEMPENHO DOS CLASSIFICADORES TREINADOS COM RIPPER APÓS SUBAMOSTRAGEM DA CLASSE NEGATIVA. VALORES DE SENSIBILIDADE (S), PRECISÃO (P) E F-MEASURE (F) MÉDIOS, COM RESPECTIVOS DESVIOS PADRÕES, DURANTE VALIDAÇÃO CRUZADA. EM DESTAQUE, PROPORÇÕES QUE LEVARAM AOS MAIORES F-MEASURES POR SUB-SUBCLASSE.....	122
TABELA 7-11: COMPARAÇÃO ENTRE STING-CSR E STING-CSR-CONSERV. COM OS MÉTODOS POOL(T+G) E POOL(T+G+C) NO CONJUNTO DE DADOS POOL160.....	150
TABELA 7-12: COMPARAÇÃO ENTRE STING-CSR E STING-CSR-CONSERV COM OS MÉTODOS EXIA E EXIA+PSSM EM VÁRIOS CONJUNTOS DE DADOS.	152
TABELA 7-13: COMPARAÇÃO ENTRE STING-CSR E STING-CSR-CONSERV E O MÉTODO C&P (CILIA & PASSERINI, 2010) EM VÁRIOS CONJUNTOS DE DADOS.	152

LISTA DE ABREVIATURAS E SIGLAS

ASA: Accessible Surface Area
AUCROC: Area Under Curve ROC
CA: Carbono alfa
CED: Contact Energy Density
CSA: Catalytic Site Atlas
CSR: Catalytic Site Residue
C β Carbono beta
DNA: Deoxyribonucleic Acid
EC: Enzyme Comission
FDA: Função de Distribuição Acumulada
FS: Feature Selection
GN: Graph Neighborhood
^JPD: Java Protein Dossier
LHA: Last Heavy Atom
PDB: Protein Data Bank
PRC: Precision-Recall Curve
RMN: Ressonância Magnética Nuclear
RNA: Rinonucleic Acid
ROC: Receiver Operating Characteristic
ROS Random OverSampling
RUS: Random UnderSampling
RWOS: Random Weighted OverSampling
SDL: STING Descriptor Library
STING_DB: STING Database
STING_RDB: STING Relational Database
SVM: Support Vector Machine
VD: Voronoi Diagram
WNA: Weighted Neighbor Average

Capítulo 1 - Introdução

Responsáveis por grande parte do funcionamento das células em todos os organismos vivos e processos metabólicos diversos, proteínas são objeto de estudo de diversas áreas do conhecimento. Conhecer e compreender os mecanismos envolvidos nesses processos representam objetivos desafiadores, demandando o desenvolvimento de novas técnicas e abordagens para investigação e análise das características das proteínas. Com o avanço da Biologia Computacional, permitindo o estudo *in silico* de macromoléculas, estudos em larga escala puderam ser desenvolvidos com drástica redução de tempo e custos.

1.1 Enzimas e resíduos de aminoácidos catalíticos

Compostas por cadeias polipeptídicas formadas pela combinação de 22 diferentes tipos de aminoácidos naturais, sendo 20 destes codificados universalmente pelo código genético, proteínas diferem entre si em diversos aspectos, tais como tamanho das cadeias peptídicas, sequências de aminoácidos, conformações tridimensionais, propriedades químicas dos aminoácidos que as compõem, entre outros. No caso dos aspectos estruturais de uma proteína, estes são classificados em quatro categorias: estrutura primária, que refere-se à sequência de aminoácidos ou sequência proteica; estrutura secundária, que representa padrões estruturais locais formados por ligações de hidrogênio, sendo as formas mais comuns folhas Beta, hélices Alfa e laços (*loops*); estrutura terciária, que corresponde à organização espacial tridimensional da proteína conforme coordenadas atômicas, e amplamente designadas somente pelo nome de estrutura proteica; e, por fim, estrutura quaternária, envolvendo a aglomeração de várias moléculas proteicas formando um complexo multiproteico (Sharma, 2009; Karp, 2008).

A predição da função de proteínas não anotadas, ou seja, com nenhuma ou pouca informação a respeito do papel biológico, ciclos metabólicos envolvidos, fontes (organismos), interações com outras moléculas, resíduos de aminoácidos funcionais, etc., é uma das tarefas principais da biologia computacional. Conhecer a função de uma proteína é importante porque permite o entendimento sistêmico de como a biomolécula interage com outros compostos químicos, e de como estas interações, que acontecem em nível molecular, influenciam no comportamento de um organismo nos vários níveis de estrutura intra e extracelular e nas interações entre organismos, e.g. patógeno e hospedeiro.

Enzimas são proteínas importantes envolvidas em uma grande variedade de funções bioquímicas através da catálise de reações relacionadas ao metabolismo de todos os organismos vivos (Freilich, et al., 2005; Kunik, et al., 2007). Nas reações enzimáticas, moléculas chamadas substratos (reagentes) são

convertidas pela catálise enzimática em moléculas diferentes: os produtos. O substrato deve ser capaz de se ligar de forma específica à enzima que, por meio desta interação, facilita a transformação do substrato em produto através da redução da energia de ativação da reação. Dessa forma, a estrutura do catalisador biológico deve favorecer o conjunto de interações que permitam a ligação ao substrato, expondo grupos químicos capazes de interagir com o mesmo e formando, transitoriamente, um complexo enzima-substrato. O **sítio ativo** ou **sítio funcional** (do inglês *active site*) é uma pequena região da proteína onde ocorrerá a reação catalisada pela enzima. Esta é, portanto, a região da enzima que contém os resíduos de aminoácidos capazes de interagir com o substrato. É nesse sítio, também, que estão localizados os resíduos de aminoácidos que diretamente participam da ruptura e estabelecimento de ligações químicas que resultam na formação do produto. Estes resíduos denominam-se **grupos de aminoácidos catalíticos** ou **resíduos de aminoácidos catalíticos** (*Catalytic Site Residues – CSR* em língua inglesa), e são atribuídos a estes a **função** desempenhada por uma enzima. Assim, o sítio ativo é fisicamente mais volumoso e abrange maior número de resíduos do que o conjunto de resíduos de aminoácidos catalíticos.

Quase todos os processos celulares necessitam de enzimas para ocorrerem em taxas significantes. Considerando que enzimas são extremamente seletivas para seus substratos, acelerando apenas uma reação específica, e raramente mais do que uma, o conjunto de enzimas produzidas na célula determina quais vias metabólicas ocorrem naquela célula.

Similarmente, podemos definir os resíduos que participam diretamente da interação de duas ou mais proteínas como **resíduos (formadores) de interface** (IFR, do inglês *Interface Forming Residues*). Estes são responsáveis pela ligação entre duas ou mais proteínas de forma que o complexo resultante possa desempenhar seu papel biológico. No caso de enzimas, os resíduos formadores de interface podem determinar a especificidade ao substrato, enquanto que a função das enzimas é determinada pelo conjunto de resíduos de aminoácidos catalíticos (Ribeiro, et al., 2010).

Na Figura 1-1, é mostrado o sítio catalítico para a Elastase de leucócito humano (PDB: 1PPF:E), uma Serino protease com tríade catalítica formada pelos resíduos de aminoácidos His57, Asp102 e Ser195. Embora não exista uma definição totalmente congruente e consistente na literatura para resíduos catalíticos, optou-se por utilizar as seguintes definições (Barlett, et al., 2002):

- Diretamente envolvidos no mecanismo catalítico, e.g. nucleófilo.
- Exercem um efeito em outro resíduo ou molécula de água que está diretamente envolvida no mecanismo catalítico auxiliando a catálise (e.g. através de ação eletrostática ou ácido-base).
- Estabilização da transição de estados intermediários.
- Exerce um efeito em um substrato ou cofator que auxilia a catálise.

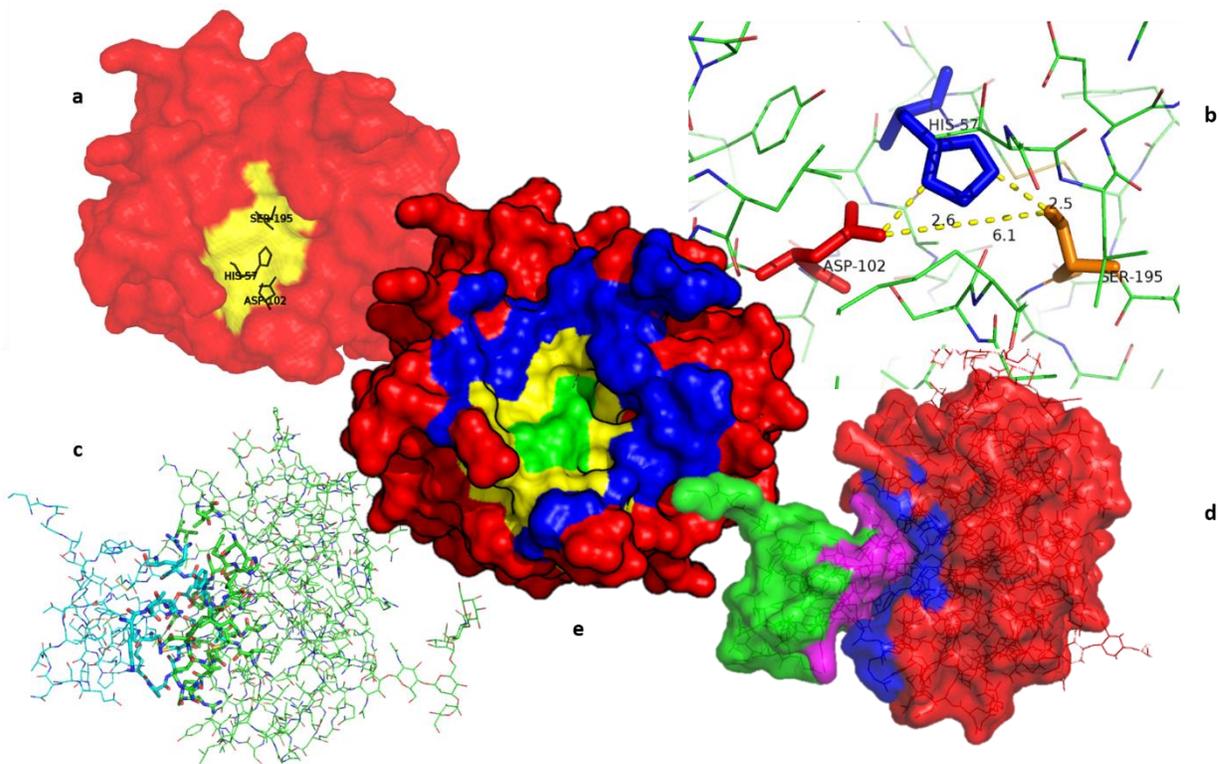


Figura 1-1: Diferentes regiões e resíduos para a elastase de leucócito humano (PDB: 1PPF cadeia E) (a) Superfície molecular em vermelho e sítio ativo em amarelo, com os resíduos catalíticos identificados em contorno preto. (b) Triade catalítica formada pelos resíduos His57, Asp102 e Ser195. (c) e (d) Resíduos formadores de interface entre a elastase de leucócito humano e o terceiro domínio do inibidor ovomucóide de *Meleagris gallopavo* (PDB: 1PPF cadeia I). As regiões marcadas em azul e rosa em (d) compõem o conjunto de resíduos formadores de interface das cadeias “E” e “I”, respectivamente. (e) Resíduos de interface em azul, sítio ativo em amarelo, resíduos catalíticos em verde e resíduos da superfície livre em vermelho. (Figura gerada com software PyMOL (Schrödinger, 2010)).

Fornecer anotação funcional para a vasta quantidade de dados de sequências e estruturas proteicas geradas por tecnologias de larga escala é uma das principais tarefas da era pós-genômica (Ofraan, et al., 2005; Shapiro & Harris, 2000; Gutteridge, et al., 2003; Kristensen, et al., 2008; Chitale & Kihara, 2011). A determinação experimental da função das proteínas é um desafio. Realizar ensaios para identificar a função de todas as proteínas ainda não caracterizadas é, no momento, inviável do ponto de vista prático. Assim, ferramentas computacionais desempenham um papel importante no atendimento de tais demandas.

A identificação dos resíduos catalíticos de uma enzima é um passo importante no entendimento do seu papel biológico e suas aplicações, particularmente devido ao fato de que um pequeno número de resíduos diretamente participa da catálise. Os arranjos espaciais bem como as propriedades físicas, químicas e físico-químicas destes resíduos determinam a reação química catalisada pela enzima. Um melhor entendimento deste complexo processo tem um significativo impacto no desenvolvimento de novas drogas, bem como na identificação de desordens genéticas e engenharia de proteínas com novas

funções (Ma, et al., 2001; Madabushi, et al., 2002).

1.2 Desafios e avanços na identificação de resíduos de aminoácidos catalíticos em enzimas

Métodos experimentais para identificação dos resíduos catalíticos em enzimas envolvem o estudo de regiões altamente conservadas nas sequências, o que exige a existência de homólogos. Duas ou mais sequências proteicas são consideradas homólogas quando compartilham uma origem evolutiva, sendo este um conceito biológico qualitativo, de forma que não é possível quantificar o quanto duas sequências são homólogas. Por outro lado, a similaridade sequencial entre sequências é uma informação quantitativa e avalia o quão parecidas duas ou mais sequências são entre si, sendo geralmente utilizada uma porcentagem de similaridade. Assim, se duas sequências possuem similaridade sequencial de 80%, isto indica que 80% de seus aminoácidos são compartilhados. Muitas vezes utiliza-se a similaridade sequencial para inferir se duas ou mais sequências são homólogas devido ao alto grau de concordância entre suas sequências. No entanto, nem sempre essa suposição é válida, sendo possível que duas sequências altamente similares tenham se originado de ancestrais diferentes, configurando uma evolução convergente.

Alinhamentos sequenciais, por sua vez, consistem em encontrar regiões similares em duas ou mais sequências, de forma que possam ser consequência de importâncias funcionais, estruturais ou evolutivas. Eventualmente, espaçamentos (*gaps*) podem ser incluídos entre os aminoácidos para que sequências semelhantes possam ser alinhadas. Em teoria da informação, a distância de edição ou distância de Levenshtein (Levenshtein, 1966), consistindo basicamente em encontrar o número mínimo de operações de forma a transformar uma sequência de caracteres em outra, considerando operações de inserção, remoção ou substituição de caracteres, é um processo análogo ao alinhamento sequencial.

Dessa forma, regiões onde os aminoácidos são perfeitamente alinhados (ou de alguma forma similares) são consideradas para uma análise mais detalhada, uma vez que esta ocorrência é um grande indício de sua importância para a proteína. O método experimental largamente empregado na identificação de resíduos catalíticos, tidos como essenciais para a catálise, são experimentos de mutagênese sítio específico ou dirigido (*site-directed mutagenesis* – na língua inglesa). Estes experimentos permitem a substituição, deleção e inserção de qualquer resíduo de aminoácido através de modificações específicas e intencionais na sequência de DNA de um gene e expressão correspondente da proteína mutante em célula hospedeira adequada (Hutchison, et al., 1978). Apesar de existirem outros métodos (e.g. análises cinéticas de reações; extrapolação por homologia), experimentos de mutagênese representam uma técnica excelente e universalmente adotada para sondar a importância de resíduos

específicos na catálise enzimática (Knowles, 1987). Mesmo com o avanço de técnicas de cristalografia e ressonância magnética nuclear, o entendimento das relações entre a estrutura e reações químicas catalisadas pelas enzimas ainda apresenta-se como um desafio à extrapolação do mecanismo catalítico considerando somente a conformação tridimensional destas biomoléculas. A realização de experimentos para detectar os mecanismos catalíticos muitas vezes é cara e demanda tempo, quando não é proibitiva. Dessa forma, métodos computacionais podem ajudar a encontrar respostas para tais desafios.

Os fatores que determinam a funcionalidade do sítio ativo de uma proteína são muito complexos e dependem de sua estrutura tridimensional, além de suas propriedades bioquímicas e biofísicas. Segundo Bobadilla e colaboradores (Bobadilla, et al., 2007) sítios funcionais podem ser entendidos como uma redondeza físico-química que acompanha uma função, em contraposição a um grupo de resíduos fixos. Assim, procura-se uma caracterização mais genérica do sítio ativo a partir de descritores estruturais de proteínas, sem considerar informações derivadas de alinhamentos sequenciais (conservação), que seja útil para a previsão dos resíduos de aminoácidos catalíticos.

Enzimas de função similar, evolutivamente relacionadas ou não, têm demonstrado compartilhar características de sequência e estrutura. Compreender a ligação entre estas características e a função da proteína é importante para o desenvolvimento de métodos de predição e compreensão da função proteica (Bray, et al., 2009). Dessa forma, o nanoambiente deste conjunto de resíduos de aminoácidos catalíticos tem suas características determinadas pela composição e geometria dos seus membros e é altamente específica, no sentido que pode ser descrito através de propriedades da estrutura, geometria, sequência e propriedades físicas, químicas e físico-químicas (Neshich, et al., 2014).

1.3 Nomenclatura para classificação hierárquica de enzimas – Número EC

Compondo grande parte do proteoma, enzimas são categorizadas de acordo com o sistema *Enzyme Commission number* (EC number), uma nomenclatura hierárquica utilizada pelo *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology* (NC-IUBMB) (Webb, 1992) que atribui um identificador único composto por quatro dígitos a cada classe de reação enzimática. O primeiro dígito representa a classe da reação catalisada: EC.1) Oxirredutases; EC.2) Transferases; EC.3) Hidrolases; EC.4) Liases; EC.5) Isomerases; e EC.6) Ligases. O segundo e terceiro dígitos (i.e. subclasse e sub-subclasse) especificam a reação catalisada, enquanto o quarto dígito define a especificidade ao substrato. A Figura 1-2 ilustra essa classificação para as enzimas Fosforribosil pirofosfato sintetase (EC.2.7.6.1).

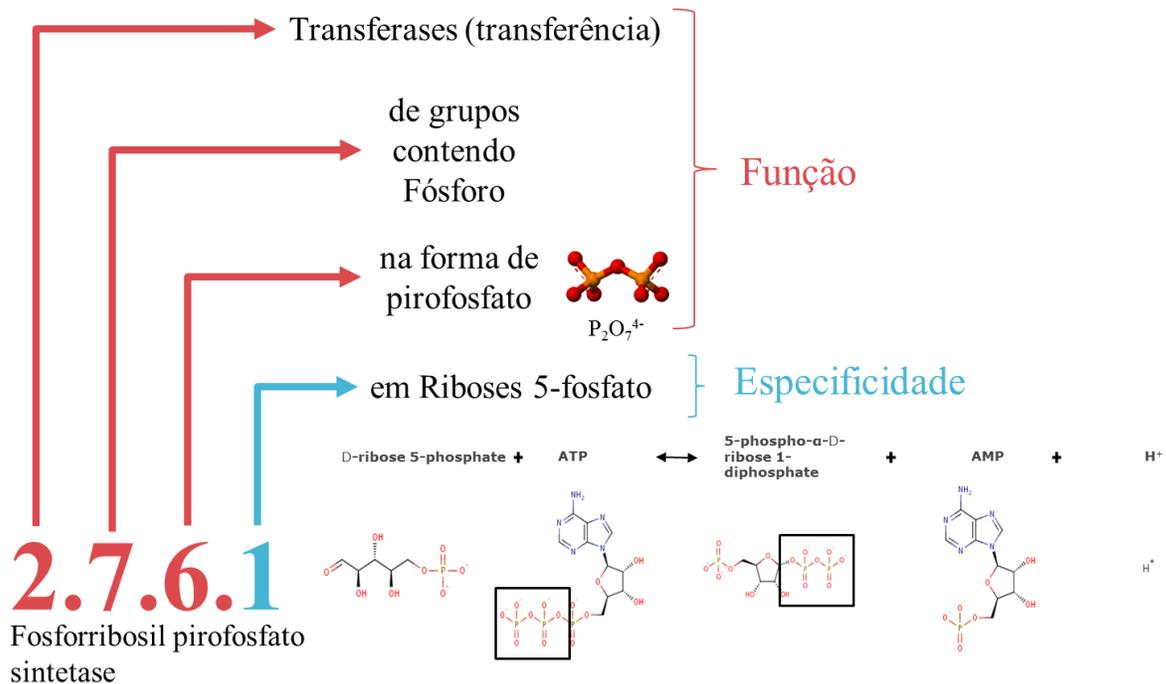


Figura 1-2: Classificação das enzimas Fosforribosil pirofosfato sintetase (EC.2.7.6.1) segundo Nomenclatura EC. Essas enzimas são responsáveis pela catálise de reações de transferência de grupos de pirofosfato entre moléculas de ATP e Ribose 5-fosfato. A função da enzima (reação genérica catalisada) é dada pelos três primeiros dígitos EC (sub-subclasse), enquanto que o quarto dígito juntamente com os outros três definem a especificidade da enzima (substrato de ligação).

Identificadores EC, portanto, não especificam enzimas, mas as reações catalisadas por elas, se diferentes enzimas (provenientes de organismos diferentes, por exemplo) são responsáveis pela catálise da mesma reação química, recebem o mesmo número EC.

Essa classificação hierárquica permite realizar uma separação das enzimas segundo reações catalisadas, possibilitando assim a diferenciação dos mecanismos catalíticos. Os primeiros três níveis da hierarquia correspondem à reação catalisada pela enzima, sendo por fim o quarto nível correspondente ao substrato da reação. Utilizando, portanto, os três primeiros níveis, é possível separar as enzimas de acordo com as reações químicas que estas catalisam, sem levar em consideração com qual substrato irão interagir, **despertando o questionamento se estas apresentam similaridades e padrões estruturais entre seus resíduos catalíticos.**

1.4 Descritores estruturais de proteínas e identificação *in silico* de resíduos de aminoácidos catalíticos

Diversos métodos têm sido propostos para obter informações sobre as propriedades físicas, químicas, físico-químicas, geométricas e espaciais dos aminoácidos (e.g. energia de contatos, potencial eletrostático, hidrofobicidade, densidade, etc.). Descritores obtidos a partir das conformações

tridimensionais das proteínas (estruturas) são denominados **descritores estruturais** de proteínas e tais propriedades permitem realizar diversas análises, segundo diferenças e semelhanças entre resíduos aminoácidos de várias regiões de uma proteína. Além de descritores estruturais, existem descritores baseados na conservação das sequências de aminoácidos das proteínas, ou seja, baseiam-se em evidências evolutivas obtidas a partir de comparações via alinhamentos sequenciais. Determinando-se quais aminoácidos são mais frequentes em certas posições da sequência, fornecem indícios da importância destes na manutenção da funcionalidade da proteína. Dessa forma, esses descritores compõem o grupo de descritores sequenciais de proteínas, ou seja, descritores obtidos a partir da sequência de aminoácidos das proteínas.

No Laboratório do Grupo de Pesquisa em Biologia Computacional (GPBC) da Embrapa Informática Agropecuária, foi desenvolvida a maior base de dados de descritores estruturais e sequenciais gratuita sobre estruturas proteicas: o STING_DB. O STING_DB é a principal fonte de informações do Blue Star STING (Neshich, et al., 2006): uma plataforma *web* que reúne uma suíte de programas com ferramentas para a visualização e análise estrutural de proteínas, disponível através do endereço <http://www.cbi.cnptia.embrapa.br/SMS>.

O grande diferencial dos descritores disponíveis no Blue Star STING deve-se ao fato destes serem obtidos seguindo uma definição do nanoambiente dos resíduos de aminoácidos. Resíduos de aminoácidos não estão isolados nas estruturas das proteínas e, por isso, realizam diversas interações com os outros resíduos em sua vizinhança. Devido a essas interações e às diferentes composições da vizinhança de cada resíduo de aminoácido, as propriedades físicas, químicas, físico-químicas e estruturais desses resíduos sofrem alterações e influência do ambiente onde estão localizados. Dessa forma, o nanoambiente de um resíduo de aminoácido é definido como sendo uma pequena região em torno de um resíduo de aminoácido, que englobe as características do resíduo de aminoácido e de seus vizinhos nessa região, capturando, assim, propriedades de um ambiente específico no qual o resíduo está inserido. No Blue Star STING esta pequena região em torno de um resíduo de aminoácido é simulada através de uma esfera imaginária colocada sobre algum átomo do resíduo de aminoácido (carbono alfa, LHA, carbono beta, etc.), denominada de **esfera sonda**. Através da esfera sonda é possível obter uma descrição do ambiente em torno do átomo central descrevendo, assim, o nanoambiente no qual um resíduo de aminoácido está inserido. Durante o cálculo dos descritores do Blue Star STING esta esfera é deslizada sobre todos os átomos, escolhidos como centrais, que compõem os resíduos de aminoácidos. Dessa forma, é possível obter uma descrição do nanoambiente de todos os resíduos de aminoácido que compõem uma proteína. É possível, ainda, utilizar diferentes tamanhos de esfera sondas, criando assim representações de nanoambientes de diversos tamanhos e características.

A partir de um conjunto de enzimas com seus resíduos de aminoácidos catalíticos devidamente identificados (e.g. experimentos de mutagênese) como estudo de caso, um modelo computacional pode ser construído usando métodos de aprendizagem de máquina, para caracterizar o nanoambiente destes resíduos segundo suas propriedades estruturais, através de descritores de proteínas extraídos do banco de dados STING_DB e, então, avaliar se tais modelos possuem capacidade para predição dos resíduos de aminoácidos catalíticos de enzimas ainda não anotadas e de novas enzimas.

O principal objetivo deste trabalho consistiu na extração e catalogação de propriedades físico-químicas e estruturais responsáveis por caracterizar as enzimas segundo as reações que estas catalisam. Para isto, foram criados diversos modelos classificadores utilizando técnicas estatísticas e de aprendizado de máquina para a obtenção de regras de forma que, ao serem aplicadas às enzimas provenientes de uma mesma sub-subclasse EC (terceiro nível na hierarquia EC), selecionem seus resíduos de aminoácidos catalíticos. A criação de um modelo classificador para cada diferente sub-subclasse EC permite a caracterização do nanoambiente necessário aos resíduos de aminoácidos, bem como um maior entendimento dos processos catalíticos, uma vez que estas regras podem ser facilmente interpretadas por especialistas, auxiliando assim estudos de certas classes de enzimas e seus mecanismos catalíticos.

No capítulo Capítulo 2, é apresentada uma revisão bibliográfica enumerando diversos métodos computacionais para predição e classificação de resíduos de aminoácidos catalíticos, estabelecendo os desempenhos das técnicas atuais e os desafios enfrentados. No capítulo Capítulo 3, são descritas as bases de dados utilizadas neste trabalho com ênfase na base de dados de descritores estruturais de proteínas. No capítulo Capítulo 4, é apresentada uma descrição detalhada dos descritores e seus diferentes tipos. Uma descrição dos métodos de aprendizado de máquina empregados neste trabalho encontra-se no capítulo 0, e a metodologia adotada para condução dos experimentos computacionais e construção dos modelos classificadores para predição de resíduos de aminoácidos catalíticos encontram-se no capítulo 0. Os resultados e discussões do emprego da metodologia adotada são apresentados no capítulo 0, seguido pela conclusão, disposta no capítulo Capítulo 8.

Capítulo 2 - Revisão Bibliográfica

Apesar deste trabalho não buscar diretamente a predição dos resíduos catalíticos, mas sim uma caracterização físico-química e estrutural dos CSR's já conhecidos, acredita-se que estas propriedades permanecerão conservadas também em enzimas não anotadas e em novas enzimas, sendo esta a premissa fundamental de qualquer método de classificação, ou seja, existência de padrões implícitos e/ou explícitos nos dados que podem ser extraídos por técnicas estatísticas ou de aprendizado de máquina, de forma que estes padrões sejam também passíveis de generalização e observáveis em outros conjuntos de dados de uma mesma população amostral. É possível que os modelos obtidos neste trabalho possam desempenhar também o papel de predição dos CSR's em novas enzimas.

2.1 Predição de resíduos de aminoácidos catalíticos a partir de sequências ou de estruturas proteicas

Diversas propostas são encontradas na literatura para a predição de resíduos de aminoácido catalíticos. Podem-se dividir estas em duas categorias: abordagens baseadas em sequência de aminoácidos que formam as enzimas e abordagens baseadas em estrutura. Enquanto métodos baseados em sequência permitem a predição de resíduos de aminoácidos catalíticos utilizando somente informações provenientes das estruturas primárias das enzimas, métodos baseados em estrutura utilizam informações adicionais extraídas da estrutura terciária da molécula. Este pode ser um ponto vantajoso para o processo de anotação funcional de uma proteína pois devido à sua complexidade, grande parte das sequências proteicas dos maiores bancos internacionais, como GenBank NR (Benson, et al., 2014), TrEMBL (Bairoch & Apweiler, 2000) e KEGG (Kanehisa, et al., 2004), possuem grandes erros de anotação funcional. Segundo o artigo publicado por SCHNOES et al. (2009), em 10 das 37 famílias enzimáticas estudadas (que possuem extensiva documentação funcional) o nível de anotações incorretas passa dos 80% das sequências destas famílias, com exceção dos bancos manualmente curados, como o SwissProt (Bairoch, et al., 2004). Eles discutem, ainda, que o nível de sequências erroneamente anotadas cresceu vertiginosamente de 1993 a 2005, principalmente com o advento das tecnologias de sequenciamento de nova geração (NGS) que ocasionou a propagação de anotações errôneas.

Na maioria dos trabalhos, leva-se em consideração a homologia entre sequências e propriedades de conservação de resíduos, bem como a propensão dos resíduos para distinguir entre resíduos de aminoácidos catalíticos e não catalíticos. A propensão ou propensidade (*propensity* em língua inglesa) é definida como sendo a razão entre o número de resíduos aminoácidos catalíticos para cada um dos 20 tipos de aminoácidos e o total de aminoácidos de cada tipo. Esta medida fornece informações de quão

propenso (probabilidade a priori) cada tipo de aminoácido é de ser catalítico. Apesar de descritores de conservação e propensão fornecerem informações relevantes para a discriminação entre os aminoácidos, neste trabalho optou-se por não utilizá-los, ao menos nos modelos finais a serem obtidos, sendo estes utilizados somente para efeito de comparação. Isto porque tais descritores não fornecem nenhum conhecimento sobre as propriedades estruturais do ambiente no qual os CSR's estão inseridos. Além do mais, inferências evolutivas nem sempre são válidas, pois proteínas podem ter diferentes funções mesmo sendo evolutivamente relacionadas (Todd, et al., 2001) e resíduos podem ser conservados devido a outras razões (e.g. estabilidade estrutural).

Resíduos de aminoácidos mostram possuir diferentes propensões para serem catalíticos. Por exemplo, resíduos de aminoácidos carregados (H, R, K, E, D) correspondem a 65% de todos os CSR's, enquanto que 27% são polares (Q, T, S, N, C, Y, W) e os outros 8% são hidrofóbicos (G, F, L, M, A, I, P, V) (Barlett, et al., 2002). Na Figura 2-1, é possível visualizar a propensão catalítica de cada resíduo, calculada a partir do banco de dados construído para uso neste trabalho. Similar ao trabalho de BARLETT et al. (2002), no banco de dados utilizado os CSR's estão distribuídos em 61% carregados, 29% polares e 10% hidrofóbicos.

Evidentemente, descritores de conservação proporcionam uma das mais poderosas características para predição de resíduos de aminoácidos catalíticos como tem sido mostrado por diversos trabalhos encontrados na literatura: PETROVA & WU (2006), CHIEN et al. (2008), GUTTERRIDGE et al. (2003), YOUN et al. (2007), LA et al. (2005), DUKKA BAHADUR & LIVESAY (2008). Adicionalmente, foi encontrado que características coevolutivas podem ser frequentemente derivadas da vizinhança de importantes sítios funcionais (Lee, et al., 2008), sendo que esta informação pode ser também utilizada para a predição de resíduos de aminoácidos catalíticos.

Características ligadas à conservação e propensão dos resíduos de aminoácidos, além de não fornecerem uma explicação para a razão pela qual tais resíduos são importantes para a funcionalidade das enzimas, possuem limitações como a necessidade de uma molécula de referência para que seja possível obter informações de similaridade e conservação da estrutura primária. Muitos trabalhos têm mostrado que, mesmo enzimas compartilhando similaridades sequenciais superiores a 50%, menos de 30% delas desempenham a mesma função (mesmo número EC). Em uma análise de 167 superfamílias homólogas retiradas de uma base de dados que fornece classificação e agrupamento de estruturas proteicas segundo diversos aspectos estruturais e sequenciais (CATH - versão 9.0) (Sillitoe, et al., 2013), quase metade possuía enzimas com diferentes números EC. Ainda que muitas destas enzimas diferenciem apenas no último nível hierárquico (especificidade ao substrato), 22 superfamílias continham membros com números EC não conservados em nenhum nível, além de diferentes atividades enzimáticas (Todd,

et al., 2001). Há ainda trabalhos que mostram casos em que a mesma função catalítica evolui independentemente (evolução convergente) (George, et al., 2004; Hegyi & Gerstein, 1999).

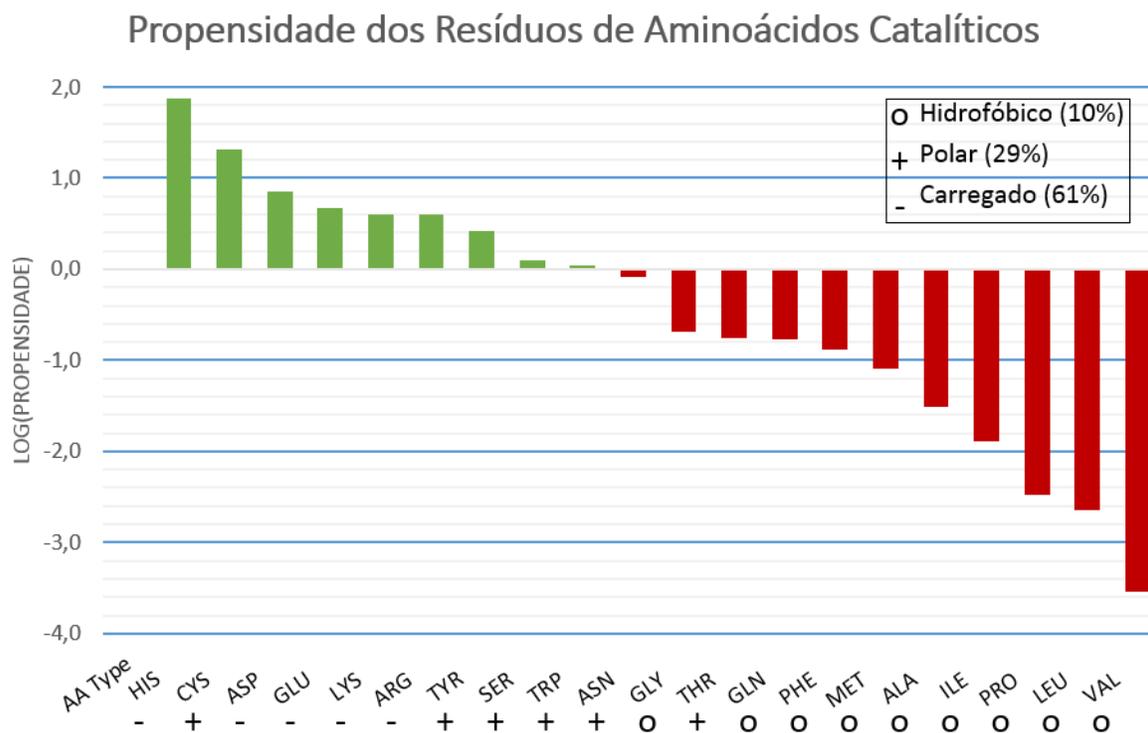


Figura 2-1: Propensidades dos resíduos de aminoácidos catalíticos obtidas a partir do banco de dados empregado neste trabalho. Aminoácidos que apresentam valores maiores do que zero podem ser considerados mais propensos para serem catalíticos do que aqueles com valores inferiores ou iguais a zero.

CHOTIA & LESK (1986) demonstraram ainda que a função das proteínas é mais conservada na estrutura terciária do que na estrutura primária. Há várias estruturas com nenhuma ou pouca similaridade de sequência que possuem similaridade estrutural, confirmando assim mais uma evidência de que a função de uma proteína está ligada à sua conformação tridimensional.

2.2 Métodos baseados em propriedades estruturais das proteínas

Métodos para predição dos CSR's exploram algumas propriedades notórias dos resíduos de aminoácidos catalíticos. Por exemplo, foi demonstrado que os resíduos catalíticos localizam-se preferencialmente no centro geométrico das estruturas proteicas (Ben-Shimon & Eisenstein, 2005) e que estes geralmente estão localizados em largas cavidades na superfície das moléculas (Barlett, et al., 2002; Ota, et al., 2003). Além disso, como muitos atuam como aceitadores ou doadores de hidrogênio no processo catalítico, as ligações de hidrogênio podem também ser utilizadas para discriminar resíduos de aminoácidos catalíticos e não catalíticos (Petrova & Wu, 2006; Tang, et al., 2008). Outras propriedades

estruturais, tais como acessibilidade ao solvente, flexibilidade de alças (Malabanan, et al., 2010) e fatores-B também têm sido utilizadas como características para a predição de resíduos catalíticos (Yuan, et al., 2003). Propriedades eletrostáticas também têm se mostrado úteis na predição de resíduos catalíticos, como características baseadas em propriedades eletrostáticas derivadas de curvas teóricas de titulação (Ondrechen, 2001; Ko, et al., 2005) e energia eletrostática dos resíduos (Elcock, 2001).

Um dos trabalhos pioneiros em predição de resíduos catalíticos utiliza uma combinação de alinhamentos múltiplos de sequências e vizinhos estruturais para calcular, além da média local da conservação ao longo da sequência, a média espacial sobre a estrutura 3D (Zvelebil & Sternberg, 1988). O Blue Star STING possui um descritor semelhante denominado Entropia Relativa 3D (apresentado na Seção 4.2.2).

Em LANDGRAF et al. (2001), agrupamentos (*clusters*) tridimensionais são utilizados para a predição de resíduos funcionais através da conservação de resíduos espacialmente adjacentes, obtendo altas taxas de predição (83%) com uma baixa taxa de erro (2%) para 15 enzimas. OTA et al. (2003) utilizaram conservação de sequência, conservação local, análise de estabilidade e localização geométrica dos resíduos para identificar 56% dos resíduos catalíticos de 49 enzimas.

Utilizando uma rede neural artificial juntamente com agrupamento espacial, considerando somente resíduos altamente conservados (D, E, K, R, H, S, T, N, Y e C), o método proposto por GUTTERIDGE et al. (2003) identificou 69% dos resíduos catalíticos com uma alta taxa de falsos positivos para 189 enzimas da base de dados CATRES (Barlett, et al., 2002), contendo anotações de resíduos catalíticos manualmente curadas. O método utiliza conservação de sequência, tipo do resíduo, e quatro parâmetros estruturais (profundidade, área relativa acessível ao solvente, localização do resíduo em uma cavidade da molécula e a área superficial da cavidade) como entradas da rede neural artificial. Após o treinamento da rede neural artificial, os resíduos identificados como catalíticos são ordenados de acordo com a saída da rede e, então, um algoritmo de clusterização é utilizado para encontrar agrupamentos entre resíduos de uma mesma enzima. Devido aos resíduos catalíticos serem encontrados em grupos, e não independentemente como é suposto durante o treinamento da rede neural, esta etapa tem por finalidade encontrar sítios ativos e filtrar falsos positivos.

Uma abordagem utilizando grafos denominada SARIG, foi realizada para identificação de resíduos funcionais em proteínas (Amitai, et al., 2004). Neste método, proteínas foram representadas como grafos de interações entre resíduos de aminoácidos, onde os aminoácidos representam nós no grafo e as interações entre eles representam arestas. Utilizando métricas de centralidade de proximidade (*closeness centrality*) (Noh & Rieger, 2004) e cardinalidade (número de vizinhos) dos nós no grafo, bem como descritores de área relativa acessível ao solvente, foi possível identificar que resíduos catalíticos, sítios

de ligação e resíduos evolutivamente conservados possuem altos valores para centralidade de proximidade nos grafos estudados, e a aplicação do método na predição de sítios ativos resultou em uma sensibilidade¹ média de 46.5% e precisão² média de 9.4%. No entanto, no caso de identificação de sítios alostéricos e de ligação de íons, cofatores e substratos, este se mostrou mais eficaz.

Apesar dos métodos pioneiros realizarem predição de um número significativo de resíduos funcionais (sensibilidade), geralmente estas predições não se mostraram seletivas, ou seja, englobam grandes frações dos resíduos das proteínas e com um número muito maior de falsos positivos (FP's) do que verdadeiros positivos (TP's). Para terem uma máxima utilidade, um classificador deve ser capaz de identificar os resíduos funcionais com sensibilidade e seletividade; neste sentido, os métodos recentemente publicados têm focado bastante na redução da taxa de falsos positivos.

O método ResBoost (Alterovitz, et al., 2009) baseia-se em regras experimentais notadamente conhecidas e amplamente empregadas por pesquisadores para selecionar resíduos candidatos à mutagênese. Tais regras baseiam-se na hipótese de que resíduos catalíticos são evolutivamente conservados, acessíveis ao solvente, encontrados em grupos, tipicamente hidrofílicos e localizados no interior de cavidades na superfície da molécula. No entanto, essas regras isoladamente não são capazes de explicar a variedade de resíduos catalíticos produzidos por bilhões de anos de evolução. Dessa forma, o método ResBoost combina os métodos *Adaptive Boosting* (Freund & Schapire, 1997) e *Alternating Decision Trees* (ADTree) (Freud & Mason, 1999). O meta-algoritmo *Adaptive Boosting* (AdaBoost) é utilizado em conjunto de outros algoritmos de aprendizado de máquina, denominados classificadores-base, a fim de aumentar o desempenho através da combinação ponderada de suas predições (*ensemble*). Iterativamente, o algoritmo ajusta os pesos das amostras erroneamente classificadas em iterações anteriores para as classificações subsequentes, de modo que os próximos classificadores-base treinados passem a dar maior relevância à predição de amostras com maior peso. O método ResBoost utiliza as regras acima como classificadores-base, e utiliza o AdaBoost para ajustar iterativamente os pesos das regras (classificadores-base) e adicioná-las como folhas na árvore gerada pelo algoritmo ADTree durante a fase de treinamento, de forma a minimizar o número de predições incorretas. Os autores reportaram taxas de sensibilidade de 55% e 85%, e precisão de 17% e 7.1%, respectivamente utilizando duas diferentes configurações do algoritmo.

O algoritmo POOL (Tong, et al., 2009) emprega o método batizado de *monotonicity-constrained maximum likelihood* para atribuir probabilidades para cada resíduo de forma a indicar sua importância

¹ Proporção de amostras positivas corretamente identificadas como positivas (cobertura).

² Proporção de amostras positivas dentre todas as amostras identificadas como positivas.

funcional. Baseando-se na suposição de monotonicidade das características utilizadas, o método realiza uma normalização não-linear destas características e encontra um estimador de máxima verossimilhança. Originalmente, o método utilizou apenas características presentes no THEMATICS (Ondrechen, 2001), que são aplicáveis somente a resíduos ionizáveis (R, D, C-H, E, H, K, Y e os N- e C- terminais). Porém extensões posteriores passaram a incorporar outros resíduos através de variáveis do ambiente (propriedades de resíduos espacialmente próximos), e mais recentemente (Somarowthu, et al., 2011) incorporam informações evolutivas filogenéticas utilizando INTREPID (Sankararaman & Sjölander, 2008); cavidades e *pockets* utilizando ConCavity (Capra, et al., 2009); e uma nova característica disponível pelo THEMATICS, *theoretical buffer range* (BR). Tais características aumentaram a predição de resíduos catalíticos, obtendo taxas de sensibilidade em torno de 80% e precisão em torno de 14%.

YAHALOM et al. (2011) propuseram um método de classificação de resíduos catalíticos que utiliza somente propriedades estruturais de proteínas, excluindo-se característica de conservação de sequência. Utilizaram Máquinas de Vetores de Suporte (SVM) (Cortes & Vapnik, 1995) para treinar classificadores utilizando diferentes conjuntos de características de forma a maximizarem o coeficiente de correlação de Matthews (MCC). Os autores reportaram taxas de sensibilidade de 61.9% (± 0.035) e precisão de 17.2% (± 0.008).

Atualmente, o método de predição de resíduos catalíticos proposto por CHIEN & HUANG (2012) talvez apresente-se como o de maior sucesso. Conhecido como EXIA (*Enzyme Catalytic residue Side-chain Arrangement*), o método baseia-se principalmente na verificação de que as cadeias laterais dos resíduos de aminoácidos catalíticos possuem orientações e arranjos específicos. Utilizando medidas dos ângulos formado pelos vetores que ligam o carbono alfa ao átomo funcional de cada resíduo de aminoácido, e vetores entre o carbono alfa a um ponto que pode representar o centro catalítico (*catalytic spot*), é possível discriminar os resíduos catalíticos dos demais resíduos. O fato das cadeias laterais dos resíduos de aminoácidos catalíticos possuírem uma certa orientação necessária para expor seus átomos funcionais, de forma a estabelecer diversas interações químicas com o substrato, foi muito bem explorado pelo método EXIA. O método ainda utiliza propensidade e um descritor de flexibilidade, juntamente com ângulo entre estes vetores, para atribuir uma pontuação a cada resíduo de aminoácido da proteína, sendo que os resíduos com maiores pontuações possuem maiores chances de serem catalíticos. O método atinge sensibilidade 71% para precisão de 18% e precisão de 22% para uma sensibilidade de 61.7%. Utilizando-se ainda informações de conservação de estrutura primária, é possível elevar a sensibilidade para 80% e precisão para 24% (Chien & Huang, 2012). Ainda que a sensibilidade seja considerada alta, tal precisão indica um elevado número de falsos positivos, ou seja, resíduos de aminoácidos não catalíticos preditos como catalíticos.

A utilização de informações provenientes da vizinhança dos resíduos catalíticos foi explorada por HAN et al. (2012). Os autores desenvolveram um novo descritor que integra informações do nanoambiente e propriedades geométricas dos resíduos de aminoácidos, chamado de MEDscore. Utilizando este descritor, foi possível perceber diferentes propensões dos resíduos de aminoácidos nos microambientes dos resíduos de aminoácidos catalíticos. Alguns resíduos (C, M, H, S, T, W, Y, F, G) localizam-se preferencialmente próximos aos resíduos de aminoácidos catalíticos, enquanto que outros (E, K, R, D, A, L, P, V, Q) estão raramente localizados em suas vizinhanças. O descritor MEDscore identificou corretamente 70% dos resíduos catalíticos com uma taxa de falsos positivos (FPR) inferior a 10%.

Os mesmos autores do método EXIA propuseram dois descritores para quantificar a rigidez estrutural local dos resíduos de aminoácidos: flexibilidade sequencial local (SEQ) e flexibilidade estrutural local (STR), ambas baseadas nas distâncias entre os resíduos de aminoácidos, considerando as vizinhanças sequenciais e estruturais, respectivamente. Utilizaram ainda uma SVM para identificação dos CSR's (Chien & Huang, 2013). Devido à necessidade de uma correta orientação dos CSR's nas moléculas para que a catálise ocorra, resíduos de aminoácidos catalíticos mostraram possuir uma maior rigidez estrutural (90% de CSR's em comparação com 14% de não CSR's), fato similarmente explorado pelo método EXIA. O método foi comparado com outros métodos, mostrando desempenhos similares, porém com maior simplicidade e custo computacional.

O método proposto por CILIA & PASSERINI (2010) utiliza três descritores baseados nas estruturas primárias das enzimas e nove baseados nas vizinhanças dos resíduos de aminoácidos das estruturas terciárias. Considerando esferas de raios fixos de 8Å centradas no centroide da cadeia lateral dos resíduos de aminoácidos, descritores codificando propriedades físicas e químicas, propensões, polaridade, número de moléculas de água, densidade atômica, e fatores-B nas vizinhanças dos resíduos de aminoácido são utilizados juntamente com conservação dos resíduos nas estruturas primárias das enzimas para treinar uma SVM, utilizando uma adaptação de uma função de *kernelização* utilizada para caracterização de pequenas moléculas (Ceroni, et al., 2007). O método mostrou ser bastante seletivo com sensibilidades acima de 60% e precisão sempre superior a 20%.

2.3 Métodos baseados em propriedades extraídas das sequências proteicas

O método *Evolutionary Trace* (ET) (Lichtarge, et al., 1996), apesar de não ser um método para este fim específico, pode ser aplicado para a predição de sítios ativos e interfaces em proteínas com estrutura conhecida. Baseado na observação de que resíduos funcionais são mais conservados nas sequências proteicas do que outros resíduos, o método encontra os resíduos mais conservados para diferentes níveis

de identidade de sequência. A partir de árvores filogenéticas e alinhamentos múltiplos considerando diferentes níveis de similaridade sequencial, o método particiona o conjunto de proteínas agrupando em cada nível da árvore sequências com alguma identidade, sendo na raiz encontradas todas as sequências e, gradativamente, a similaridade mínima é incrementada até as folhas, onde têm-se sequências com alto grau de homologia. O método então calcula os consensos dos alinhamentos múltiplos em cada nível da árvore para cada grupo de sequências. Estas sequências de consensos são então alinhadas e os resíduos nas diferentes posições da sequência são classificados como conservados, caso apresentem-se como invariantes no alinhamento; ou classe-específicos (*class-specific*) caso o tipo do resíduo apresente variação no alinhamento entre sequências de consenso. Considerando-se os diferentes níveis da árvore (identidade sequencial) o método consegue obter os resíduos mais conservados em todo um conjunto de proteínas e também aqueles conservados intragrupos (*classe-específicos*).

Enquanto o método ET tem mostrado sucesso em muitos estudos de caso (Chakravarty, et al., 2005; Innis, et al., 2000; Zhu, et al., 2004), a necessidade de uma inspeção manual na versão original do algoritmo não é aplicável para a automatização em larga escala. Modificações e versões automatizadas do método foram propostas e testadas para dois conjuntos de dados. Em um dos estudos (Aloy, et al., 2001), os resíduos catalíticos foram corretamente preditos para 62 de 80 (77.5%) enzimas. Em outro estudo (Yao, et al., 2003) aproximadamente 60% dos resíduos catalíticos foram corretamente preditos para 29 enzimas experimentalmente analisadas.

Dentre os métodos que utilizam somente informações da estrutura primária e desenvolvidos especificamente para predição de CSR's, destacam-se os métodos CRpred (Zhang, et al., 2008) e L1pred (Dou, et al., 2012). Apesar de não incorporarem informações estruturais, tais métodos atingem taxas de desempenho similares às de muitos métodos baseados em estruturas, com taxas de sensibilidade em torno de 40% e precisão de 18%. Isto deve-se ao fato de que, mesmo em métodos baseados em estruturas, as propriedades de maior poder discriminativo entre CSR's e o restante dos resíduos são as propriedades derivadas da comparação entre sequências (conservação). Mesmo que utilizem informações estruturais como tentativa de melhorarem as predições, poucos descritores possuem capacidades preditivas similares aos descritores de conservação. Por outro lado, estes métodos, por não necessitarem da estrutura terciária das enzimas, podem ser empregados em enzimas com estruturas ainda não resolvidas. No entanto, a existência de sequências homólogas é essencial e necessária para realizar as predições.

Métodos baseados em transferência de anotação usando somente similaridade de sequência geralmente necessitam de um alto grau de similaridade. Para se transferir todos os quatro dígitos do número EC com uma taxa de erro de até 10%, são necessários pelo menos 60% de similaridade sequencial (Tian & Skolnick, 2003), e somente por volta de 60% das proteínas puderam ser anotadas por homologia

com proteínas contendo informações funcionais experimentalmente identificadas para 62 proteomas (Rost, et al., 2003).

Os métodos baseados em estrutura, apesar de trazerem ganhos relativos à sensibilidade em comparação com métodos baseados em sequência, configurando uma maior cobertura dos resíduos de aminoácidos catalíticos, ainda apresentam baixas taxas de precisão, assim como métodos baseados em sequência, resultando um alto número de resíduos preditos erroneamente como catalíticos. Isto dificulta a aplicação prática destes métodos, uma vez que ainda assim são necessários muitos ensaios experimentais para a identificação correta dos resíduos de aminoácidos catalíticos e para a eliminação dos falsos positivos.

Apesar de vários métodos já terem sido propostos, uma comparação direta entre tais métodos é bastante difícil, dadas as diferentes medidas de desempenho utilizadas e principalmente os diferentes conjuntos de dados com diferentes tamanhos e qualidades. Métodos mais recentes têm utilizado estruturas escolhidas a partir do *Catalytic Site Atlas* (CSA) (Furnham, et al., 2014) e do banco de dados CATRES (Barlett, et al., 2002), o que permite uma melhor comparação entre métodos. Porém, ainda é difícil saber qual dos métodos é o mais eficaz, pois a maioria dos métodos utiliza apenas um subconjunto das amostras. No entanto, a acurácia da predição de resíduos catalíticos permanece por volta de 70%, com altas taxas de falsos positivos (Petrova & Wu, 2006).

2.4 Separação de enzimas segundo reações catalisadas (número EC) e desafios à predição dos CSR's

Alguns trabalhos na literatura buscam, ao invés da predição de resíduos catalíticos para obterem informações funcionais das enzimas, classificarem proteínas inteiras segundo suas classes enzimáticas e outros níveis da hierarquia EC. Recentemente, VOLPATO et al. (2013) utilizaram uma rede neural artificial construída usando uma nova arquitetura desenvolvida para estes propósitos, denominada *N-to-1 Artificial Neural Network*. Utilizando somente informações de sequência para classificar as enzimas dentre as seis diferentes classes enzimáticas (números EC. de 1 a 6), o método mostrou-se bastante eficaz com taxas de sensibilidade de 80% e taxa de falsos positivos sempre menores que 7% para todas as seis classes. VOLKAMER et al. (2013) utilizaram um esquema de cascata para a predição dos quatro níveis da hierarquia EC. Utilizando diversas SVM's, uma para cada nível da hierarquia EC de acordo com a disponibilidade de estruturas em cada nível, foi possível identificar o primeiro nível EC com acurácia variando de 37.8% para EC.6 até 75.3% para EC.1. Para as subclasses, a acurácia das predições varia de 62.8% para 15 subclasses do número EC.1 até 80.9% para 6 subclasses do número EC.4.

Realizar uma separação das enzimas segundo suas classes, sub-classes e sub-subclasses enzimáticas,

e criar modelos classificadores específicos para cada uma delas pode trazer ganhos em relação à utilização de um modelo geral para predição de resíduos de aminoácidos catalíticos. No entanto, a predição dos CSR's envolve diversos desafios, como o alto desbalanceamento entre as classes de resíduos (catalíticos e não catalíticos), onde cerca de menos de 1% dos resíduos são anotados como catalíticos. Além disto, a dificuldade em se obter uma definição consistente para resíduos de aminoácidos catalíticos e a caracterização dos resíduos como grupos funcionais e não como resíduos isolados vêm também sendo apresentadas como desafios para métodos existentes. Unindo-se o fato de que as bases de dados sobre resíduos catalíticos disponíveis não são completas, ou seja, contêm um número menor de CSR's do que realmente existem, devido à ausência de experimentos que comprovem propriedades catalíticas, os métodos classificadores veem-se prejudicados, devido ao uso de rótulos incorretos para ajuste dos modelos. Não é raro defrontar-se com trabalhos indicando propriedades catalíticas de resíduos antes tomados como não catalíticos e ainda não incorporados às bases de dados (Schreier & Höcker, 2010; Wedekind, et al., 1995; Jordan, et al., 1999). É interessante notar que as informações contidas nestas bases de dados não informam se um resíduo não é catalítico, ou seja, se não existe informação para um resíduo nestas bases, isto não significa que tal resíduo deva ser tomado como não catalítico. No entanto, em métodos de aprendizado de máquina supervisionados é necessário a existência de rótulos para cada amostra utilizada no aprendizado. Por esta razão, tomam-se os resíduos não contidos nas bases de dados como não catalíticos, mas na maioria das vezes estes rótulos estão incompletos ou incorretos. Por essas razões, as taxas de falsos positivos apresentadas pelos métodos classificadores, apesar de estarem em conformidade com os conjuntos de dados utilizados, não necessariamente representam falsas predições.

O desbalanceamento entre as classes apresenta-se como um grande desafio para as técnicas atuais de aprendizado de máquina, sendo necessário o emprego de técnicas de pré-processamento como subamostragem (*undersampling*) ou sobreamostragem (*oversampling*). A maioria dos métodos proposto para predição de resíduos catalíticos realizam a subamostragem dos resíduos de aminoácidos não catalíticos de forma a reduzirem o desbalanceamento (Petrova & Wu, 2006; Tang, et al., 2008; Chien & Huang, 2013). No entanto, nenhum utiliza a sobreamostragem como forma de aliviar o desbalanceamento.

Outro desafio apresenta-se na caracterização dos resíduos de aminoácidos catalíticos como grupos funcionais. Geralmente, enzimas possuem um número baixo de resíduos catalíticos (entre 1 e 10, sendo em média de 3 a 4 resíduos). Dessa forma, quando se busca uma caracterização dos resíduos catalíticos, classificação através de regras, por exemplo, seria interessante obter explicações sobre as propriedades catalíticas do grupo de resíduos e não somente de um resíduo. Obviamente, resíduos catalíticos desempenham papéis diferentes durante a catálise e, portanto, devem apresentar características físico-

químicas e estruturais distintas de modo a complementarem seus papéis. No entanto, alguma característica presente nos nanoambientes destes resíduos deve existir ao ponto de tais resíduos serem importantes para a funcionalidade da enzima.

Assim, buscou-se obter uma caracterização do “por quê” tais resíduos apresentam-se como catalíticos por meio de propriedades estruturais compartilhadas entre os grupos de resíduos. Em alguns métodos classificadores, esta busca vê-se limitada devido às características do próprio método, como, por exemplo, em Redes Neurais Artificiais, onde a partir de uma rede treinada é difícil ou até mesmo impossível obter uma caracterização interpretável dos grupos de resíduos. No entanto, regras facilitam a interpretação e podem fornecer uma caracterização de grupos de resíduos quando estes são corretamente considerados.

Capítulo 3 - Conjunto de dados

Todas as informações em relação à posição tridimensional dos átomos das estruturas estudadas foram retiradas de banco de dados públicos. O *Protein Data Bank* (PDB) (Berman, et al., 2000) é um repositório de todas as estruturas tridimensionais de proteínas que foram resolvidas experimentalmente e, portanto, é uma das fontes mais importantes para a realização desse trabalho, uma vez que é a partir das informações obtidas do PDB que os descritores estruturais de proteínas são calculados.

3.1 Base de dados de resíduos catalíticos – Catalytic Site Atlas

O banco de dados CSA (da sigla em inglês para *Catalytic Site Atlas*) (Furnham, et al., 2014; Porter, et al., 2004) foi utilizado para identificar os resíduos de aminoácidos catalíticos das enzimas provenientes do PDB. O CSA contém dois tipos de entradas: entradas manualmente anotadas, derivadas da literatura através de estudos experimentais; e entradas obtidas a partir da transferência de anotação por homologia com as primeiras, encontradas pelo *software* PSI-BLAST (*Position-Specific Iterative Basic Local Alignment Search Tool*; (Altschul, et al., 1997)). Como em outros trabalhos, considerou-se somente entradas no CSA obtidas a partir de evidências experimentais, excluídas aquelas obtidas por homologia, devido à propagação e introdução de erros que podem reduzir a confiabilidade dos dados.

As anotações provenientes de evidências experimentais totalizam 6262 entradas de resíduos catalíticos identificados para 928 estruturas do PDB e, apesar de ser o maior banco de dados sobre resíduos catalíticos disponível gratuitamente, obviamente não é completo, no sentido de que não possui identificação de todos os resíduos catalíticos para as enzimas presentes no PDB. Como discutido no capítulo Capítulo 1, devido à dificuldade em identificar os resíduos catalíticos de uma proteína, principalmente sem auxílio de métodos computacionais para guiar experimentos de mutagênese, muitas vezes nem todos os resíduos catalíticos são avaliados e identificados. No entanto, o CSA apresenta-se como um dos únicos bancos de dados que reúne informações sobre resíduos catalíticos, sendo utilizado pela grande maioria dos trabalhos que buscam o estudo desses resíduos. Ressaltando-se que, apesar das taxas obtidas pelos métodos classificadores que utilizam esta base de dados serem consideradas virtuais, elas fornecem uma forma de avaliar e comparar os desempenhos, sendo possível utilizar um estudo de caso para averiguar melhor o poder preditivo dos modelos construídos.

Ainda que as anotações no CSA originadas a partir de evidências experimentais forneçam uma maior qualidade, comparada à qualidade das anotações derivadas da transferência por homologia, estas também estão susceptíveis a erros e incoerências. Por exemplo, a enzima *Ubiquinol oxidase* (EC. 1.10.3.10) de *Escherichia coli* (código PDB: 1FFT cadeia A) apresenta no CSA anotações para 23 resíduos. Esta grande

quantidade de resíduos de aminoácidos identificados como catalíticos é bastante anormal, dado que a média de CSR's em toda a base de dados é em torno de 3 a 4 resíduos, sendo que de um total de 1906 cadeias enzimáticas 78% apresentam quatro ou menos resíduos identificados como catalíticos, e apenas 13 cadeias (0.68%) apresentam mais de 10 CSR's. Uma análise no banco de dados de sequências proteicas Uniprot (The UniProt Consortium, 2014) mostra anotações sobre experimentos de mutagênese realizados em alguns aminoácidos da enzima (1FFT), sendo que dos 23 resíduos identificados pelo CSA como catalíticos apenas 9 aparecem no Uniprot como causadores da inativação da enzima. O restante dos resíduos não apresenta correspondência nos dois bancos.

Existem ainda no CSA anotações para 78 cadeias com mutações nos resíduos de aminoácidos catalíticos. Após experimentos de mutagênese serem realizados considerando a substituição de alguns resíduos de aminoácidos, as estruturas mutantes, assim como as estruturas nativas (*wild type structures*), são depositadas no PDB. O CSA então utiliza a estrutura mutante para rotular os resíduos de aminoácidos identificados como catalíticos na estrutura nativa, o que introduz incoerências nas anotações. A Figura 3-1 ilustra o caso de anotações de mutantes para a enzima de código PDB 1DKI cadeia A, onde no CSA o resíduo de aminoácido SER47 aparece como catalítico. No entanto, este somente é encontrado na estrutura mutante e o resíduo correto seria a CYS47, como apresentado na própria descrição da entrada.

Ainda que tais anotações possam ser facilmente detectadas e corrigidas, ferramentas computacionais que não consideram estas incoerências podem utilizar informações incorretas e fornecer resultados imprecisos. Todos os casos semelhantes foram detectados e as respectivas anotações foram corrigidas.

Residue	Chain	Number	UniProtKB Number	Functional Part	Function	Target	Description
Ser	A	47	192	macie:sideChain			Cys 47 (appears as Ser 47 in pdb file) acts as nucleophile for initial attack on the peptide substrate to form an acyl enzyme intermediate. Also stabilises the tetrahedral intermediate via oxyanion hole.

Figura 3-1: Anotação de resíduo de aminoácido catalítico no CSA de enzima mutante (PDB 1DKI cadeia A). O resíduo de aminoácido CYS47 foi experimentalmente identificado como catalítico. No entanto na enzima 1DKI (inativa) encontra-se um aminoácido do tipo Serina (SER47) na mesma posição.

3.2 Base de dados de números EC - PDBSprotEC

Para obter anotações sobre números EC das cadeias proteicas contidas no CSA, utilizou-se o banco de dados PDBSprotEC (Martin, 2004; Martin, 2005). O PDBSprotEC fornece um mapeamento entre enzimas do PDB e o banco de dados sobre informações relacionadas à nomenclatura das enzimas (números EC) ENZYME (Bairoch, 2000). Por sua vez, o banco de dados ENZYME fornece anotações sobre números EC de sequências presentes no Swiss-Prot (banco de dados de sequências proteicas (Bairoch, et al., 2004; Bairoch & Apweiler, 2000)). Dessa forma, um mapeamento entre estruturas do

PDB e seqüências no Swiss-Prot é utilizado pelo PDBSprotEC para transferir anotações do ENZYME para as enzimas do PDB. Dentre todos os mapeamentos disponíveis no PDBSprotEC, utilizou-se somente aqueles em que os quatro níveis EC foram devidamente identificados e ainda somente entradas com um único número EC.

3.3 Base de dados catalítica separada por números EC

Uma vez que as cadeias tiveram seus resíduos de aminoácidos catalíticos bem como seus respectivos números EC identificados, foi removida a redundância sequencial entre essas cadeias. Devido a muitas cadeias enzimáticas presentes no CSA apresentarem alta similaridade sequencial, sendo a grande maioria devido a homo-oligômeros (complexos proteicos formados por vários monômeros ou cadeias idênticas), faz-se necessário remover estruturas muito similares a fim de não introduzir vieses durante as fases de treinamento e teste dos classificadores. Para isto, utilizou-se um valor de 95% de similaridade sequencial máxima limitando as similaridades entre quaisquer duas cadeias da base de dados para este valor máximo. Num primeiro momento, este número pode parecer alto. No entanto, o objetivo da remoção de redundância foi de obter um conjunto de enzimas livres de homo-oligômeros e cadeias mutantes, sendo que uma análise do número de cadeias não redundantes considerando diferentes valores de similaridade sequencial mostrou que há uma drástica redução do número de cadeias à medida que se considera a remoção de cadeias com 100% de similaridade (homo-oligômeros). Porém, esta redução não apresenta o mesmo comportamento quando diminui-se ainda mais a similaridade, como apresentado no histograma da Figura 3-2. Portanto, um conjunto contendo somente enzimas com similaridade sequencial máxima de 95% não é tão redundante quanto um conjunto contendo enzimas com similaridade máxima de 40%, havendo uma diferença de apenas 40 cadeias (4%).

Mesmo sendo uma etapa preliminar, a escolha do conjunto de estruturas proteicas é uma etapa fundamental para a realização de qualquer trabalho. As etapas subsequentes de análise e posterior discussão dos resultados dependem criticamente e podem ser facilitadas devido a uma boa escolha do espaço amostral. Atualmente, existem 105,097 (13 de Dezembro de 2014) estruturas no PDB, sendo que 57,290 destas entradas possuem pelo menos identificação do primeiro nível da nomenclatura EC (classe enzimática) envolvendo 47,645 arquivos PDB distintos (de Beer, et al., 2014), ou seja, pouco mais de 54% das estruturas depositadas no PDB são enzimas. No entanto, removendo-se a redundância sequencial em 30% apenas 23,801 estruturas são retornadas pelo PDB (utilizando-se *Advanced Search* com o critério *Macromolecule Type Contains Protein* igual a “Yes” e “Retrieve only representatives at 30% sequence identity”). Destas 2,629 estão anotadas como “*unknown function*”, ou seja, é possível que até 11% das estruturas não redundantes no PDB não possuam anotações funcionais. Ainda que mais da

metade das estruturas no PDB são consideradas enzimas, a distribuição das estruturas em relação às classes EC não é homogênea, sendo que as classes das Oxirredutases (EC.1), Transferases (EC.2) e Hidrolases (EC.3) correspondem a 85% do total de enzimas no PDB, e os 15% restantes pertencentes às outras três classes (Liasas, EC.4; Isomerases, EC.5; e Ligases, EC.6), como pode ser visto pelo gráfico da Figura 3-3. No entanto, somente uma pequena fração destas enzimas possuem informações sobre seus resíduos catalíticos no CSA. A nova versão do CSA (versão 2.0) (Furnham, et al., 2014) contém 6.262 entradas de resíduos de aminoácidos catalíticos e 335 entradas de cofatores anotadas a partir de evidências experimentais obtidas da literatura, sendo estas correspondentes somente a 1.906 cadeias proteicas (928 estruturas), enquanto que no PDBSProtEC encontram-se 71.980 entradas atribuindo números EC às estruturas do PDB.

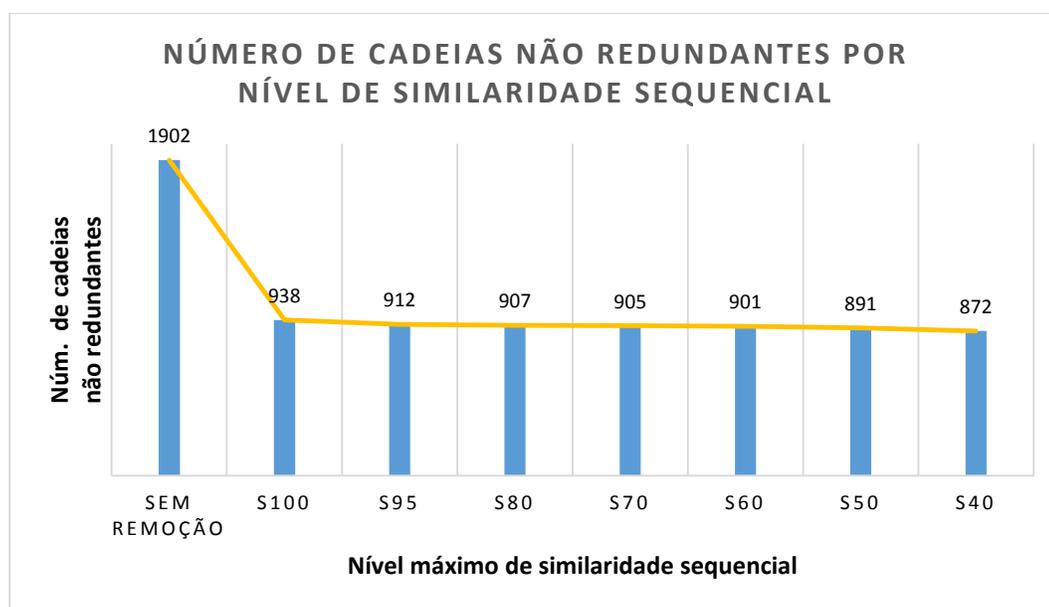


Figura 3-2: Gráfico do número de cadeias não redundantes contidas no CSA para diferentes níveis de similaridade sequencial.

Somente cadeias no PDBSProtEC com apenas um número EC e os quatro níveis da nomenclatura devidamente identificados foram consideradas, o que resultou em um total de 64.858 cadeias. O cruzamento das informações entre PDBSProtEC e CSA resultou num total de 1.555 cadeias com números EC anotados e os respectivos resíduos catalíticos identificados. A remoção de similaridade sequencial foi realizada utilizando-se o *software* CD-HIT (Huang, et al., 2010) para dois níveis de similaridade: 95% e 40%. Após a remoção de redundância, 751 cadeias apresentaram similaridade sequencial inferior a 95%, e 723 cadeias similaridade inferior a 40%.

Dada a dificuldade em se determinar a qualidade das estruturas geradas pelas diferentes técnicas (e.g. difração de raios X, RMN, etc.), foram seguidas algumas indicações apresentadas por

LASKOWSKI (2005). Quanto melhor a resolução de um modelo resolvido a partir da cristalografia por difração de raios-X, maior o nível de detalhamento e maior a acurácia do modelo final. No entanto, não existe uma identificação clara de qual valor adotar para eliminar estruturas com grande quantidade de erros. E para estruturas resolvidas por outras técnicas, como Ressonância Magnética Nuclear (RMN), não existe uma medida similar que possa ser utilizada. Apesar de ser a medida mais clara de qualidade de um modelo, a resolução, por não possuir uma definição única, tende a ser inconsistente, e seu valor pode ser exagerado (Weissig & Bourne, 1999).

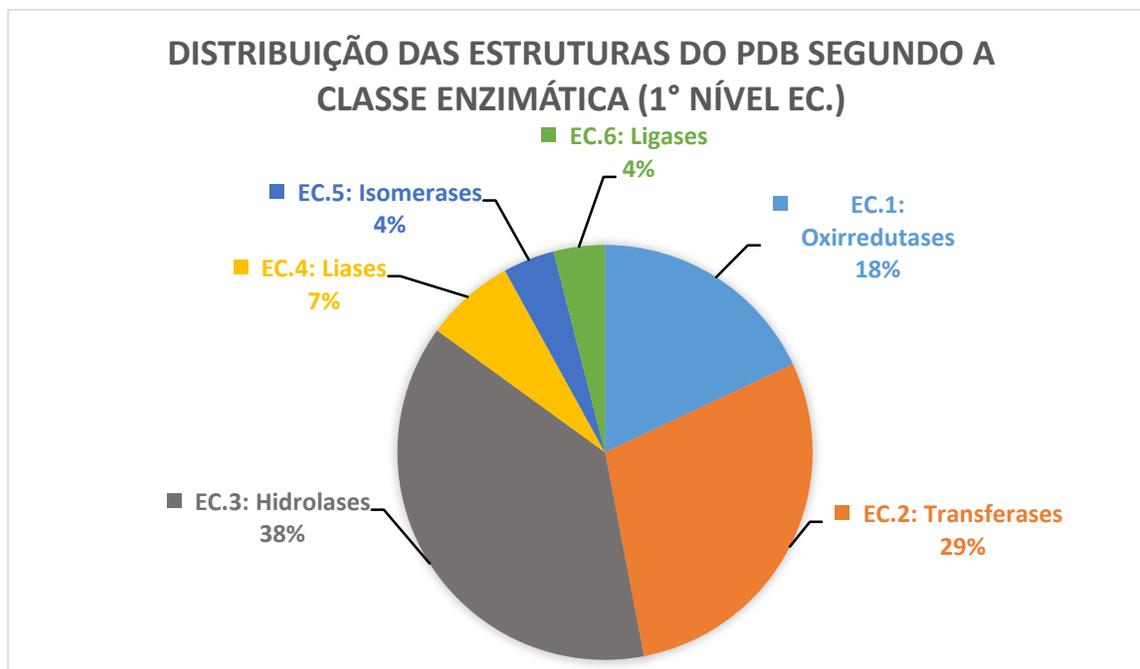


Figura 3-3: Distribuição das estruturas no PDB por classe enzimática (primeiro nível EC).

Outras medidas somente disponíveis a partir da cristalografia por difração de raios X são os fatores-R (R_{value}) e fatores-R livre (R_{free}). Tais fatores reportam a diferença entre os fatores da estrutura calculados a partir do modelo e aqueles obtidos a partir dos dados experimentais. Valores altos do fator-R correspondem a uma baixa concordância entre os dados. Valores no intervalo de 0.40 a 0.60 podem ser obtidos de estruturas totalmente aleatórias. Portanto, valores acima de 0.40 devem ser tratados com cuidado (Laskowski, 2005).

Os fatores-B fornecidos pelas estruturas depositadas no PDB são uma maneira de determinar precisão das coordenadas atômicas. Apesar de esses fatores estarem intimamente relacionados com erros posicionais dos átomos, esta relação não é simples de formular (Tickle, et al., 1998b). Porém, como sugerido por Laskowski, átomos com fatores-B que excedem 40.0 podem ser excluídos para análises que envolvem regiões específicas da molécula.

Dentre as 751 estruturas selecionadas neste trabalho, 744 foram resolvidas por cristalografia de

raios-X e apenas 7 por Ressonância Magnética Nuclear (RMN). As estruturas resolvidas por difração de raios X possuem uma resolução média de 2.12Å sendo 3.4Å a menor resolução (pior qualidade), configurando um conjunto de estruturas com resoluções aceitáveis, uma vez que resoluções piores não fornecem uma boa precisão das cadeias laterais e detalhamento atômico, essenciais para o estudo de pequenas regiões (nanoambiente) das moléculas, como é caso de sítios ativos. Apresentam ainda *fatores-R* e *fatores-R livres* sempre menores que 0.40, sendo os valores máximos de 0.321 e 0.365, respectivamente, indicando que tais estruturas apresentam alta concordância entre os modelos finais e dados experimentais (Kleywegt & Jones, 1997).

Optou-se por considerar apenas estruturas obtidas por cristalografia de raios-X, removendo-se as sete estruturas resolvidas por RMN, por não haver uma medida similar para avaliar a qualidade de estruturas e não haver outras estruturas resolvidas por cristalografia que pudessem substituí-las. Um esquema dos filtros aplicados para a construção da base de dados utilizada neste trabalho encontra-se presente na Figura 3-4.

Como resultado da curadoria dos dados, obteve-se um conjunto de dados final composto por 744 cadeias proteicas totalizando 233.081 resíduos não catalíticos e 2.193 resíduos catalíticos (0.93% do total) divididas segundo suas respectivas sub-subclasses EC (terceiro nível na nomenclatura EC). Como a distribuição das cadeias não é a mesma para todos os números EC, sendo que a maioria destes não contém nenhum representante, é necessário estipular uma quantidade mínima de cadeias que um nível da hierarquia EC deva possuir, isto porque para números EC com poucas cadeias os métodos aplicados neste trabalho serão bastante afetados, reduzindo suas capacidades de replicação (aplicação em outras cadeias – generalização). Assim, escolheu-se somente para estudo as sub-subclasses EC com mais de 15 cadeias não redundantes, resultando em um total de 12 sub-subclasses EC contendo 275 cadeias (37% do total de cadeias), distribuídas segundo mostra a Tabela 3-1.

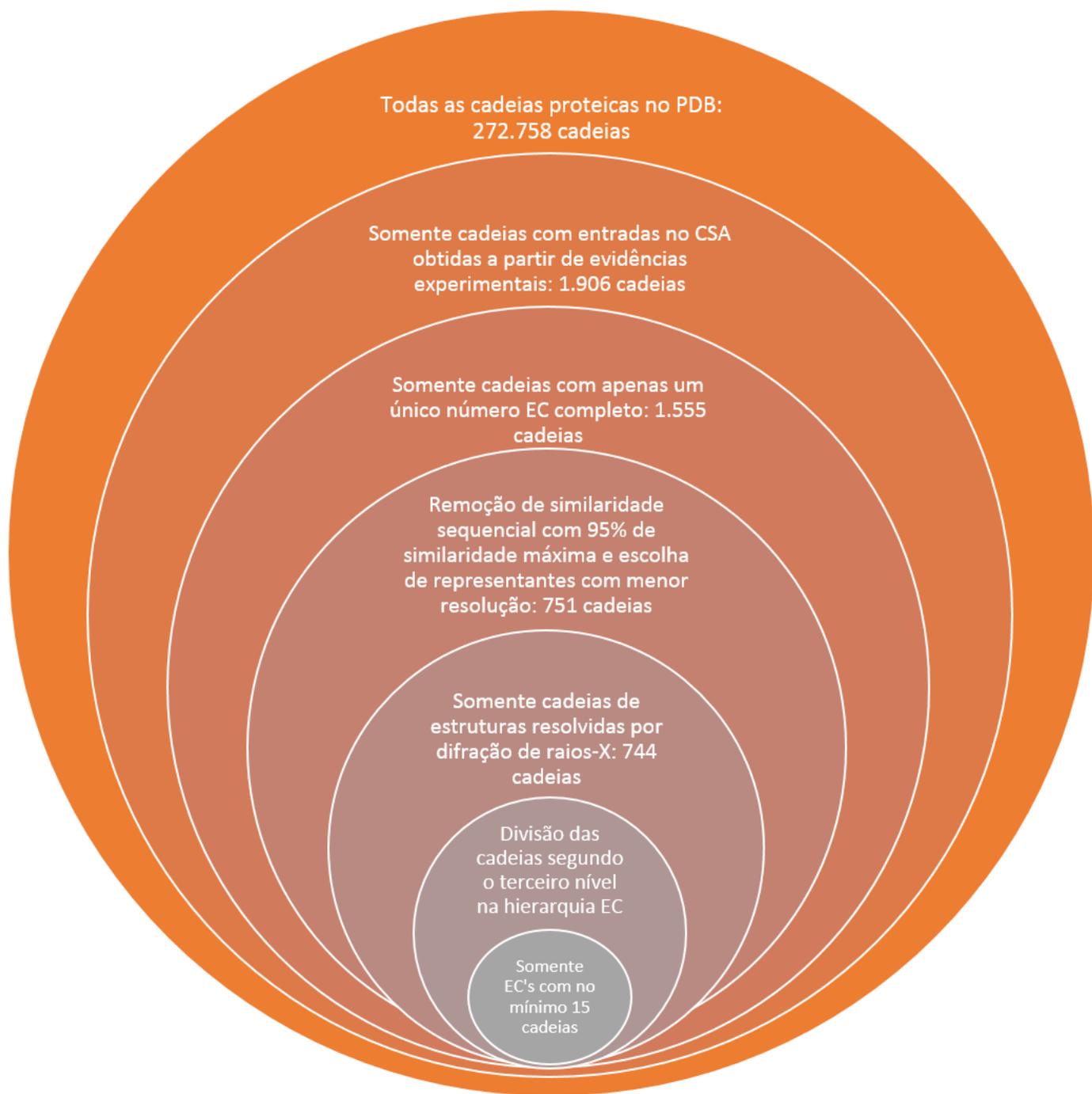


Figura 3-4: Diagrama do processo de filtragem e escolha das cadeias utilizadas neste estudo.

Tabela 3-1: Sub-subclasses EC selecionadas neste estudo para extração de regras e caracterização dos resíduos de aminoácidos catalíticos. Na tabela encontram-se o número EC, nome da sub-subclasse EC, importância e função das enzimas e o número de resíduos de aminoácidos catalíticos e não catalíticos.

NÚMERO EC	NOME	IMPORTÂNCIA/FUNÇÃO	#CADEIAS (#CSR / #ÑCSR)
EC 3.2.1	Glicosidases	Catalisam reações de hidrólise de ligações glicosídicas de polissacarídeos	51 (148/21616)
EC 1.1.1	Desidrogenase tendo NAD ⁺ ou NADP ⁺ como aceitadores	Catalisam reações de oxidação (transferência de elétrons) que têm como doadores grupos CH-OH e aceitadores NAD ⁺ ou NADP ⁺ .	26 (69/9424)
EC 3.1.1	Hidrolases de ester carboxílico	Catalisam reações de hidrólise de ésteres de ácidos carboxílicos	24 (103/7720)
EC 2.3.1	Aciltransferase transferindo grupos não amino-acils.	Catalisam reações de transferência de grupos acil com exceção de grupos amino-acil.	23 (80/8068)
EC 4.2.1	Hidro-liases	Catalisam reações de clivagem de ligações químicas por meios outros mecanismos que não incluem hidrólise e oxidação formando moléculas de água.	23 (68/6159)
EC 2.4.2	Pentossiltransferases	Catalisam reação de transferência de pentoses.	22 (56/6220)
EC 4.1.1	Carboxi-liases	Catalisam reações de descarboxilação, adicionando ou removendo grupos carboxilas de compostos orgânicos.	21 (65/7688)
EC 2.1.1	Metiltransferases	Catalisam reações de transferência de grupos metil (metilação).	18 (42/4937)
EC 2.5.1	Transferase de grupos alquil e aril, não incluindo grupos metil.	Catalisam reações de transferência de grupos alquil e aril, não incluindo grupos metil.	18 (55/5143)
EC 2.7.1	Fosfotransferase tendo álcool como aceitador	Catalisam reações de transferência de grupos fosfato (fosforilação) tendo como aceitador um álcool.	17 (52/5890)
EC 3.1.3	Monoester Fosfórico Hidrolases	Grupo de hidrolases que catalisam a hidrólise de ésteres monofosfóricos com a produção de um mol de ortofosfato.	17 (68/4888)
EC 3.4.21	Serino endopeptidases	Catalisam reações de clivagem de ligações peptídicas nos quais um dos seus aminoácidos catalíticos é uma Serina.	15 (56/4444)

3.4 Blue Star STING, STING_DB e STING_RDB

Blue Star STING é uma plataforma *web* com ferramentas para visualização e análise estrutural de moléculas biológicas (Neshich, et al., 2006). Seus módulos estão concentrados em um único pacote, que visa oferecer um instrumento completo para estudos das macromoléculas, suas estruturas e as relações

estrutura-função. Através dos diversos descritores estruturais de proteínas, o Blue Star STING fornece uma plataforma poderosa para estudos de proteínas, através de mapas de contatos (*Java Table of Contacts – JTC*) (Mancini, et al., 2004), gráficos de parâmetros estruturais para múltiplas estruturas alinhadas (*Multiple Structure Single Parameter – MSSP*) (a ser publicado), e distribuição dos valores dos descritores (*Java Protein Dossier – JPD*) (Neshich, et al., 2004), entre outros. Informações como posição dos resíduos de aminoácido na sequência e na estrutura, busca de padrões, identificação de vizinhança, ligações de hidrogênio, ângulos e distâncias entre átomos são facilmente obtidas, além de dados sobre natureza e volume dos contatos atômicos inter e intracadeias, conservação e relação entre os contatos intracadeias e parâmetros funcionais. Esta plataforma pode ser acessada via web pelo endereço: <http://www.cbi.cnptia.embrapa.br/SMS>.

Para armazenar informações sobre os descritores estruturais de proteínas o Blue Star STING utiliza uma base de dados composta de centenas de milhares de arquivos texto formatados de modo a permitirem a leitura e exibição das informações pelos módulos do Blue Star STING. Conhecida como STING_DB, esta é a maior base de dados de descritores estruturais atualmente disponível de forma gratuita. Em sincronia com o PDB, novas estruturas semanalmente depositadas são obtidas e seus respectivos descritores estruturais são calculados. No entanto, o STING_DB foi desenvolvido para dar suporte ao Blue Star STING, de forma que análises complexas em larga escala são bastante difíceis de realizar, dada a necessidade de interpretar diversos arquivos texto com formatações diferentes. Geralmente, em trabalhos que necessitam de uma análise aprofundada dos dados, pequenos programas eram criados para extrair informações desses arquivos textos e inseri-los em um banco de dados relacional, de forma a facilitar a análise e interoperabilidade da informação. Baseado nas necessidades dos estudos abordados neste trabalho e outros trabalhos do grupo de pesquisa, um banco de dados relacional foi projetado e todas as informações do STING_DB foram utilizadas para criar o STING_RDB (*STING Relational DataBase*). Da mesma forma que o STING_DB, o STING_RDB é sincronizado com o PDB, e semanalmente novas estruturas são adicionadas. Foram incorporados também ao STING_RDB novos descritores estruturais de proteínas que serão discutidos na seção 0, além de algumas variantes dos descritores estruturais que eram calculadas em tempo real quando os módulos eram carregados. Por exemplo tem-se os parâmetros da classe *Sliding Window*, onde uma janela de tamanhos de 3 a 9 aminoácidos é utilizada para percorrer a sequência e calcular descritores de vizinhança sequencial, através da soma ponderada dos valores dos descritores dos resíduos central e seus vizinhos. Tais variantes foram incorporadas e estão disponíveis também no STING_RDB. Sendo uma das principais contribuições deste trabalho, o STING_RDB é ainda maior que o STING_DB em número de descritores estruturais de proteínas, oferecendo uma alternativa rápida, unificada e estruturada das informações

disponíveis no Blue Star STING.

Entre os módulos do Blue Star STING, há o *Java Protein Dossier*, ou simplesmente ^JPD, uma ferramenta de visualização dos descritores estruturais de proteínas em forma de gráfico, com uso extensivo de cores, onde cada descritor é apresentado em um gradiente de cores de forma intuitiva e de fácil visualização. Com o ^JPD, é possível realizar uma comparação dos descritores de duas cadeias de uma proteína ou até mesmo duas cadeias de proteínas diferentes estruturalmente alinhadas pelo próprio módulo. O ^JPD compõe uma das ferramentas que motivaram a realização deste trabalho, como será discutido na seção 7.1.

3.5 Biblioteca de descritores estruturais de proteína – *STING Descriptor Library (SDL)*

Os descritores calculados e utilizados pelo Blue Star STING em seus módulos são gerados utilizando-se programas desenvolvidos pelo próprio grupo de pesquisa e também programas de terceiros, disponíveis em forma de código aberto ou gratuitos para finalidades acadêmicas e de pesquisa. Estes programas recebem como entrada arquivos em formato PDB contendo, entre outras informações, as posições tridimensionais dos átomos que compõem os resíduos das estruturas biológicas. Na saída, geralmente são produzidos arquivos textos com informações sobre os cálculos e os descritores para cada resíduo da molécula, sendo estes arquivos mantidos na base de dados do Blue Star STING (*STING_DB*) para posterior leitura e utilização por seus módulos. Este processo permite ao Blue Star STING integrar o cálculo dos descritores com diversas ferramentas para suas visualizações. No entanto, estes programas não se encontram unificados em uma biblioteca que possibilite o cálculo parametrizado dos descritores, além de dificultarem uma possível expansão do Blue Star STING. Pensando nisto, neste trabalho foi criada uma biblioteca para unificar e organizar os cálculos e operações de entrada e saída de dados para os descritores do Blue Star STING, bem como oferecer uma plataforma capaz de calcular descritores estruturais de proteínas considerando diferentes parametrizações dos algoritmos.

A biblioteca foi implementada utilizando-se a linguagem de programação *Java* de forma a aumentar a interoperabilidade entre as rotinas de cálculo e os módulos do Blue Star STING (*Java Applets*). Sendo uma linguagem orientada a objetos e polimórfica, é possível reduzir a redundância de códigos e facilitar a criação de novos descritores.

Intitulada de *STING Descriptor Library (SDL)* a biblioteca reúne rotinas para cálculo, leitura e escrita dos descritores estruturais. Na *SDL*, todo descritor realiza a implementação da interface *Java Descriptor*, que fornece métodos para a recuperação dos valores disponíveis para um dado descritor. Há ainda classes específicas que facilitam a implementação de novos descritores, como a classe

ResidueDescriptor, que é utilizada por descritores definidos ao nível de resíduos. Associado a cada descritor, existem classes específicas que realizam o cálculo de cada descritor (extensões da classe abstrata *AbstractDescriptorCalculator*), que fornecem possibilidade de parametrização dos cálculos e algoritmos, como por exemplo, raios das esfera sondas, tamanhos dos segmentos de sequências, átomos centrais, etc. A leitura e escrita dos descritores tanto em arquivos textos como em bancos de dados relacionais, são realizadas através de objetos que implementam a interface *DescriptorProvider*. Esta interface possui apenas dois métodos (*read* e *write*), sendo que a classe *FlatFileDescriptorProvider* lê e escreve em arquivos textos e objetos da classe *DatabaseDescriptorProvider* realizam estas operações em um banco de dados relacional. Esta representação auxilia da manipulação dos descritores, uma vez que todos são objetos pertencentes a uma mesma superclasse e implementam métodos comuns, explorando o polimorfismo da linguagem.

A partir desta hierarquia, foi possível criar classes específicas para cálculos de descritores derivados de outros descritores, como o caso de descritores de vizinhança. Na SDL, classes que estendem a superclasse abstrata *NeighborCalculator* fornecem métodos para cálculo de descritores de vizinhança, utilizando como entrada objetos que implementam a interface *NeighborDescriptors*. Atualmente, estão implementados os descritores de vizinhança discutidos na seção 4.3. No entanto, novas implementações podem facilmente ser criadas através da interface *NeighborCalculator* e da superclasse.

Dentre todos os descritores do Blue Star STING, aqueles produzidos pelo laboratório de pesquisa foram reimplementados na SDL utilizando linguagem Java. Os demais descritores são gerados utilizando-se programas executados via chamadas aos processos externos em Java (*Process API* e *Runtime API*), sendo suas saídas devidamente processadas. Uma alternativa seria o uso da *Java Native Interface API* (JNI) que permite executar código em outras linguagens sem a necessidade de criação de novos processos. No entanto, seu uso restringe a programas onde se tem acesso ao código-fonte, pois são necessárias alterações nos códigos dos programas nativos para permitirem comunicação com o programa em Java.

Anteriormente à SDL, o cálculo dos descritores do Blue Star STING era realizado utilizando-se as unidades assimétricas contidas nos arquivos em formato PDB disponíveis no *Protein Data Bank*. A unidade assimétrica corresponde à menor porção de uma estrutura cristalizada que através da aplicação de operações simétricas (translação, rotação ou combinação de ambas) sejam capazes de gerar uma célula unitária completa (unidade repetitiva do cristal). A célula unitária por sua vez pode ser transladada no espaço de forma a se obter todo o cristal (Figura 3-5). Sendo assim, muitas vezes uma unidade assimétrica pode não representar o arranjo biológico natural de uma proteína, mas uma porção deste arranjo ou ainda múltiplos arranjos. Por outro lado, os arranjos biológicos (*biological assembly*) são arranjos

macromoleculares que representam (ou acredita-se representar) a forma funcional de uma molécula, e são obtidos através da aplicação de operações de translação, rotação ou ambas em unidades assimétricas. Um arranjo biológico pode ser obtido de uma ou múltiplas cópias de uma unidade assimétrica ou apenas uma porção da unidade assimétrica. Na SDL, foi considerada a leitura de arquivos nos formatos PDB e mmCIF (Fitzgerald, et al., 2005), além de arquivos contendo as unidades assimétricas ou biológicas utilizando a biblioteca de código aberto BioJava (versão 3.0.8) (PrliĆ, et al., 2012). Segundo nota publicada em Maio de 2013 no *website* do PDB, modificações nos processos de submissão e armazenamento dos arquivos de grandes estruturas foram implementadas, entrando em produção no início de 2014. Estas modificações incluem a disponibilidade dos arquivos anteriormente divididos em diversos arquivos menores no formato PDB, e um único arquivo nos formatos mmCIF e PDBx [http://www wwpdb.org/news/news_2013.html#22-May-2013]. Assim, a SDL garante a continuidade das operações do Blue Star STING ao prover a leitura destes novos formatos.

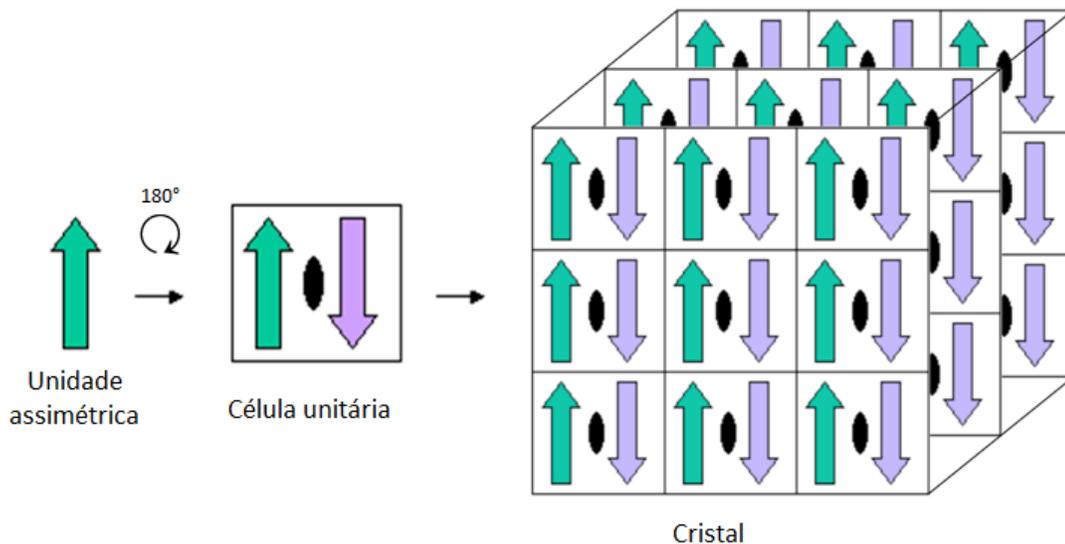


Figura 3-5: Esquema didático da composição de um cristal e aplicação de operações de translação e rotação a partir de uma unidade assimétrica para obtenção da célula unitária (unidade repetitiva do cristal). Figura baseada naquela disponível nos recursos educacionais no website do PDB (http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/bioassembly_tutorial.html).

Outra vantagem da SDL é que seus cálculos foram tornados totalmente parametrizáveis, de forma que os descritores podem ser calculados segundo as necessidades do usuário. É possível utilizar raios arbitrários para esferas sondas, tamanhos de janelas de aminoácidos, diferentes tabelas com valores experimentais, distâncias máximas e mínimas entre átomos, entre outros parâmetros. Assim, os módulos do Blue Star STING podem ser reimplementados considerando o cálculo dos descritores em tempo real, segundo parâmetros definidos pelo usuário através de uma interface apropriada. Estes descritores são calculados pela SDL e podem ser visualizados sem a necessidade de criação e armazenamento de

arquivos no STING_DB ou registros no STING_RDB.

Para garantir a qualidade das informações geradas, a SDL ainda possui classes responsáveis por validar e checar os valores calculados para seus diversos descritores. Como uma expansão do *STING Quality Assessment* (STING_QA) (Neshich, et al., 2006), é possível recuperar as mesmas estatísticas sobre os descritores bem como outras informações considerando diversos níveis de granularidade, como por exemplo médias e desvio padrões dos descritores, e resíduos com valores nulos para cada descritor.

Dessa forma, a *STING Descriptor Library*, utilizando uma linguagem de programação moderna, possibilita e facilita a expansão do Blue Star STING, contribuindo para uma melhoria dos serviços gratuitos disponíveis pela plataforma.

3.6 Base de dados relacional de descritores estruturais e físico-químicos de proteínas: STING_RDB

Como descrito anteriormente, o Blue Star STING opera utilizando uma base de dados composta por arquivos textos divididos por tipo de descritor. Assim, para cada estrutura do PDB, tem-se um arquivo texto para cada descritor presente no Blue Star STING.

Dada a dificuldade em realizar análises em larga escala sobre estes dados, a criação de um modelo relacional e posterior implementação em um sistema computadorizado traz diversas vantagens. Por este motivo, foi criado um banco de dados relacional para abrigar todos os descritores já presentes no STING_DB e novos descritores criados a partir deste trabalho.

Anteriormente ao STING_RDB, os dados de todos os descritores estruturais de proteínas eram agrupados por uma ferramenta em um único arquivo XML, que permitia sua posterior leitura e interpretação pelos módulos do Blue Star STING. Esta ferramenta realiza a tarefa de ler os arquivos textos existentes para uma estrutura dada como entrada e agrupar todos os valores lidos em um arquivo XML, descrevendo toda a proteína através de seus descritores estruturais. Porém, para que isto seja possível, é preciso calcular e gerar os arquivos textos de cada descritor e, assim, diversos programas escritos em diversas linguagens de programação, tais como C++, Perl e Fortran, são executados por outra ferramenta que realiza o cálculo geral dos descritores. Esse cálculo gera os arquivos e armazena-os nos locais apropriados para cada descritor permitindo, assim que posteriormente o arquivo XML seja gerado.

Para alimentar o modelo relacional, é preciso extrair as informações destes arquivos textos utilizando a ferramenta descrita acima e, então, realizar uma interpretação do XML gerado e preencher as tabelas apropriadas no banco de dados. Ou ainda, é possível realizar o cálculo dos descritores e salvá-los diretamente no banco de dados, sem a necessidade de gerar arquivos XML's intermediários. Escolheu-se utilizar esta segunda abordagem através do uso da *STING Descriptor Library* (SDL).

Através da implementação adequada da interface *DescriptorProvider*, foi criada a classe *DatabaseDescriptorProvider*, que oferece implementações para os métodos de leitura e escrita em um banco de dados relacional. Toda a comunicação e mapeamento entre os objetos em Java e Tabelas no banco de dados foram realizadas utilizando-se o padrão de mapeamento objeto-relacional (*Object Relational Mapping* - ORM), implementados pelo *framework* de código aberto Hiberante ORM (<http://hibernate.org/orm/>). Dessa forma, o Hibernate realiza um mapeamento entre objetos Java e tabelas no banco de dados de forma totalmente transparente, instanciando objetos Java a partir da leitura de tabelas e salvando conteúdo dos objetos em tabelas apropriadas. O Hibernate permite a integração com qualquer sistema de gerenciamento de banco de dados, além de agilizar o processo de produção de novas aplicações. Um diagrama contendo o modelo relacional do STING_RDB encontra-se no Apêndice A.

Capítulo 4 - Descritores de proteínas

O estudo *in silico* de proteínas é muitas vezes realizado através de propriedades extraídas a partir dos modelos computacionais das estruturas proteicas. Essas propriedades recebem, portanto, o nome de descritores de proteínas, uma alusão aos descritores de imagens nas áreas da Computação de Visão Computacional e Processamento de Imagens.

Descritores de proteínas são utilizados para descrever o aspecto dessas proteínas segundo diferentes tipos de propriedades, sejam propriedades químicas, físicas, físico-químicas, estruturais, geométricas ou de conservação de estrutura primária. Estes descritores são fundamentais para o entendimento *in silico* das proteínas e, possibilitam a realização de diversos estudos e análises, tanto por especialistas quanto por programas de computador.

Neste capítulo são apresentados os diferentes tipos de descritores de proteínas disponíveis na base de dados utilizada neste trabalho: o Blue Star STING.

4.1 Descritores físico-químicos e estruturais de proteínas

Atualmente, o Blue Star STING apresenta 27 tipos diferentes e independentes de descritores estruturais de proteínas, sendo que um total de 1,175 variações destes descritores estão pré-calculados, utilizando diferentes parametrizações e armazenados no STING_RDB. Dentre estes descritores, foram escolhidos para utilização neste trabalho aqueles que apresentam maiores probabilidades de estarem associados com processos de reconhecimento de padrões em proteínas. Os descritores escolhidos estão listados na Tabela 4-1 e detalhados nas seções 0, 4.2 e 4.3. Procurando uma definição de nanoambiente dos resíduos catalíticos através dos descritores físico-químicos e estruturais, num primeiro momento descartaram-se descritores referentes à conservação dos aminoácidos, uma vez que estes parâmetros são uma medida de um conjunto de proteínas homólogas, e não refletem nenhuma característica presente na estrutura proteica.

4.1.1 Potencial Eletrostático

Os átomos que compõem os aminoácidos das proteínas podem, em determinadas condições, apresentar carga elétrica, que interagem com outras regiões carregadas da própria proteína ou ainda com outras moléculas e/ou íons de seu ambiente. Portanto, em um determinado ponto do espaço é possível calcular o potencial eletrostático devido às cargas presentes nas macromoléculas ao redor deste ponto.

Blue Star STING utiliza o programa *Delphi* (Rocchia & Neshich, 2007) para cálculos de potencial eletrostático para todas as proteínas presentes no PDB, através da resolução da equação linearizada de Poisson-Boltzman (PBE) para uma solução de sal monovalente simétrico:

$$\vec{\nabla}[\varepsilon_r(\vec{r})\vec{\nabla}\varphi(\vec{r})] = \frac{1}{\varepsilon_0} \left[\rho^{carga}(\vec{r}) - \frac{\varepsilon_{solv}\kappa^2(\vec{r})}{4\pi} \varphi(\vec{r}) \right] \quad \text{Equação 4-1}$$

onde $\varepsilon_r(\vec{r})$ é a constante dielétrica relativa local, ε_{solv} é a constante dielétrica do solvente, $\rho^{carga}(\vec{r})$ é a distribuição de carga sobre o soluto, e κ^2 é o parâmetro de Debye, que é igual a:

$$\kappa^2 = \frac{8\pi e^2 C}{\varepsilon_{solv} k_B T} \quad \text{Equação 4-2}$$

se \vec{r} aponta para o solvente ou zero caso contrário, onde C é a concentração em massa do sal, k_B é a constante de Boltzmann e T é a temperatura absoluta.

O Blue Star STING mantém pré-calculados os valores de potencial eletrostático para todos os átomos de cada resíduo de aminoácido da proteína, além de outras quatro variações: potencial eletrostático no carbono- α , no último átomo pesado da cadeia lateral (LHA da sigla em inglês *Last Heavy Atom*), valor médio dos átomos do resíduo, e soma dos potenciais dos átomos na superfície da molécula para cada resíduo. Assim, para cada resíduo de aminoácido, quatro tipos de descritores de potencial eletrostático são gerados (Rocchia & Neshich, 2007). Através da SDL, é possível recuperar estas informações para todos os resíduos de uma proteína ou até mesmo para todos os átomos de uma proteína.

Tabela 4-1: Subconjunto de descritores estruturais de proteínas presentes no STING_RDB e utilizados neste trabalho.

Descritor	Total de atributos	Atributos de vizinhança
Acessibilidade em isolamento	2	0
Espongicidade (C α e LHA raios 3 a 7Å)	10	40
Densidade (C α e LHA raios 3 a 7Å)	10	40
Energia de contatos estabelecidos	15	60
Energia de contatos não usados	15	60
Potencial eletrostático	4	16
Hidrofobicidade	2	0
Densidade energética de contatos (C α e LHA raios 3 a 7Å)	10	40
Ordem de Cross Link	27	108
Ordem de Cross Presence	27	108
Grafos de contatos	17	0
Cavidade e Pockets	19	0
Distâncias	3	0
Subtotal	161	472
TOTAL	633 descritores	

4.1.2 Hidrofobicidade

Aminoácidos são classificados como hidrofóbicos ou hidrofílicos, dependendo de suas afinidades para estabelecerem ligações de hidrogênio com as moléculas de água. Existem diversas escalas que definem a hidrofobicidade relativa dos resíduos de aminoácidos: quanto mais positivo é seu valor, mais

hidrofóbicos são os aminoácidos localizados em dada região da molécula. No cálculo de hidrofobicidade no Blue Star STING são utilizadas as escalas definidas por RADZICKA & WOLFENDEN (1988) e também a escala definida por KYTE & DOOLITTLE (1982) apresentadas na Tabela 4-2.

A Hidrofobicidade de um resíduo de aminoácido i do tipo t é calculada utilizando-se o valor estipulado na escala adotada para o tipo de aminoácido considerado ($hydro(t)_i^{scale}$), ponderado pela acessibilidade relativa ao solvente (razão entre a acessibilidade ao solvente do resíduo i (Acc_i) e a acessibilidade máxima apresentada pelo aminoácido de tipo t ($AccMax(t)$, calculada conforme descrito na seção 4.1.4)), conforme mostra a Equação 4-3.

$$Hidrofobicidade_i = \frac{Acc_i}{AccMax(t)} hydro(t)_i^{scale} \quad \text{Equação 4-3}$$

Utilizando a SDL, é possível obter tais descritores de hidrofobicidade empregando as escalas já definidas na biblioteca, ou ainda utilizando outras escalas fornecidas para cálculo.

Tabela 4-2: Escala de Hidrofobicidade definida por RADZICKA & WOLFENDEN (1988) e KYTE & DOOLITTLE (1982) para cada tipo de aminoácido, com suas características físico-químicas.

Aminoácido	Polaridade Da Cadeia Lateral	Carga Da Cadeia Lateral Ph 7,4	Constante De Hidrofobicidade de Radzicka	Constante De Hidrofobicidade de Kyte-Doolittle
ALA	Apolar	Neutro	1,81	1,80
ARG	Polar	Positiva	-14,92	-4,50
ASN	Polar	Neutro	-6,64	-3,50
ASP	Polar	Negativa	-8,72	-3,50
CYS	Apolar	Neutro	1,28	2,50
GLN	Polar	Neutro	-5,54	-3,50
GLU	Polar	Negativa	-6,81	-3,50
GLY	Apolar	Neutro	0,94	-0,40
HIS	Polar	Neutro/Positiva	-4,66	-3,20
ILE	Apolar	Neutro	4,92	4,50
LEU	Apolar	Neutro	4,92	3,80
LYS	Polar	Positiva	-5,55	-3,90
MET	Apolar	Neutro	2,35	1,90
PHE	Apolar	Neutro	2,98	2,80
PRO	Apolar	Neutro	3,5	-1,60
SER	Polar	Neutro	-3,4	-0,80
THR	Polar	Neutro	-2,57	-0,70
TRP	Apolar	Neutro	2,33	-0,90
TYR	Polar	Neutro	-0,14	-1,30
VAL	Apolar	Neutro	4,04	4,20

Fontes (Radzicka & Wolfenden, 1988) e (Kyte & Doolittle, 1982)

4.1.3 Contatos e Densidade de Energia de Contatos

A maior parte dos resíduos de aminoácidos estabelece diferentes contatos com outros resíduos de aminoácidos, além da ligação peptídica que une a sequência primária de organização das proteínas. A plataforma Blue Star STING diferencia 14 tipos de contatos intra e intercadeias e mais 12 tipos de

contatos entre cadeias proteicas e moléculas de DNA/RNA, que são armazenados em sua base de dados em nível atômico, ou seja, possui informações sobre quais pares de átomos estão estabelecendo contato e a energia equivalente ao tipo de contato. Blue Star STING define contatos baseados nas distâncias relativas entre átomos ou aminoácidos (no caso de contatos aromáticos), que estejam dentro do intervalo de mínimo e máximo estabelecido para cada tipo de contato, e os tipos dos átomos considerados (da Silveira, et al., 2009). Na Tabela 4-3 encontram-se os 14 tipos de contatos internos (entre resíduos de aminoácidos de uma mesma cadeia) utilizados neste trabalho, juntamente com distâncias mínimas e máximas e energia associada a cada contato, pré-definidas pelo Blue Star STING.

O descritor definido como Densidade de Energia de Contatos (ou CED – *Contacts Energy Density*, em inglês) utiliza uma sonda esférica centrada no carbono- α ou LHA de cada resíduo. Todos os contatos entre resíduos de aminoácidos dentro da sonda esférica têm suas energias somadas e divididas pelo volume da sonda esférica. É possível que um átomo dentro da sonda esférica estabeleça contato com outro átomo fora da sonda esférica. Neste caso, o contato não é considerado no cálculo, uma vez que ambos os átomos estabelecendo contato necessitam estar dentro da sonda esférica. Ainda que haja uma contribuição energética deste contato para o nanoambiente delimitado pela sonda, não faz sentido considerar somente uma parcela da energia do contato, fragmentando sua contribuição energética. Por outro lado, uma sonda esférica de maior raio irá, por fim, acabar englobando estes contatos excluídos, reduzindo a sobreposição dos valores calculados utilizando-se sondas esféricas de diferentes raios.

Dessa forma, para cada resíduo, considerando apenas uma esfera sonda, há 4 atributos para este descritor, pois se consideram de forma separada os contatos internos (entre resíduos de uma mesma cadeia) e externos (entre resíduos de cadeias diferentes). Para este trabalho, utilizaram-se somente os descritores de contatos internos e esfera de raio 3 a 7Å, com incremento de 1Å (valores pré-calculados presentes no STING_RDB), totalizando de 10 atributos para a Densidade de Energia de Contatos.

Utilizando-se a SDL, é possível adotar outros raios para as sondas esféricas (além dos pré-definidos pelo Blue Star STING), bem como diferentes configurações para distâncias mínima e máxima entre dois átomos de diferentes resíduos para o estabelecimento de contato e suas respectivas energias.

4.1.4 Acessibilidade

Nem todos os aminoácidos que constituem as proteínas estão acessíveis ao solvente. Alguns estão enterrados no núcleo das estruturas proteicas e, normalmente, estão associados à estabilidade de tais estruturas, promovendo contatos que não são estabelecidos com o solvente, como por exemplo, os contatos hidrofóbicos entre átomos de carbono dos resíduos de aminoácidos.

Tabela 4-3: Diferentes tipos de contatos internos armazenados no STING_RDB e seus respectivos valores de distâncias mínimas e máximas para estabelecimento de contato e energia de interação.

Tipo de Interação	Sub-tipo	Distâncias (min-max) (Å)	Energia Associada (kcal/mol)
Hidrofóbica		2.0-3.8	0.06
Carregada	<i>Atrativa</i>	2.0-12.0	10.00
	<i>Repulsiva</i>		
Ligação de hidrogênio	<i>Entre átomos das cadeias principais</i>	2.0-3.2	2.60
	<i>Entre átomos das cadeias principais contendo uma molécula de água</i>		
	<i>Entre átomos das cadeias principais contendo duas moléculas de água</i>		
	<i>Entre átomos da cadeia principal e lateral</i>		
	<i>Entre átomos da cadeia principal e lateral contendo uma molécula de água</i>		
	<i>Entre átomos da cadeia principal e lateral contendo duas moléculas de água</i>		
	<i>Entre átomos das cadeias laterais</i>		
	<i>Entre átomos das cadeias laterais contendo uma molécula de água</i>		
	<i>Entre átomos das cadeias laterais contendo duas moléculas de água</i>		
Aromática		0.0-8.0	1.50
Ligação Dissulfeto		1.5-2.8	85.00

Fonte: http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/energy_contacts_table.html

Os limites pré-definidos para o volume dos átomos é tido como o volume de van der Waals, ou seja, cada átomo é representado como uma esfera rígida de raio pré-estabelecido, de acordo com o tipo do átomo. A sobreposição de todos os volumes do arranjo tridimensional de átomos da estrutura de cada proteína gera sua superfície de van der Waals. Para calcular a área da superfície acessível que cada resíduo fornece para a superfície proteica, utiliza-se uma sonda esférica com raio igual ao da molécula de água (1.4Å) que percorre toda a superfície de van der Waals. A superfície gerada pelo centro geométrico da esfera sonda é a superfície acessível ao solvente (Figura 4-1). Utilizar uma esfera de raio zero irá retornar, portanto, a superfície de Van der Waals (Lee & Richards, 1971). Da mesma forma, a superfície molecular pode ser obtida pela intersecção da superfície da esfera sonda com a superfície dos átomos que compõem a molécula.

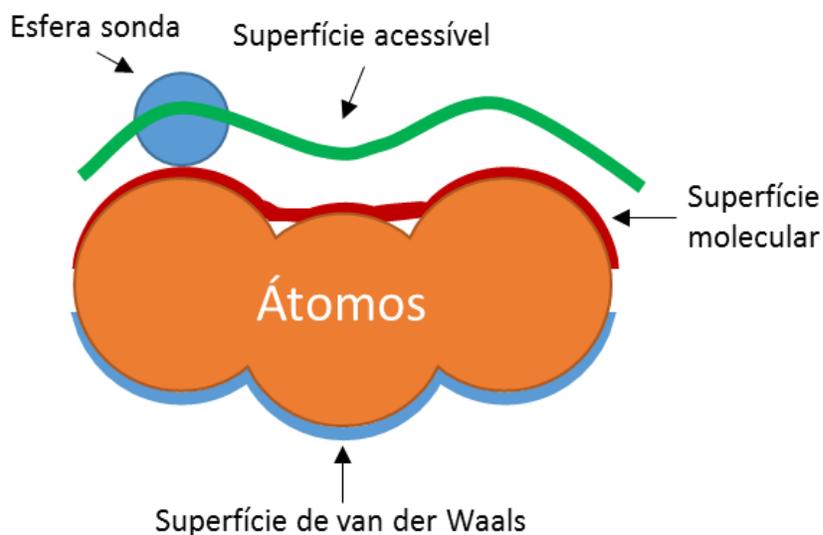


Figura 4-1: Área acessível ao solvente de acordo com a superfície de van der Waals.

Para calcular a área acessível que cada resíduo contribui para a superfície proteica, a plataforma Blue Star STING utiliza o algoritmo *SURFV* (Sridharan, et al., 1992). A partir do cálculo fornecido pelo *SURFV*, é possível obter informações sobre a acessibilidade de cada resíduo em isolamento (cadeia isolada) e em complexo (todas as cadeias) e, então, calcular a diferença entre as acessibilidades para obter informações sobre a presença de tais resíduos nas regiões de interface da proteína, quando a estrutura representa um complexo proteico.

O Blue Star STING também utiliza valores em isolamento para o cálculo da acessibilidade relativa, definida como a razão entre a acessibilidade em isolamento e a acessibilidade máxima estipulada para cada tipo de resíduo de aminoácido. Para encontrar a acessibilidade máxima de cada tipo de aminoácido, Blue Star STING emprega uma abordagem em que quatro estruturas contendo um aminoácido de mesmo tipo, localizado no C-terminal da cadeia, são editadas de forma que o C-terminal seja isolado no “vácuo”. Utilizando-se então o programa *SURFV*, a acessibilidade é calculada para as quatro estruturas. Dos quatro valores de acessibilidade para o C-terminal, são extraídas a acessibilidade mínima e máxima, sendo a diferença percentual entre estes valores utilizada como fator de multiplicação para ajustar o valor máximo obtido, e estabelecer assim a acessibilidade máxima de cada tipo de aminoácido. A Tabela 4-4 traz os valores das acessibilidades máximas obtidas pelo processo acima, para os 20 tipos de aminoácidos.

Na SDL, foi possível a utilização de outros programas para o cálculo de acessibilidade, como o *Surface Racer* (Tsodikov, et al., 2002) (já utilizado pelo Blue Star STING para o cálculo de Curvatura) e NACCESS© (não publicado, S. Hubbard and J. Thornton 1992-6). Também foi implementado na SDL o algoritmo para o cálculo de acessibilidade de Shrake-Rupley (Shrake & Rupley, 1973). Estes algoritmos calculam a superfície acessível aproximada, uma vez que obter tal superfície é computacionalmente

proibitivo, assim, diferentes algoritmos podem resultar em diferentes valores de área acessível. Portanto, possuir o cálculo da área acessível utilizando diversos algoritmos possibilita uma comparação e até mesmo um consenso entre a acessibilidade de um átomo.

Tabela 4-4: Acessibilidades máximas por tipo de aminoácido. Acessibilidades obtidas a partir do software SURFV para os 20 tipos de aminoácidos de quatro estruturas apresentando o aminoácido de mesmo tipo no C-terminal da cadeia (ASA1, ASA2, ASA3 e ASA4), bem como a diferença percentual entre os valores mínimo e máximo calculados. A diferença percentual é então utilizada para ponderar a acessibilidade máxima de cada tipo de aminoácido, estabelecendo um valor limite para cada tipo de aminoácido.

Tipo AA	ASA 1	ASA 2	ASA 3	ASA 4	$[ASA_{max} - ASA_{min}]%$	ASA Max.
ALA	231.680	232.768	233.585	234.345	2.0	239.0
ARG	362.479	366.063	371.165	372.958	3.0	384.1
ASN	277.159	278.898	282.891	284.687	3.0	293.2
ASP	271.343	273.849	276.055	276.827	2.0	282.4
CYS	255.831	256.499	260.818	262.35	3.0	270.2
GLN	302.331	303.265	309.626	310.915	3.0	320.2
GLU	292.397	303.215	304.521	312.052	7.0	333.9
GLY	202.875	205.294	205.443	205.471	2.0	209.6
HIS	303.173	305.122	307.905	317.645	5.0	333.5
ILE	302.749	302.944	305.736	306.417	2.0	312.5
LEU	306.523	310.613	311.242	311.438	2.0	317.7
LYS	339.108	340.594	343.988	347.742	3.0	358.2
MET	317.302	317.461	321.639	325.357	3.0	335.1
PHE	325.05	327.259	340.984	342.137	5.0	359.2
PRO	265.778	265.784	266.094	266.453	1.0	269.1
SER	241.856	242.445	243.84	245.213	2.0	250.1
THR	265.434	266.091	267.756	269.504	2.0	274.9
TRP	363.036	368.853	385.297	387.278	7.0	414.4
TYR	352.526	355.254	359.056	360.794	3.0	371.6
VAL	276.272	278.396	279.453	280.703	2.0	286.3

Fonte: http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/solvent_accessible_area.html

4.1.5 Ordem de Cross Link

Devido ao dobramento da sequência de aminoácidos na estrutura tridimensional, os resíduos distantes na sequência primária são colocados próximos na conformação assumida pela proteína funcional, e podem, portanto, interagir entre si. O parâmetro *Cross Link Order*, presente na plataforma Blue Star STING, leva essa característica em consideração. A palavra “ordem” no nome do descritor é devido à relação do número de contatos estabelecidos entre segmentos de aminoácidos, separados por um número mínimo de aminoácidos na estrutura primária da proteína, e que foram encontrados dentro de uma sonda esférica de raio r centrada nos carbonos α e β e no LHA de qualquer resíduo de uma proteína. Para cada segmento somente é contabilizado um contato, assim, se o resíduo central estabelecer

mais de um contato com outros átomos dentro da esfera, que estão distantes na sequência, somente um dos contatos será considerado (Figura 4-2). No Blue Star STING, estão pré-calculadas e armazenadas 27 variações deste descritor, considerando segmentos de tamanhos mínimos com 15, 20 e 30 aminoácidos, e esferas com raios de 3.5, 5.0 e 8.5Å.

4.1.6 Ordem de Cross Presence

Seguindo a mesma definição do descritor *Ordem Cross Link*, apresentado na seção 4.1.5, o descritor *Cross Presence* contabiliza a presença dos aminoácidos dentro de sondas esféricas de raios r , centradas nos carbono- α , carbono- β e LHA, que estão distantes na estrutura primária da proteína por no mínimo l aminoácidos (tamanho do segmento), mesmo que os resíduos não estejam estabelecendo nenhum contato entre si. É contabilizado somente um resíduo para cada segmento. Da mesma forma que para o descritor de *Ordem de Cross Link* o Blue Star STING, mantém pré-calculado 27 variações deste descritor segundo os mesmos parâmetros.

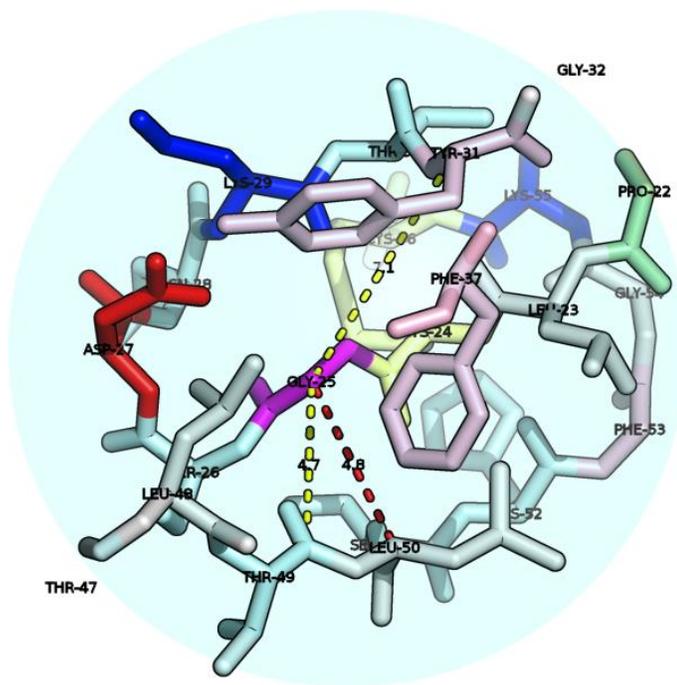


Figura 4-2: Esfera sonda centrada no carbono alfa do resíduo GLY-25 (magenta), que estabelece contatos com os resíduos THR-31, THR-49 e THR-50 (linhas tracejadas). Considerando segmentos de 5 aminoácidos o resíduo GLY-25 possui Ordem de Cross Link igual a 2, pois os resíduos THR-49 e THR-50 estão a uma distância na sequência menor que o tamanho do segmento (5 aminoácidos). Assim, são contabilizados somente os dois contatos em amarelo.

4.1.7 Contatos não Usados

Na seção 4.1.3, discute-se como são definidos e armazenados os descritores de contatos estabelecidos pelos resíduos na estrutura tridimensional das proteínas e complexos proteicos. No Blue Star STING, são registradas estatísticas do número máximo de contatos estabelecidos para cada um dos

14 tipos de contatos internos e para cada um dos 20 aminoácidos. À medida que novas estruturas são depositadas no PDB, estes valores podem ser modificados, caso algum aminoácido de uma nova estrutura estabeleça um número maior de contatos de certo tipo que o previamente obtido de toda a base de dados. No entanto para atualização destes valores, somente são consideradas estruturas resolvidas através da cristalografia por difração de raios-x e com resolução melhor que 2.0Å. A Tabela 4-5 apresenta os valores dos números máximos de contatos estabelecidos para os 20 aminoácidos atualmente utilizada pelo Blue Star STING. Comparando os valores do número de contatos estabelecidos por cada aminoácido em todas as proteínas presente no PDB, é possível calcular quantos contatos não foram estabelecidos em relação ao valor máximo encontrado para os mesmos tipos de aminoácido e contato. O número de contatos **não** utilizados estabelece um **potencial** de contatos, que não leva em conta o meio ambiente em que cada aminoácido está inserido.

Tabela 4-5: Número máximo de contatos estabelecidos por cada tipo de aminoácido. O descritor de contatos não usados é calculado subtraindo-se o número de contatos estabelecidos pelos resíduos de aminoácidos do valor máximo presente na tabela, sendo que isto fornece uma ideia de potencial de contatos. Esta tabela é atualizada sempre que novas estruturas são depositadas no PDB.

Tipo de Interação	A	R	N	D	C	Q	E	G	H	I
Hidrofóbica	13	18	12	12	10	15	16	13	20	17
Carregada atrativa	0	46	0	46	0	0	46	0	42	0
Carregada repulsiva	0	45	0	38	0	0	48	0	42	0
Ligação de Hidrogênio cadeia principal – cadeia principal	7	6	6	6	7	6	6	8	6	6
Ligação de Hidrogênio cadeia principal – cadeia lateral	4	7	6	7	3	6	7	4	6	3
Ligação de Hidrogênio cadeia lateral – cadeia lateral	0	4	5	5	0	5	4	0	3	0
Ligação de Hidrogênio cadeia principal – cadeia principal (1-H ₂ O)	5	4	4	4	5	5	4	6	4	5
Ligação de Hidrogênio cadeia principal – cadeia lateral (1-H ₂ O)	5	9	6	6	4	6	7	5	6	3
Ligação de Hidrogênio cadeia lateral – cadeia lateral (1-H ₂ O)	0	4	6	6	0	5	6	0	3	0
Ligação de hidrogênio cadeia principal – cadeia principal (2-H ₂ O)	6	5	5	6	4	6	4	5	4	4
Ligação de Hidrogênio cadeia principal – cadeia lateral (2-H ₂ O)	6	17	11	5	4	6	7	6	5	6
Ligação de Hidrogênio cadeia lateral – cadeia lateral (2-H ₂ O)	0	8	8	6	0	6	7	0	4	0
Aromática	0	0	0	0	0	0	0	0	0	0
Ligação dissulfeto	0	0	0	0	1	0	0	0	0	0
Tipo de Interação	L	K	M	F	P	S	T	W	Y	V
Hidrofóbica	19	18	14	30	17	21	13	28	27	14
Carregada atrativa	0	22	0	0	0	0	0	0	0	0
Carregada repulsiva	0	22	0	0	0	0	0	0	0	0
Ligação de hidrogênio cadeia principal – cadeia principal	6	6	6	6	5	7	6	6	6	6
Ligação de Hidrogênio cadeia principal – cadeia lateral	3	5	4	3	3	5	5	4	5	3
Ligação de Hidrogênio cadeia lateral – cadeia lateral	0	4	0	0	0	4	4	2	4	0
Ligação de Hidrogênio cadeia principal – cadeia principal (1-H ₂ O)	5	5	4	5	4	4	5	3	4	5
Ligação de Hidrogênio cadeia principal – cadeia lateral (1-H ₂ O)	4	6	4	4	4	7	5	4	6	5
Ligação de Hidrogênio cadeia lateral – cadeia lateral (1-H ₂ O)	0	4	0	0	0	6	5	4	5	0
Ligação de hidrogênio cadeia principal – cadeia principal (2-H ₂ O)	5	5	5	6	4	5	4	5	5	4
Ligação de Hidrogênio cadeia principal – cadeia lateral (2-H ₂ O)	7	8	6	4	6	10	9	5	10	5
Ligação de Hidrogênio cadeia lateral – cadeia lateral (2-H ₂ O)	0	5	0	0	0	7	9	6	7	0
Aromática	0	0	0	0	0	0	0	0	0	0
Ligação dissulfeto	0	0	0	0	0	0	0	0	0	0

Fonte http://www.cbi.cnptia.embrapa.br/DB/unused_contacts/maxcon.txt

4.1.8 Densidade

A densidade local de cada aminoácido é calculada utilizando uma abordagem de sonda esférica, assim como para o descritor de *densidade de energia de contatos* (seção 4.1.3). Para cada resíduo, uma sonda esférica de raio r é centrada no carbono- α ou no LHA. As massas dos átomos internos à sonda esférica são somadas e divididas pelo volume da sonda esférica. Utilizando diferentes raios para a sonda esférica (de 3 a 7Å, com incremento de 1 Å), centrada em dois átomos diferentes, e distinguindo entre cadeias isoladas e em complexos proteicos, o Blue Star STING mantém pré-calculado em seus bancos de dados 20 variações deste descritor.

Durante a implementação deste descritor na SDL, algumas modificações foram realizadas no cálculo das densidades. Átomos dos resíduos de aminoácidos são considerados esferas rígidas de raios iguais ao de van der Waals, podendo se sobrepor ou estarem parcialmente inseridos na esfera sonda. Para evitar uma sobre-estimativa da massa encontrada dentro das esferas sonda, utiliza-se somente a porção não sobreposta ou parcialmente inserida na esfera sonda durante o cálculo. A porção da massa a ser considerada corresponde à diferença entre as massas total (m) e de sobreposição (m_{\cap}), sendo esta última proporcional à razão entre o volume de intersecção (V_{\cap}) de duas esferas de massa uniforme e raios R e r (Equação 4-4) e o volume total do átomo considerado (Equação 4-5).

$$V_{\cap} = \frac{\pi(R + r - d)^2(d^2 + 2dr - 3r^2 + 2dR + 6rR - 3R^2)}{12d} \quad \text{Equação 4-4}$$

$$m_{\cap} = \frac{V_{\cap}}{V} m \quad \text{Equação 4-5}$$

Ainda que não seja realístico o cálculo da densidade considerando estas modificações, é mais preciso e menos sensível às sobreposições, aumentando o poder de descrição do agrupamento (*packing*), submetido aos nanoambientes de diferentes resíduos de aminoácidos da estrutura proteica.

No entanto, átomos localizados na superfície da molécula têm seus valores de densidade ligeiramente menores que aqueles presentes no interior da molécula. Ao centrar a esfera sonda em um átomo na superfície, a porção do volume da esfera fora das fronteiras da molécula (superfície molecular) não contém átomos (massa), porém, mesmo assim todo o volume da esfera é considerado. Isto cria um viés para os átomos da superfície, pois estes terão uma tendência em apresentar valores de densidade menores que aqueles localizados no interior da molécula, devido a inclusão do espaço externo à proteína.

4.1.9 Esponjicidade

Seguindo a mesma abordagem utilizada para cálculo do descritor de *Densidade* (seção 4.1.8), mas ao invés de se somar a massa total ou parcial dos átomos internos à sonda esférica, soma-se o volume ocupado por cada átomo (utilizando o raio de van der Waals e desconsiderando-se o volume de sobreposição calculado através da Equação 4-4). Esse volume é então subtraído e normalizado pelo volume da sonda esférica, resultando em uma medida do espaço vazio no nanoambiente de cada resíduo. Sendo semelhante ao cálculo da densidade, a esponjicidade também introduz o mesmo viés para aqueles átomos localizados na superfície da molécula.

Da mesma forma que o descritor de densidade, o Blue Star STING possui pré-calculados 20 tipos de variações para a *Esponjicidade*, sendo utilizados esfera sondas de raios de 3 a 7Å centradas nos carbonos- α e LHA de cada resíduo. São ainda consideradas cadeias isoladas e em complexo.

4.1.10 Distâncias

Para cada resíduo de aminoácido de uma proteína, são calculados três descritores de distância. Em cada cadeia proteica, tem-se definidos os aminoácidos N-terminal e C-terminal, que correspondem as duas extremidades da sequência proteica. As distâncias euclidianas no espaço tridimensional entre um resíduo qualquer da cadeia proteica e os resíduos nas duas extremidades da sequência são calculadas, e esses dois valores são tomados como descritores. Utiliza-se também a distância entre o resíduo de aminoácido e o centro de massa da cadeia proteica.

4.1.11 Cavidades e Pockets

Originalmente o Blue Star STING utiliza o algoritmo *Pocket* (Edelsbrunner & Koehl, 2003) (disponível em: <http://csb.stanford.edu/~koehl/ProShape/>) para detecção de cavidades e *pockets* nas superfícies das moléculas. No entanto, muitos outros algoritmos existem e estes diferem quanto aos *pockets* e cavidades detectadas. Por esta razão, recentemente foi adicionado ao Blue Star STING outros dois algoritmos para detecção: *NanoShaper* (Decherchi & Rocchia, 2013) e *FPocket* (Guilloux, et al., 2009), sendo o *FPocket* considerado um dos métodos mais precisos na detecção de *pockets* (Volkamer, et al., 2010).

Utilizando três algoritmos diferentes, o Blue Star STING é capaz de detectar átomos de resíduos de aminoácidos que estão presentes em diferentes cavidades das proteínas, além de informações descrevendo propriedades das próprias cavidades, tais como área e volume ocupados. O algoritmo *FPocket* define ainda diversas outras propriedades e pontuações que podem ser utilizadas para descrever os *pockets* encontrados com maior detalhamento, como proporção entre átomos polares e apolares,

drogabilidade, flexibilidade, entre outras (ainda não disponível no ¹PD).

4.1.12 Curvatura

Proteínas e outras macromoléculas não possuem superfícies totalmente planas, sendo estas compostas também por regiões protuberantes (convexas) e invaginadas (côncavas). A presença destas regiões é, em especial, bastante frequente em enzimas, conferindo uma especificidade devido ao tamanho dos ligantes que podem realizar interações com estas regiões da superfície da molécula.

O Blue Star STING utiliza do programa *Surface Racer* (Tsodikov, et al., 2002) para cálculo da curvatura de cada ponto na superfície proteica, medindo o desvio da área local da superfície em relação a um plano. Regiões que interagem em duas cadeias diferentes de um complexo de proteínas devem ter curvaturas complementares, de forma a favorecer a interação entre elas, como ilustrado na Figura 4-3.

A curvatura é definida em nível atômico, ou seja, a cada átomo da proteína é atribuído um valor de curvatura correspondente à região onde este se localiza na superfície da molécula, sendo atribuído um valor negativo àqueles átomos em regiões côncavas, positivo para átomos em regiões convexas e o valor zero aos átomos enterrados (presentes no interior da proteína). Para atribuir valores aos resíduos de aminoácidos de cada proteína, o Blue Star STING realiza o cálculo da curvatura média considerando-se apenas os resíduos na superfície da proteína (curvatura > 0). Considera-se ainda cada cadeia isoladamente e em complexo, sendo que cada resíduo possui dois valores de curvatura.

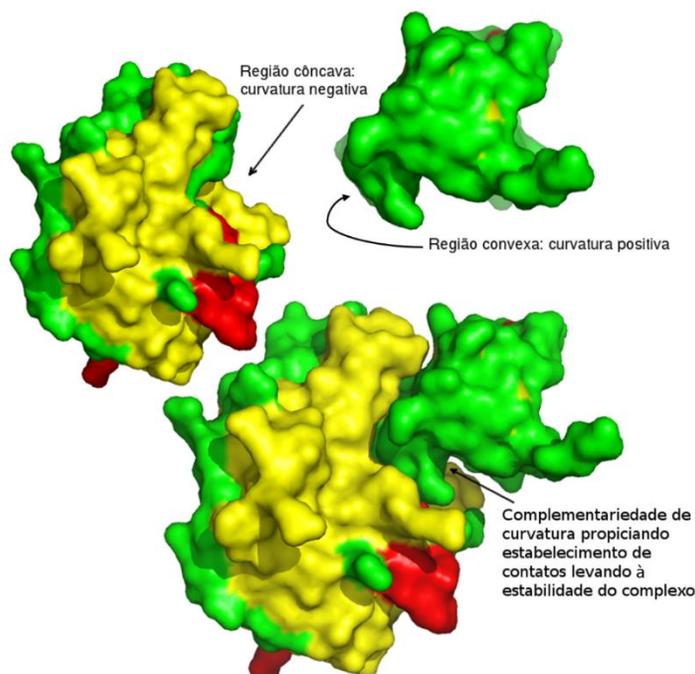


Figura 4-3: Complexo formado pela proteína elastase de leucócitos humanos e o terceiro domínio do inibidor de ovomucóide de *Meleagris gallopavo* (código PDB: 1PPF). Complementariedade de curvatura entre cadeias proteicas propicia uma região com maior superfície de contato entre as moléculas, aumentando a estabilidade do complexo

4.1.13 Patches Hidrofóbicos

Diferentes regiões da superfície das proteínas apresentam diferentes características físico-químicas. Dessa forma, estas regiões podem ser fragmentadas (patches) em regiões majoritariamente hidrofóbicas, polares, apolares, etc. O descritor de patches hidrofóbicos emprega a técnica desenvolvida por LIJNZAAD et al. (1996) para encontrar tais fragmentos na superfície das proteínas, que configuram-se tipicamente como hidrofóbicas. Tais regiões são também denominadas de *hot-spots*, devido à sua importância nas interações entre macromoléculas, uma vez que estas regiões não estabelecem ligações com a moléculas de água, sendo predominante locais de alta energia de contatos entre a proteína e seu ligante e proteína-proteína.

4.1.14 Ligand Pocket e Water Contact Residues – LPR e WCR

Proteínas podem interagir com outras moléculas de diferentes naturezas, como por exemplo moléculas de água. No Blue Star STING, os descritores *Ligand Pocket Residue* (LPR) e *Water Contact Residue* (WCR) descrevem tais interações. Resíduos de aminoácidos que encontram-se próximos de átomos de um ligante, considerando um valor máximo entre suas distâncias, são considerados como LPR's, enquanto resíduos de aminoácidos que estão próximos de moléculas de água cocrystalizadas com a estrutura são considerados WCR's. O Blue Star STING define uma distância máxima para os dois casos de 4Å. Porém, a SDL permite estabelecer outros valores para as distâncias máximas destes descritores.

4.1.15 Rotâmeros

Os átomos dos resíduos de aminoácidos se organizam conformacionalmente formando ângulos entre si. Para os átomos da cadeia principal (nitrogênio, carbono- α , carbono-carbonila e oxigênio) há dois ângulos: PHI (ϕ) e PSI (ψ). Para os átomos da cadeia lateral, podem haver até cinco ângulos, χ -1 até χ -5 (CHI). Uma vez que as cadeias laterais dos aminoácidos possuem comprimentos diferentes, nem todos possuem os mesmos grupos de ângulos. Por exemplo, o ângulo χ -5 é calculado apenas para resíduos de aminoácidos Arginina, enquanto que para resíduos de aminoácido Glicina não há nenhum dos ângulos χ , uma vez que sua cadeia lateral é um átomo de hidrogênio apenas. A Figura 4-4 representa os ângulos diedrais para o aminoácido Arginina.

Assim como os ângulos ϕ e ψ , que apresentam preferências de configuração devido a possíveis choques estéricos entre os aminoácidos, os ângulos da cadeia lateral (χ) também apresentam algumas preferências, podendo apresentar um padrão de orientações (Schrauber, et al., 1993). Baseado nas orientações das cadeias laterais via análise dos ângulos diedrais (χ), rotâmeros são definidos como aquelas conformações que possuem baixa energia. Assim, com base nos valores apresentados por estes

ângulos para um resíduo de aminoácido qualquer da proteína, é possível estabelecer uma classificação para a conformação apresentada pela sua cadeia lateral comparando-se com valores pré-calculados disponíveis na literatura (Lovell, et al., 2000; Shapovalov & Dunbrack Jr., 2011).

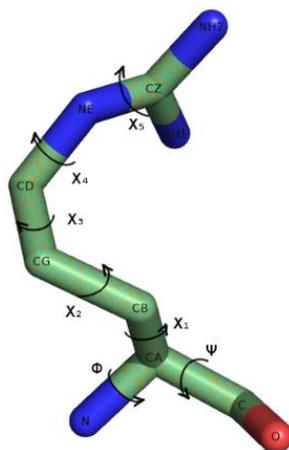


Figura 4-4: Ângulos diedrais da cadeia principal e lateral do aminoácido Arginina (único que se estende até χ_5).

A partir dos diferentes valores dos ângulos ϕ e ψ , é possível construir diagramas denominados gráficos de Ramachandran, em homenagem ao bioquímico *Gopalasamudram Narayana Ramachandran*, que inicialmente propôs seu uso. Estes gráficos possibilitam a análise das conformações diedrais de todos os resíduos de aminoácidos presentes na estrutura e definem regiões consideradas permitidas e proibidas, de acordo com efeitos estéricos para cada tipo de aminoácido. Assim, através de gráficos de Ramachandran, é possível avaliar a qualidade dos modelos gerados e também obter uma aproximação das estruturas secundárias, uma vez que certos tipos de estruturas secundárias possuem preferências de pares dos ângulos (ϕ , ψ). O Blue Star STING possibilita visualização de gráficos de Ramachandran utilizando a definição proposta por LOVELL et al. (2003).

4.1.16 Contatos proteína-ligante

Outro descritor de contatos entre resíduos de aminoácido e o ligante são calculados utilizando-se o método proposto por SOBOLEV et al. (1999). Neste método, cada átomo da proteína e do ligante recebe uma classificação baseada no tipo do átomo (C, O, H, etc.) e das ligações químicas estabelecidas (covalente, dupla, tripla, etc.). Os pares de átomos entre resíduos de aminoácidos e do ligante, que estão suficientemente próximos entre si, são rotulados de acordo com as classes atômicas dos átomos participantes da interação em seis diferentes tipos de contatos: ligação de hidrogênio (Hb), hidrofóbico (Ph), envolvendo átomos em anéis aromáticos (Ar), hidrofílico-hidrofóbico (HH), entre aceitadores (AA) e entre doadores de elétrons (DD), como é mostrado na Tabela 4-6.

Tabela 4-6: Tipos de contatos entre átomos de diferentes classes para o cálculo do descritor de contatos proteína-ligante. Hb: ligação de hidrogênio; Ph: contato hidrofóbico; Ar: contato aromático-aromático; HH: contato hidrofílico-hidrofóbico; AA: contato aceitador-aceitador; DD: contato doador-doador; -: outro contato.

CLASSE ATÔMICA	I	II	III	IV	V	VI	VII	VIII
I - HIDROFÍLICO	Hb	Hb	Hb	HH	-	-	-	-
II - ACEITADOR	Hb	AA	Hb	HH	-	-	-	-
III - DOADOR	Hb	Hb	DD	HH	-	-	-	-
IV – HIDROFÓBICO	HH	HH	HH	Ph	Ph	Ph	Ph	Ph
V – AROMÁTICO	-	-	-	Ph	Ar	-	-	-
VI – NEUTRO	-	-	-	Ph	-	-	-	-
VII – NEUTRO-DOADOR	-	-	-	Ph	-	-	-	-
VIII – NEUTRO-ACEITADOR	-	-	-	Ph	-	-	-	-

4.1.17 Estrutura secundária

Conformações tridimensionais locais dos resíduos de aminoácidos recebem diferentes denominações, de acordo com as ligações de hidrogênio estabelecidas pelos resíduos de aminoácidos nestas regiões, criando diferentes conformação estruturais. Diferentemente da estrutura terciária, a estrutura secundária não define posições tridimensionais específicas dos resíduos de aminoácidos, mas sim conformações locais de um grupo de resíduos que interagem entre si para formar diferentes combinações estruturais. Entre as mais encontradas, têm-se hélices alfa, folhas beta e laços (*loops*).

No Blue Star STING, existem três diferentes atribuições de estrutura secundária para um resíduo de aminoácido, através de informações contidas nos próprios arquivos PDB (quando disponível) e as calculadas pelos *softwares* DSSP (Kabsch & Sander, 1983; Joosten, et al., 2010) e STRIDE (Frishman & Argos, 1995). Muitas vezes, os tipos de estruturas secundárias, bem como os tamanhos e resíduos iniciais e finais, divergem entre as diferentes classificações. Sendo assim, com a presença destas três fontes distintas, é possível obter um consenso e encontrar regiões com estruturas secundárias mais confiáveis.

4.1.18 Energia de solvatação

A energia de solvatação, corresponde à energia entre ligações de átomo do soluto e do solvente. No caso do solvente ser água, esta também é chamada de energia livre de hidratação. A partir da área relativa acessível ao solvente de cada átomo, é possível calcular aproximadamente a energia livre de hidratação presente nas interações entre tais átomos e as moléculas de água (solvente) (Ooi, et al., 1987). A energia livre de hidratação do *i*-ésimo átomo de um resíduo de aminoácido é calculada como sendo o produto entre seu parâmetro de solvatação atômico (g_i), determinado experimentalmente, e a área relativa acessível ao solvente ($ASA_{relativa}$). O somatório das energias livre de hidratação dos átomos que compõem

o resíduo r e os átomos na vizinhança de r , normalizado pela soma das áreas relativas acessíveis ao solvente de todos os átomos considerados, fornece a energia de solvatação do resíduo r (G_r) na forma:

$$G_r = \frac{\sum_i g_i ASA_{relativa}}{\sum_i ASA_{relativa}} \quad \text{Equação 4-6}$$

Considerando-se valores de solvatação atômicos experimentais obtidos para seis diferentes temperaturas, cada resíduo possui um total de seis descritores de solvatação. O Blue Star STING calcula e armazena descritores de solvatação utilizando raios de 3 a 9Å, totalizando 42 descritores por resíduo de aminoácido.

4.1.19 Descritores baseados em grafos

Cadeias proteicas podem ser representadas como grafos não direcionados onde o conjunto de vértices é composto pelos resíduos de aminoácidos ou os átomos de uma cadeia proteica, enquanto que as arestas do grafo representam interações entre estes resíduos ou átomos. No Blue Star STING, através da SDL utilizou-se como conjunto de vértices os resíduos de aminoácidos da proteína, sendo o conjunto de arestas definido de duas formas: utilizando-se contatos interatômicos previamente calculados ou considerando-se resíduos de aminoácidos que encontram-se próximos, respeitando uma distância máxima pré-estabelecida entre seus carbonos- α .

A partir de um grafo, é possível obter diversas métricas e medidas que extraem e descrevem o comportamento das cadeias proteicas como redes de interações entre resíduos de aminoácidos. Representando propriedades geométricas e topológicas das cadeias proteicas, estas métricas podem ser consideradas como descritores estruturais de proteínas, uma vez que os grafos são construídos baseando-se em informações estruturais das cadeias proteicas. Atualmente 17 descritores são extraídos de grafos disponíveis para cálculo utilizando-se a SDL:

- Centralidade por proximidade (*closeness*): medida do quanto um resíduo está próximo dos demais. Este descritor mede o quão central um resíduo é na cadeia proteica, sendo os resíduos centrais aqueles presentes num maior número de caminhos (sequência de vértices) entre quaisquer dois resíduos de aminoácidos no grafo. Esta medida fornece informações de quão próximo um resíduo encontra-se do centro geométrico da cadeia proteica. Assim, resíduos de aminoácidos localizados em cavidades na superfície da molécula terão valores mais elevados de centralidade por proximidade do que a média dos resíduos da superfície (Figura 4-5a). A centralidade de proximidade é calculada como média da distância geodésica (menor

distância) entre um resíduo e os demais resíduos da mesma cadeia proteica, conforme a Equação 4-7:

$$Closeness_i = \frac{n - 1}{\sum_j d_{ij}} \quad \text{Equação 4-7}$$

onde d_{ij} é a distância geodésica entre os resíduos i e j .

- Excentricidade: de um resíduo de aminoácido v representa a distância de v ao resíduo mais distante deste no grafo, conforme mostrado na Equação 4-8. Resíduos centrais são os resíduos com valores de excentricidade mínimos.

$$ecc_v = \max_{u \in V} \{d_{v,u}\} \quad \text{Equação 4-8}$$

onde V é o conjunto de todos os vértices e $d_{v,u}$ é a distância geodésica entre os vértices v e u .

- Centralidade por proximidade local (*local closeness*): similar à centralidade de proximidade, a centralidade por proximidade local considera apenas um subconjunto dos resíduos de aminoácidos, distantes de no máximo m arestas, para calcular a centralidade de um resíduo de aminoácido (Equação 4-9), ao invés de todos os vértices do grafo. Trata-se de uma medida de centralidade local, em contraposição a uma medida global obtida pelo descritor de centralidade por proximidade (Mitternacht & Berezovsky, 2011). Por ser uma métrica local, a centralidade por proximidade local fornece uma medida da profundidade que um resíduo se encontra na superfície da molécula. Assim, resíduos de aminoácido localizados em cavidades mais profundas terão maiores valores para a centralidade local (Figura 4-5b).

$$LC_m = \sum_{k=1}^m \frac{n_k}{k^2} \quad \text{Equação 4-9}$$

onde n_k é o número de vértices cuja distância mais curta ao vértice de interesse é exatamente k .

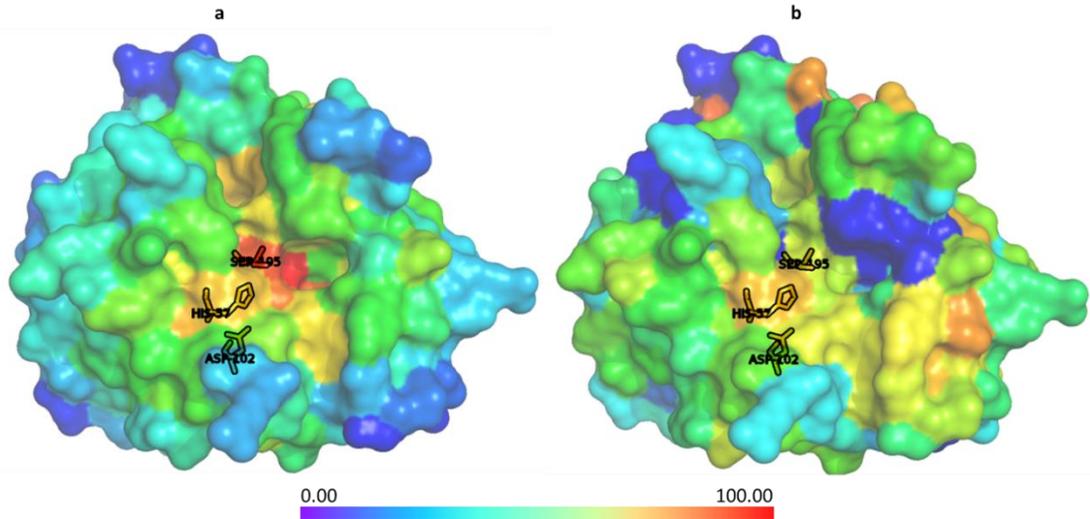


Figura 4-5: (a) Centralidade por proximidade (*closeness*) para enzima Tripsina de *Salmo salar* (PDB: 1A0J cadeia A), com triade catalítica destacada em contorno preto (HIS57, ASP102, e SER195). Resíduos de aminoácido catalíticos e outros resíduos próximos ao centro geométrico da cadeia apresentam valores mais elevados de *closeness*. (b) Centralidade de proximidade local para enzima Tripsina de *Salmo salar* (PDB: 1A0J cadeia A). A centralidade local considera a vizinhança de cada resíduo de aminoácido, sendo indiferente à sua localização em relação ao centro de massa da proteína.

- Centralidade radial: similar a centralidade por proximidade. No entanto, atribui valores mais altos de centralidade aos vértices que estão mais próximos de todos os outros vértices no grafo, com relação ao diâmetro do grafo (maior distância entre dois vértices do grafo), conforme a equação:

$$RC_i = \frac{\sum_j (d - d_{ij} + 1)}{(n - 1)d} \quad \text{Equação 4-10}$$

onde d é o diâmetro do grafo, d_{ij} a distância entre os vértices i e j e n é o número de vértices no grafo.

- Conectividade local média (LAC): este descritor avalia a conectividade local de resíduo de aminoácido baseada na média das conectividades de seus vizinhos. Quando um resíduo e sua vizinhança possuem um alto número de conexões (arestas/contatos), este resíduo e seus vizinhos possuirão valores médios de conectividade local mais elevados do que vértices em regiões com poucas arestas em suas vizinhanças (Li, et al., 2011). Este descritor assemelha-se ao descritor de densidade de energia de contatos, mas o LAC não considera as energias dos contatos, somente os números de interações realizadas pelo par de resíduos, como mostra a Equação 4-11.

$$LAC_v = \frac{\sum_{w \in N_v} deg^{C_v}(w)}{|N_v|} \quad \text{Equação 4-11}$$

onde N_v é o conjunto dos vértices vizinhos ao vértice v , e $deg^{C_v}(w)$ é a conectividade local

(número de arestas) do vértice w no subgrafo induzido a partir dos vizinhos de v : C_v .

- Máxima componente de vizinhança (MNC) e densidade da máxima componente de vizinhança (DMNC): são descritores que utilizam cálculo de componentes conexas nos grafos para a obtenção de descritores. Enquanto MNC (*Maximum Neighborhood Component*) de um vértice v é definido como o tamanho da máxima componente no grafo induzido a partir dos vizinhos de v , DMNC de um vértice v é a razão entre o número de arestas na máxima componente na vizinhança de v (MNC_v) pelo número de vértices na mesma componente, elevado a um fator de regularização da rede, dado por E/N^ϵ , onde $1 \leq \epsilon \leq 2$ (Lin, et al., 2008). Componentes conexas em grafos não direcionados são definidas como subgrafos nos quais existe pelo menos um caminho entre quaisquer dois vértices do subgrafo induzido a partir do grafo original.
- Coeficiente de agrupamento (*cluster coefficient* ou CC): de um resíduo v representa a fração dos vizinhos de v que também são vizinhos entre si. Vértices com altos valores de CC possuem um agrupamento maior entre seus vizinhos, de forma que estas regiões podem ser consideradas regiões de alta concentração de ligações interatômicas.
- Grau ou cardinalidade (*degree*): é o descritor mais simples, correspondendo ao número de arestas conectadas ao vértice v , i.e. número de contatos.
- *Betweenness centrality*: número de caminhos mais curtos entre todos os pares de vértices do grafo que passam pelo vértice v . Este descritor mede a importância de um vértice de acordo com sua localização entre os caminhos mais curtos no grafo.
- Gargalo (*bottleneck*): para cada vértice v no grafo, uma árvore de caminhos mínimos é construída tendo v como raiz: T_v . A partir da árvore, define-se o peso de um vértice w como sendo igual ao número de descendentes de w na árvore T_v . Assim, um vértice é considerado como um gargalo se seu peso não for menor que $n/4$, onde n é o número de nós na árvore. O valor de gargalo de um vértice w é então igual ao número de vértices v no qual w é um gargalo na árvore T_v (Yu, et al., 2007; Przulj, et al., 2004).
- Centralidade Alfa: medida de centralidade onde os vértices são ponderados devido a fontes “externas”, como apresentado na Equação 4-12. De acordo com a definição é possível controlar a influência externa ajustando o parâmetro α na Equação 4-12. Quando $\alpha = 0$, somente as importâncias externas são consideradas no cálculo da centralidade, e onde α possui valores altos maior peso é dado à conectividade dos vértices do grafo. Utilizou-se como importâncias externas de cada resíduo de aminoácido a distância euclidiana, no espaço tridimensional, deste ao centro de massa da cadeia considerada.

$$\alpha C_v = (I - \alpha A^T)^{-1} e \quad \text{Equação 4-12}$$

onde A é matriz de adjacências para o grafo, I a matriz identidade, e é o vetor com as importâncias externas de cada vértice do grafo e α é o parâmetro de ponderação.

- Grau médio da vizinhança: de um resíduo v é a média dos graus (*degree*) dos vizinhos de v .
- Baricentro: soma normalizada de todas as distâncias entre um vértice v e todos os outros

vértices do grafo. Em uma proteína os resíduos de aminoácidos (vértices) que encontram-se em regiões “centrais” da molécula terão valores de baricentro menores que aqueles localizados em regiões periféricas da molécula, uma vez que, resíduos de aminoácidos “centrais” estarão a uma distância média à todos outros resíduos inferior àqueles em regiões mais distantes do centro de massa da proteína.

- *Random Walk Betweenness*: similar ao descritor *betweenness*, porém utiliza caminhos aleatórios. Mede o número esperado de vezes que um nó é visitado por caminhos aleatórios entre todos os pares de vértices.

4.2 Descritores de conservação de estrutura primária

4.2.1 Entropia e entropia relativa

A partir de um conjunto de sequências proteicas, é possível realizar alinhamentos múltiplos de sequência (*Multiple Sequence Alignment* ou simplesmente MSA) de forma que seja possível extrair informações sobre a importância de certos aminoácidos presentes nas sequências. Se numa dada posição do alinhamento múltiplo um mesmo tipo de aminoácido aparece em todas as sequências alinhadas, há um grande indício de sua importância para a proteína, justificando sua manutenção. Aminoácidos presentes em uma quantidade significativa de sequências na mesma posição do alinhamento são ditos aminoácidos conservados e, por isso, descritores de proteína que utilizam alinhamentos múltiplos são denominados descritores de conservação de sequência ou estrutura primária.

Dada uma proteína alvo, é possível encontrar valores de entropia e entropia relativa para cada um de seus aminoácidos, através do alinhamento desta proteína com outras sequências proteicas contidas num banco de dados de sequências, como por exemplo o Uniprot (The UniProt Consortium, 2014). A partir do alinhamento múltiplo, é calculado para cada aminoácido da sequência alvo um valor de entropia definindo o grau de “desordem” da respectiva coluna no alinhamento múltiplo. Se esta desordem for baixa, este aminoácido provavelmente possui uma importância alta para a proteína, como por exemplo um aminoácido catalítico. Caso contrário, se a desordem for alta, há grandes chances deste aminoácido não ter um grande efeito sobre a proteína e, por isso, foi modificado em outras sequências homólogas sem prejuízo para a funcionalidade da proteína.

No Blue Star STING, encontram-se presentes descritores de conservação derivados de duas fontes de dados. Além dos valores diretamente extraídos dos arquivos presentes na base de dados HSSP (Joosten, et al., 2010), descritores de conservação são gerados pelo próprio Blue Star STING (Higa, et al., 2006).

A partir dos arquivos do HSSP, são extraídos para cada resíduo de aminoácido de uma proteína valores de entropia, entropia relativa, e frequência de cada um dos 20 tipos de aminoácido na mesma

posição no MSA. Os mesmos descritores são calculados pelo Blue Star STING utilizando MSA's gerados localmente, com auxílio das ferramentas BLAST (Altschul, et al., 1990) e CLUSTAL OMEGA (Sievers, et al., 2011). A entropia associada a um resíduo pode ser obtida pela Equação 4-13, enquanto que a entropia relativa pode ser obtida a partir da entropia como mostrado na Equação 4-14. Tanto para os alinhamentos múltiplos extraídos dos arquivos HSSP, quanto para os MSA's locais, é calculada a confiança (*reliability*) da informação, como mostrado na Equação 4-15: quanto maior o número de sequências alinhadas, maior será a confiança dos valores de entropia e entropia relativa e, assim, maior será a confiança na inferência sobre a importância evolutiva de um resíduo.

$$S_i = - \sum_{R_i}^{20} f_{R_i} \ln f_{R_i} \quad \text{Equação 4-13}$$

onde f_{R_i} é a frequência do aminoácido R na posição i do alinhamento, e S_i é a entropia do aminoácido na posição i .

$$relent_i = \frac{S_i}{\ln 20} \quad \text{Equação 4-14}$$

$$reliability_i = \begin{cases} 0 & \text{se } ncc_i \leq 5 \\ ncc_i / |MSA| & \end{cases} \quad \text{Equação 4-15}$$

onde ncc_i é a quantidade de sequências em que foi encontrado um aminoácido na mesma posição (*match*) e $|MSA|$ é o número de sequências alinhadas. Se $ncc_i = |MSA|$, então a confiança será de 100% na informação para o resíduo na posição i .

Além destes descritores, foi incorporado à SDL outro descritor de conservação, considerando as frequências naturais dos aminoácidos obtidas através da ferramenta ProtScale (Gasteiger, et al., 2005), no cálculo da entropia relativa. Na entropia relativa disponível no HSSP e pelo próprio Blue Star STING, todos os aminoácidos recebem a mesma probabilidade de serem encontrados numa certa posição da sequência. Por esta razão, na Equação 4-14 entropia é dividida por $\ln 20$ para encontrar a entropia relativa que, por definição, segue a Equação 4-16 mostrada abaixo. Assim, utilizando as frequências naturais dos aminoácidos para cálculo da entropia relativa (denominada de entropia relativa de *background*) procura-se melhorar a precisão das informações de conservação dos resíduos de aminoácidos (Wang & Samudrala, 2006).

$$backgroundRelent_i = \sum_R^{20} f_{R_i} \ln \frac{f_{R_i}}{f_{R_b}} \quad \text{Equação 4-16}$$

onde f_{R_i} é a frequência do aminoácido R na posição i do alinhamento e f_{R_b} é a frequência natural ou de *background* do aminoácido R , extraída a partir da ferramenta ProtScale (Gasteiger, et al., 2005).

4.2.2 Densidade de entropia 3D

Da mesma forma que os descritores de Densidade de Energia de Contatos discutidos na seção 4.1.3, o descritor de Densidade de entropia emprega uma esfera sonda de raio r , centrada nos carbonos- α e LHA de cada resíduo da proteína, e calcula a densidade de entropia dentro da esfera somando-se a entropia relativa dos resíduos dentro da esfera e dividindo-se pelo volume da esfera sonda. Dessa forma, busca-se captar não somente resíduos conservados através da entropia relativa de cada resíduo, mas também nanoambientes conservados através da densidade de entropia relativa dentro das esferas sondas.

Originalmente, este descritor utilizava somente valores de entropia relativa provenientes dos arquivos HSSP. Durante a implementação da SDL, foi incorporado também o cálculo dos descritores de entropia 3D utilizando a entropia relativa calculada pelo próprio Blue Star STING, bem como adicionado o descritor de Densidade de entropia relativa de *background*, calculado utilizando a entropia relativa de *background* discutida na seção anterior. Assim como os descritores de densidade e esponjicidade, o descritor de densidade de entropia 3D também possui viés em relação àqueles resíduos de aminoácidos localizados na superfície da molécula. Sendo que esses resíduos terão tendência em possuir valores de densidade de entropia inferiores aos resíduos no interior da proteína.

O Blue Star STING possui valores de densidade de entropia pré-calculados para esferas de raios de 3 a 7Å, com incremento de 1Å.

4.2.3 Pressão evolutiva

Diferentes aminoácidos em uma proteína estão sujeitos a diferentes taxas de variação, ou biologicamente falando, diferentes pressões evolutivas. Um aminoácido em uma posição da sequência que evolui lentamente é dito conservado, enquanto que aqueles em posições que evoluem rapidamente são ditos “variáveis”. Assim, é possível calcular a taxa evolutiva relativa de cada resíduo de aminoácido da proteína.

O Blue Star STING utiliza o programa *Rate4Site* (Mayrose, et al., 2004) para o cálculo da taxa evolutiva relativa, fornecendo como entrada do programa os alinhamentos múltiplos de sequências obtidos dos arquivos HSSP e os gerados localmente pelo Blue Star STING.

4.3 Descritores de vizinhança dos aminoácidos

Os descritores presentes no STING_DB são calculados levando-se em consideração apenas propriedades individuais dos aminoácidos, ou uma pequena vizinhança definida pelas esferas sondas. Porém, ao considerar o estudo de regiões específicas das proteínas, as propriedades dos resíduos de aminoácido nestas regiões tomadas em conjunto podem trazer ganhos expressivos. Por esta razão, foram adicionados ao Blue Star STING quatro diferentes descritores de vizinhança utilizando diferentes abordagens e definições para a vizinhança de um resíduo de aminoácido: descritores ponderados pela área relativa acessível ao solvente, distâncias espaciais, “forças de contatos”, *sliding window* e arestas no grafo representando a proteína

4.3.1 Descritores Ponderados pela Acessibilidade Relativa e Distâncias espaciais

Inspirado pelo trabalho de POROLLO & MELLER (2007), foram criados e incorporados à SDL os descritores denominados *Weighted Neighbor Average* ou simplesmente WNA. Estes descritores definem dois valores para cada descritor originalmente no Blue Star STING. O primeiro deles utiliza valores da acessibilidade relativa dos resíduos para ponderar os descritores (*WNASurface*), e o segundo utiliza o inverso da distância entre o resíduo central e seus vizinhos (*WNA Distance*). Para definição de vizinhança, utilizou-se uma esfera de raio 15Å centrada no carbono- α de um resíduo de aminoácido. Da forma como proposto originalmente no artigo de POROLLO & MELLER (2007) o raio da esfera é mantido fixo em 15Å. Porém, a partir da SDL, é possível utilizar outros valores de raio e criar diversos descritores com diferentes tamanhos de vizinhança. No caso dos descritores WNA que utilizam a acessibilidade relativa como fator de ponderação, somente são considerados os resíduos com área relativa acessível ao solvente maior que 5%, ou seja, resíduos que não estão totalmente no interior da proteína.

Devido a diferenças na densidade nas vizinhanças dos resíduos de aminoácidos, optou-se por normalizar os valores destes descritores pela densidade atômica das esferas sondas (vizinhanças), como mostrado na Equação 4-17 e na Equação 4-18.

$$WNA_{distance_r} = \left(D_r + \sum_{i \in viz(r)} \frac{D_i}{d_{ri}} \right) / \rho_r \quad \text{Equação 4-17}$$

onde D_r é o valor do descritor para o resíduo alvo (r), D_i é o valor do descritor para o resíduo i na vizinhança de r ($viz(r)$), ρ_r é a densidade da vizinhança do resíduo alvo e d_{ri} é distância euclidiana entre os carbonos- α dos resíduos r e seu i -ésimo vizinho.

$$WNA_{surface_r} = \left(\sum_{\{i | i \in viz(r) \wedge rsa_i > 5\% \}} RSA_i * D_i \right) / \rho_r \quad \text{Equação 4-18}$$

onde D_i é o valor do descritor para o resíduo i pertencente ao conjunto formado pelos resíduos na vizinhança de r (incluindo o próprio r) com área relativa acessível ao solvente (RSA_i) superior a 5%, e ρ_r é a densidade da vizinhança do resíduo alvo.

4.3.2 Descritores Ponderados pela “força de contato”

Outra forma de calcular descritores de vizinhança é utilizando diagramas de Voronoi (Segura, et al., 2011). Nesta abordagem, os resíduos considerados vizinhos são aqueles que compartilham uma mesma aresta no diagrama de Voronoi construído a partir dos átomos de cada resíduo. Assim, um par de resíduos é dito em contato se pelo menos um par de átomos pesados de cada resíduo possui uma face em comum. Os autores alertam que esta definição de contato difere da definição clássica, que implica interações atômicas. No entanto, reportam que esta definição traz melhores resultados que outros métodos baseados em distâncias sequenciais ou estruturais entre os resíduos, além de não necessitarem do estabelecimento de um valor limite (*threshold*).

Nos diagramas de Voronoi, alguns resíduos terão um número maior de contatos do que outros, dependendo de suas localizações. Por isso, é calculado um parâmetro chamado de “força de contato” (c_{ij}). Seja r_i um resíduo qualquer da proteína e r_j ($j=1..n$) os vizinhos de r_i , e seja N_{ij} número de arestas compartilhadas ou os pares de contatos entre r_i e r_j . Então, pode-se calcular o número total de pares de contatos para o resíduo de aminoácido r_i conforme segue:

$$N_i = \sum_{j=1}^n N_{ij} \quad \text{Equação 4-19}$$

A partir de N_i , a força de contato c_{ij} entre os resíduos r_i e r_j pode ser calculada de acordo com a Equação 4-20, sendo então utilizada para ponderar a influência dos descritores dos resíduos de aminoácidos vizinhos a r_i (f_{jk}) e fornecer o descritor de vizinhança do resíduo r_i (ef_{ik}), onde k é o descritor a ser considerado, conforme Equação 4-21.

$$c_{ij} = \frac{N_{ij}}{N_i} \quad \text{Equação 4-20}$$

$$ef_{ik} = \sum_{j=1}^n c_{ij} f_{jk} \quad \text{Equação 4-21}$$

4.3.3 Descritores de vizinhança sequenciais (Janelas deslizantes)

No caso de descritores de vizinhança por janelas deslizantes (*sliding window*), utiliza-se uma definição de vizinhança baseada nas estruturas primárias das proteínas. Primeiramente, define-se um tamanho (L , geralmente sendo ímpar) para a janela que irá “deslizar” sobre a estrutura primária da proteína, definindo assim uma vizinhança para o aminoácido central dentro desta janela. Dessa forma, a vizinhança de um aminoácido a_i é definida como sendo os aminoácidos adjacentes dentro da janela de tamanho L : $a_{i-L/2}, \dots, a_{i-1}, a_{i+1}, \dots, a_{i+L/2}$. O descritor de vizinhança é então composto pela média aritmética dos descritores dos resíduos de aminoácidos vizinhos e o descritor do resíduo de aminoácido a_i .

No Blue Star STING, é possível obter descritores de *Sliding Window* utilizando janelas de tamanhos 3, 5, 7 e 9 aminoácidos. Porém utilizando a SDL é possível estipular qualquer tamanho para o cálculo.

4.3.4 Descritores de vizinhança a partir de grafos de contatos

Da mesma forma como foram construídos os grafos para representar cadeias proteicas para a extração dos descritores apresentados na seção 4.1.19, pode-se utilizá-los para definição de uma vizinhança e, então, calcular descritores de vizinhança.

Para isto, define-se um valor máximo para o tamanho da vizinhança (K) que será utilizado. Este tamanho é medido em número de arestas ou tamanho do caminho mínimo entre o resíduo central e seus vizinhos, ou seja, para $K=1$, apenas são considerados vizinhos de um resíduo de aminoácido os vértices imediatamente adjacentes a este (possuem uma aresta ligando ele a seu vizinho). No caso de $K=2$, são considerados os vizinhos adjacentes mais os vizinhos destes vizinhos, e assim sucessivamente para valores maiores de K . Seja portanto r_i um resíduo de aminoácido qualquer e f_i um descritor para este resíduo, então pode-se calcular o descritor de vizinhança (g_{f_i}) a partir do grafo $G(V, E)$, onde V é o conjunto de vértices ou resíduos de aminoácidos e E é o conjunto de arestas ou contatos. Considerando como vizinhos do vértice r_i os resíduos de aminoácido visitados por um caminho mínimo de tamanho máximo igual a K , sendo d_{ij} a distância (número de arestas) entre o resíduo e seu vizinho j , o cálculo de g_{f_i} é realizado considerando-se diferentes pesos para as camadas vizinhas ($k=1, 2, 3, \dots, K$) de r_i como segue na Equação 4-22.

$$g_{f_i} = f_i + \sum_{k=1}^K \frac{\sum_{\forall j | d(i,j) < k} f_j}{k^2} \quad \text{Equação 4-22}$$

Capítulo 5 - Aprendizado de Máquina

Segundo MITCHELL (1997), “um programa de computador é dito aprender a partir de uma experiência E com respeito a uma classe de tarefas T e uma métrica de desempenho P , se o seu desempenho na tarefa T , medido por P , aumenta com a sua experiência E ”. Formalmente, aprendizado de máquina opera em dois subespaços, sendo um o espaço de variáveis (ou dados) X e outro o espaço de possíveis modelos de aprendizado de máquina Θ . A partir de um conjunto de treinamento $T_r \subseteq X$, um algoritmo de aprendizado de máquina seleciona parâmetros de um modelo $\theta \in \Theta$. Diferentes algoritmos utilizam diferentes métricas e critérios para a seleção de um modelo, ou mais de um em alguns casos (*ensembles*).

Algoritmos de aprendizado de máquina podem ser divididos de acordo com o método de aprendizagem utilizado: aprendizado supervisionado, aprendizado não supervisionado ou aprendizado semisupervisionado.

No aprendizado supervisionado, toda entrada é acompanhada de uma saída desejada que o método deve ser capaz de reproduzir (também chamada de rótulo ou classe em problemas de classificação). Assim, dado um conjunto de treinamento T_r composto por pares (\mathbf{x}_i, y_i) , onde $\mathbf{x}_i \in X$ são as entradas para a i -ésima amostra do conjunto de treinamento e $y_i \in Y$ a saída correspondente para esta mesma amostra, o treinamento supervisionado busca, portanto, encontrar uma função $y = f(\mathbf{x}, \theta)$, com menor erro. Geralmente, além de um conjunto de treinamento T_r , trabalha-se com um conjunto de validação $V_a \subseteq X$. Ajusta-se o modelo $\theta \in \Theta$ empregando T_r e o objetivo é minimizar o erro junto a V_a , visando homogeneizar a capacidade de generalização do modelo θ . Predição e classificação são aplicações mais usuais de treinamento supervisionado (Mohri, et al., 2012).

No aprendizado não supervisionado, o conjunto de entrada é composto somente pelos atributos das amostras ($\mathbf{x}_i \in X$), ou seja, a saída desejada é considerada desconhecida a *priori*. Diversas técnicas de aprendizado não supervisionado buscam explicar e resumir as principais características dos dados, como por exemplo *Clustering*, Modelos ocultos de Markov, redução de dimensionalidade dos dados (e.g. análise de componentes principais), redes neurais artificiais (e.g. mapas auto-organizáveis), etc. (Mohri, et al., 2012).

Como um caso intermediário entre as duas abordagens, tem-se o aprendizado semisupervisionado, que é útil quando somente uma pequena parcela dos dados possui rótulos, porém são insuficientes para realizar um aprendizado supervisionado. Assim, utiliza-se, neste caso, informações provenientes dos dois tipos de amostras, as rotuladas e as não rotuladas (Chapelle, et al., 2006).

5.1 Modelos Classificadores

Modelos classificadores podem ser descritos como um mapeamento que, a partir de um conjunto de dados de entrada, retorna ou um valor numérico (e.g. probabilidade) ou simplesmente associa o indivíduo a uma classe (Fawcett, 2006). O treinamento de modelos classificadores usualmente é realizado utilizando-se treinamento supervisionado, embora outros tipos de aprendizado possam também ser utilizados. Assim, procura-se encontrar um modelo que seja capaz de classificar novas amostras para as quais os rótulos são ainda desconhecidos, seja atribuindo um rótulo ou retornando uma probabilidade da amostra pertencer a cada classe do problema.

Um classificador de resíduos de aminoácidos catalíticos é um classificador binário, pois busca, a partir de descritores estruturais de proteínas e um conjunto de amostras como entrada, atribuir um rótulo positivo (catalítico) ou negativo (não catalítico) através da função de mapeamento selecionada a partir dos dados de treinamento. A partir das saídas fornecidas, aplicando-se o modelo classificador ao conjunto de treinamento, ou mais comumente a um conjunto validação (não utilizado no treinamento), podem-se obter quatro resultados possíveis quando há apenas duas classes (catalítico x não catalítico):

- um resíduo catalítico é classificado como tal pelo modelo classificador (verdadeiro positivo, ou TP, da sigla em inglês *true positive*);
- um resíduo não-catalítico é classificado como tal (verdadeiro negativo, ou TN, da sigla em inglês *true negative*);
- um resíduo catalítico é classificado como não catalítico (falso negativo, ou FN, da sigla em inglês *false negative*);
- ou ainda um resíduo não-catalítico é classificado como catalítico (falso positivo, ou FP, da sigla em inglês *false positive*).

Com estes quatro resultados, pode-se construir uma matriz de confusão (também chamada de tabela de contingência) utilizada por qualquer métrica de avaliação de modelos classificadores para mensurar o desempenho de um classificador, como mostra a Figura 5-1.

Diversas metodologias para treinamento e seleção de modelos classificadores podem ser utilizadas e estão disponíveis na literatura. A princípio, não se pode afirmar qual metodologia é superior em classificação. Dependendo do problema a ser resolvido, pode-se ter desempenhos comparativos diferentes entre os modelos classificadores.

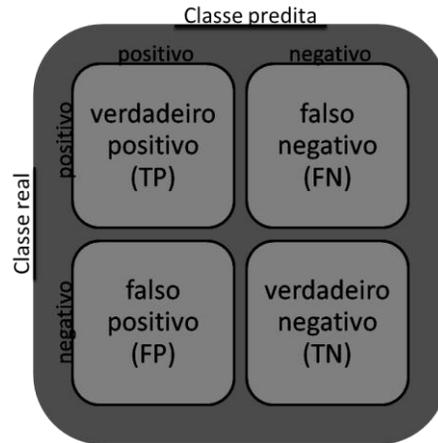


Figura 5-1: Matriz de confusão para classificação binária (duas classes).

De acordo com os objetivos deste trabalho, obter regras que sejam interpretáveis, restringe o uso de métodos classificadores àqueles capazes de construir modelos que realizam o mapeamento entre o conjunto de dados e as classes, utilizando-se de simples comparações aplicadas aos descritores escolhidos para compor as regras. Os métodos classificadores mais utilizados para estes propósitos envolvem árvores de decisão e classificação, e métodos para indução de regras. Apesar de ser possível extrair regras de outros métodos, como redes neurais artificiais (*Artificial Neural Networks* – ANN, em inglês), regras obtidas a partir destes métodos tendem a ser menos parcimoniosas e compreensíveis do que as obtidas por métodos especializados (Andrews, et al., 1995; Tickle, et al., 1998a). Isto traz um maior desafio à classificação, uma vez que, para manter o classificador o mais interpretável possível, não é permitido utilizar métodos que realizem combinações lineares ou não lineares das entradas. Em métodos para a construção de árvores de decisão ou indução de regras, o conjunto de dados é subdividido de acordo com um teste aplicado a um único descritor, de forma que este teste fornece a melhor separação possível entre as classes. Para avaliar a qualidade de um teste, é utilizada uma métrica de impureza, que define o grau de homogeneidade dos subconjuntos gerados pela aplicação do teste, segundo os rótulos das amostras contidas em cada uma desses subconjuntos.

5.2 Desbalanceamento entre classes

Em problemas de classificação, quando uma das classes apresenta um número muito maior de amostras comparado às demais classes, então diz-se que os dados estão desbalanceados. Na classificação envolvendo duas classes, uma classe pode ser majoritária (geralmente a classe negativa) e outra minoritária (geralmente a classe positiva e a classe de interesse). Assim, o número de amostras da classe negativa vai ser superior ao número de amostras da classe positiva. Uma definição para a proporção entre as classes que leva um problema a ser considerado desbalanceado é bastante divergente. A princípio,

qualquer problema em que se tenha uma distribuição desigual entre as classes pode ser considerado desbalanceado. Existem certos problemas em que o desbalanceamento é ainda mais acentuado, de forma que o número de amostras da classe positiva não supera 5% do número de amostras negativas. Este caso extremo, conhecido como classificação de classes raras ou eventos raros, é ainda mais desafiador. WEISS (2004) demonstrou que problemas de mineração de classes raras e dados desbalanceados enfrentam muitos desafios em comum e ambos podem se beneficiar de técnicas similares para a correção das distribuições entre as classes. No entanto, apresentam objetivos um pouco diferentes: enquanto em problemas desbalanceados o foco principal são as distribuições desbalanceadas entre as classes, em problemas de mineração de classes raras a classe minoritária é tida como anormal e requer uma atenção especial.

O desbalanceamento entre classes fornece um grande desafio aos métodos classificadores, pois geralmente tais métodos são construídos utilizando-se o erro médio ou acurácia, de forma que tomando um conjunto de dados onde 99% das amostras pertencem à classe negativa e somente 1% pertencem à classe positiva, e realizando a predição de todas as amostras como negativas, independente da entrada, irá produzir um desempenho de 99% de acertos. Por isso, ainda que junto a problemas em que o desbalanceamento entre classes não seja muito alto seja possível utilizar a acurácia como métrica de avaliação e aprendizado, em problemas de mineração de classes raras ou altamente desbalanceados deve-se utilizar outras métricas que tenham um enfoque maior na classe rara, como por exemplo precisão e sensibilidade (Han, et al., 2009). Assim, deve-se utilizar métricas alternativas à acurácia para avaliar e guiar o aprendizado de métodos classificadores junto a problemas desta natureza, como por exemplo as métricas mostradas na Tabela 5-1.

Enquanto a precisão aborda a porcentagem de verdadeiros positivos dentre todas as amostras que foram preditas como positivas, a sensibilidade é a porcentagem de verdadeiros positivos que foram corretamente preditos como positivos. Portanto, a precisão mede a exatidão do classificador, enquanto a sensibilidade mede a cobertura do mesmo. Dessa forma, essas métricas estão relacionadas diretamente com o critério de desempenho da classificação de problemas desbalanceados, quando aplicadas à classe minoritária.

O aprendizado a partir de dados desbalanceados tem recebido bastante atenção nos últimos anos, sendo considerado por muitos como um dos tópicos mais desafiadores em aprendizado de máquina (Stefanowski, 2013; Wu, et al., 2010). Diversos métodos foram propostos para tratar este problema (Chawla, 2005; Chawla, et al., 2004; He & Garcia, 2009; Weiss, 2004)), sendo que eles podem ser divididos em duas categorias:

- Métodos de pré-processamento, que alteram a distribuição entre as classes e utilizam este

novo conjunto de dados no aprendizado; e

- Métodos que realizam alterações nos algoritmos de aprendizado, utilizando novas estratégias, construção de *ensemble* de classificadores ou ainda classificação com diferentes custos entre as classes (*cost sensitive classifiers*).

Tabela 5-1: Métricas de avaliação mais apropriadas para avaliar classificadores em problemas com desbalanceamento entre classes.

Precisão	$\frac{TP}{TP + FP}$
Sensibilidade	$\frac{TP}{TP + FN}$
F-measure ou F-score	$2 * \frac{\text{precisão} * \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}}$
Geometric-mean or g-mean	$\sqrt{\text{especificidade} * \text{sensibilidade}}$
Matthews Correlation Coefficient ou MCC	$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$
Área abaixo da curva ROC (AUCROC)	Área abaixo da curva no gráfico entre Sensibilidade x 1 – Especificidade.

Técnicas de pré-processamento geralmente envolvem realizar uma subamostragem (*undersampling*) da classe majoritária e sobreamostragem (*oversampling*) da classe minoritária, seguindo diversas estratégias diferentes para criar um novo conjunto de dados balanceado. Subamostragem e sobreamostragem aleatórias (RUS e ROS) são amplamente empregadas nesses casos, sendo que tais técnicas consistem em remover (replicar) amostras da classe majoritária (minoritária) escolhidas aleatoriamente, até que se obtenha um conjunto de dados balanceado. No entanto, remoção e replicação de amostras apresentam diversas desvantagens. A remoção aleatória de amostras negativas pode acabar por remover amostras importantes do conjunto de dados e degradar ainda mais a capacidade de generalização (predição de novas amostras) dos classificadores. Por outro lado, a replicação de amostras positivas não introduz nenhum dado novo ao conjunto de amostras e, por fim, não há garantias de uma melhor definição das regiões de fronteira entre as classes. Para superar estas desvantagens, diversos métodos de pré-processamento foram propostos, dentre eles o SMOTE (Chawla, et al., 2002), provavelmente um dos mais conhecidos e empregados. O método SMOTE, abreviação para *Synthetic Minority Over-sampling Technique*, busca realizar a sobreamostragem da classe minoritária introduzindo novas amostras sintéticas criadas na vizinhança de amostras minoritárias do conjunto de dados. No

algoritmo, escolhe-se o número k de vizinhos da classe minoritária que serão utilizados para construir as amostras sintéticas, bem como a taxa de sobreamostragem (número de amostras sintéticas criadas). Recentemente, uma modificação do SMOTE foi proposta de forma a adaptá-lo a problemas desbalanceados. Uma das desvantagens do SMOTE, quando aplicado a dados desbalanceados, é que amostras da classe minoritária podem estar relativamente distantes no espaço de variáveis, de forma que novas amostras sintéticas sejam criadas em uma região entre estes vizinhos que contenha muitas amostras negativas em sua vizinhança, dada a baixa representatividade das amostras positivas (ver Figura 5-2). O algoritmo LN-SMOTE (Maciejewski & Stefanowski, 2011) realiza a criação de novas amostras sintéticas para a classe positiva somente em regiões consideradas “seguras” pelo algoritmo, minimizando assim o risco de se criar amostras positivas muito próximas às amostras negativas. Para identificar estas regiões consideradas “seguras”, o algoritmo rotula cada amostra do conjunto de dados (positivas e negativas) de acordo com a classificação realizada pelos seus k vizinhos mais próximos (KNN). Se a classificação de uma amostra condiz com sua classe, esta é considerada segura. Caso contrário, é considerada insegura, ou seja, para ser considerada segura a vizinhança de uma amostra necessita possuir uma maioria de amostras da mesma classe. Se a amostra é considerada estar em uma região insegura, qualquer tentativa de gerar uma amostra sintética nesta região tende a resultar na introdução de amostras em regiões onde há maior concentração de amostras negativas. Posteriormente, amostras inseguras da classe negativa (majoritária) são removidas, e as amostras consideradas seguras da classe positiva são utilizadas para criar amostras sintéticas. O LN-SMOTE pode ser aplicado mais de uma vez ao conjunto de dados, pois, após a remoção de amostras negativas inseguras, amostras positivas anteriormente consideradas inseguras podem vir a ser rotuladas como seguras, devido à remoção de amostras negativas em sua vizinhança.

O método de subamostragem *Nearest Cleaning Rule* (NCL) (Laurikkala, 2001), no qual o LN-SMOTE baseia-se para a remoção de amostras negativas, consiste em remover amostras da classe majoritária consideradas ruidosas (possuem uma classificação diferente de seus três vizinhos mais próximos). Para isto, utiliza-se o conceito de *Edited Nearest Neighbor* (ENN) (Wilson & Martinez, 2000) para identificar amostras ruidosas. No entanto, em problemas com alto desbalanceamento, há uma tendência natural de haver poucas ou até mesmo nenhuma amostra negativa com um número de vizinhos da classe positiva maior do que da classe negativa, de forma que pouquíssimas amostras negativas são removidas. Consequentemente, o desbalanceamento não é superado.

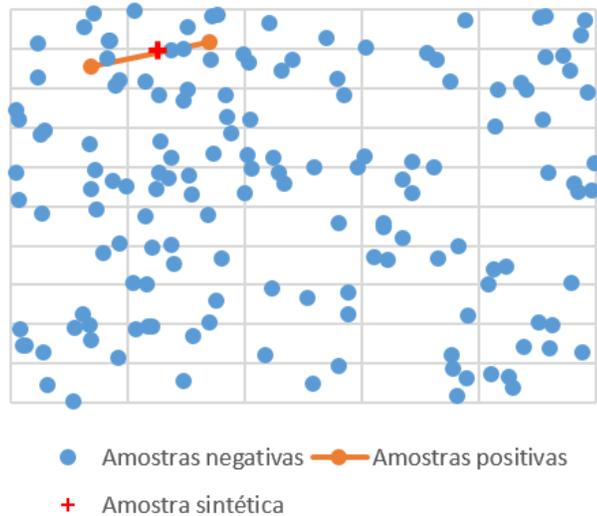


Figura 5-2: Esquema didático da aplicação do SMOTE, ocasionando a introdução de amostras sintéticas da classe positiva próximas a amostras negativas. Amostra sintética (em vermelho) é criada próxima a amostras negativas (azuis) pelo SMOTE dada a baixa representatividade da classe minoritária (em laranja).

JAPKOWICZ (2003) realizou um estudo para avaliar a influência de três parâmetros em problemas com alto desbalanceamento entre classes: complexidade, tamanho do conjunto de treinamento e grau de desbalanceamento. Para isto, foram gerados 125 domínios de dados considerando diversas combinações destes três parâmetros, de forma que cada conjunto de dados é dividido em intervalos de igual tamanho, sendo que intervalos contíguos contêm amostras de classes opostas (Figura 5-3). Conjuntos de dados com maior complexidade possuem maior número de intervalos contíguos, sendo que, para cada intervalo, amostras são geradas a partir de uma distribuição uniforme e o número de amostras em cada intervalo é dado pelo grau de desbalanceamento e tamanho do conjunto de dados desejado. Utilizando o algoritmo C5.0 (Quinlan, 2001), versão comercial e possivelmente melhorada do C4.5 (Quinlan, 1993), foi possível observar que, nos casos em que o desempenho do classificador foi considerado baixo, o grau de desbalanceamento não foi o maior responsável pela degradação de desempenho. Em conjuntos de dados com baixa complexidade (linearmente separáveis), não importando o tamanho dos conjuntos de dados considerados e o grau de desbalanceamento entre classes, o desbalanceamento em si não mostrou ser um obstáculo para a classificação. No entanto, à medida que a complexidade aumenta, os classificadores tornam-se mais sensíveis ao desbalanceamento. Assim, em problemas desbalanceados, o desbalanceamento pode não ser o maior desafio, sendo grau de complexidade e tamanho dos conjuntos de dados os causadores da degradação do desempenho dos classificadores. Os autores mostraram também que, fixando o número de amostras positivas por intervalo em 50, e variando-se somente o grau de complexidade e grau de desbalanceamento, quando a classe minoritária encontra-se dividida em um

número maior de pequenos subconceitos (altas complexidades) o desempenho dos classificadores é bastante afetado, mesmo nos casos de baixo grau de desbalanceamento. Assim, concluem que nem sempre é o próprio desbalanceamento o maior problema para a classificação, mas sim, a presença de pequenas disjunções (*small disjunctions*) da classe minoritária, ou seja, não são complexidade, grau de desbalanceamento ou tamanho do conjunto de treinamento que causam esta degradação. A degradação está muito mais ligada ao fato da classe minoritária estar dividida em pequenos subgrupos (*clusters*). São propostas duas abordagens baseadas em *clusterização* e divisão do problema em subproblemas de classificação, como tentativa de aliviar o impacto destes problemas.

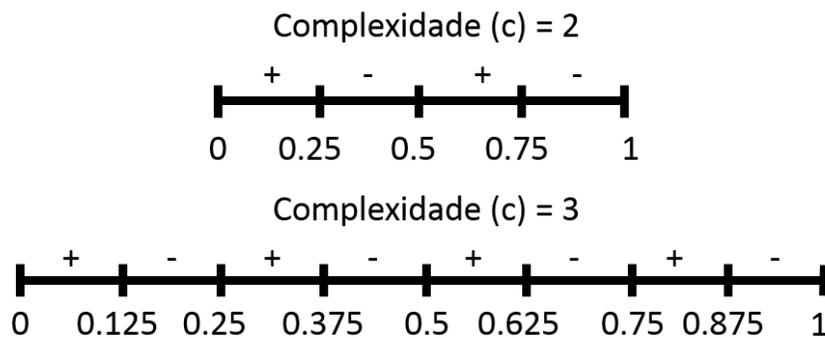


Figura 5-3: Divisão dos dados sintéticos para dois diferentes níveis de complexidade utilizados em (Japkowicz, 2003) Sendo o número de intervalos igual a 2^c , de forma que, com aumento da complexidade o número de intervalos aumenta exponencialmente e, como consequência as amostras sofrem maior segregação em intervalos adjacentes contendo classes opostas. +: amostras positivas; -: amostras negativas.

PRATI *et al.* (2004) estudaram o impacto da sobreposição entre classes minoritária e majoritária. Para o estudo, foram gerados conjuntos de dados artificiais onde as classes minoritária e majoritária foram representadas por dois *clusters* de dimensão 5, utilizando-se uma distribuição gaussiana para gerar amostras em torno dos centroides. Variando-se a taxa de desbalanceamento e distância entre os centroides, de forma que as classes pudessem ser movidas de uma total separação para uma alta sobreposição, os autores mostraram, utilizando um classificador C4.5, que a área abaixo da curva ROC (AUCROC) sofre uma influência maior do aumento de sobreposição entre as classes do que da taxa de desbalanceamento. No caso em que as classes estão claramente separadas (nenhuma sobreposição), a taxa de desbalanceamento não possui efeito algum sobre as medidas de desempenho do classificador. Neste mesmo trabalho, os autores ainda conduziram experimentos para avaliar cinco técnicas de pré-processamento para balanceamento de dados diferentes, compreendendo: sobreamostragem aleatória (ROS), subamostragem aleatória (RUS), *Nearest Cleaning Rule* (NCR), SMOTE e SMOTE + ENN (SMOTE + *Edited Nearest Neighbor Rule*) (uma união dos métodos SMOTE e ENN que permite realizar também sub-amostragem da classe majoritária, similar ao LN-SMOTE). A partir de dados artificiais,

concluíram que sobreamostragem tem grande impacto no desempenho dos classificadores, melhorando-o consideravelmente quando comparado com subamostragem. Porém, quando o grau de sobreposição é muito alto, não é possível definir quais dos métodos é o melhor. No entanto, para todos os outros casos, o SMOTE + ENN obteve um alto desempenho mesmo quando havia alguma sobreposição entre as classes. A Figura 5-4 traz um esquema simplificado de sobreposição entre classes e de pequenas disjunções.

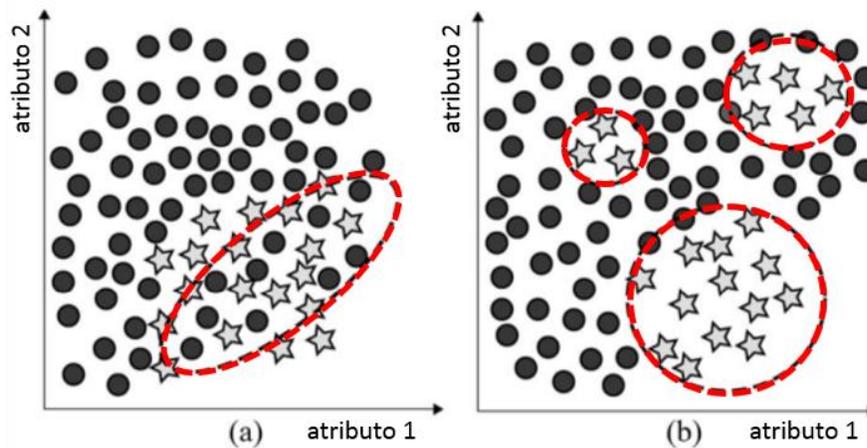


Figura 5-4: Desafios na classificação de dados desbalanceados. (a) sobreposição entre classes; (b) pequenas disjunções. Figura baseada em GALAR (2012).

Diversas abordagens evolutivas também têm sido utilizadas para promover tanto a redução na dimensionalidade dos dados (seleção de atributos) (Whitley, et al., 1998; Guerra-Salcedo, et al., 1999; Papadakis & Theocharis, 2006; Wang, et al., 2007; Sikora & Piramuthu, 2007) quanto para a seleção de protótipos (Cano, et al., 2003; Galar, et al., 2013).

CANO *et al.* (2007) propõem um algoritmo evolutivo para realizar seleção de amostras de forma estratificada, particionando o conjunto de dados em um número pré-estabelecido de subconjuntos, sendo que a seleção é realizada em cada subconjunto separadamente. Ao final do algoritmo, as melhores soluções são agrupadas, formando um conjunto de treinamento contendo um subconjunto das amostras iniciais. Neste trabalho, testou-se o emprego do método proposto, mas com pequenas alterações no formato dos subconjuntos (estratos). Devido ao desbalanceamento entre as classes, ao invés de considerar todas as classes para gerar os subconjuntos de amostras, realizou-se somente a estratificação das amostras negativas, sendo que as amostras positivas não foram subamostradas. Inicialmente, as amostras negativas são divididas aleatoriamente em k subconjuntos e cada subconjunto é submetido a um processo evolutivo para a seleção de amostras negativas que, juntamente com todas as amostras positivas, resultem em ganhos de desempenho dos classificadores treinados. A Figura 5-5 ilustra o procedimento. O algoritmo

CHC (Eshelman, 1991) é empregado como abordagem evolutiva. Neste esquema, a cada geração $N/2$ indivíduos de uma população de N indivíduos são escolhidos aleatoriamente para a aplicação do operador de recombinação e gerar uma nova população de N indivíduos. O algoritmo utiliza recombinação metade uniforme (HUX – *Half Uniform Crossover*), que consiste em trocar metade dos bits, aleatoriamente escolhidos, que diferem nos indivíduos pareados (pais). CHC também implementa um método para a prevenção de incesto, calculando a distância de *Hamming* antes de realizar a recombinação entre dois indivíduos: o operador somente é aplicado caso essa distância seja maior que um valor limite estabelecido. O valor inicial para este limite é igual a $L/4$, onde L é o número de bits em cada indivíduo, de forma que este valor é reduzido em 1 caso nenhuma nova solução seja gerada pelo operador de recombinação. O algoritmo não utiliza operador de mutação. Ao invés disto, se não há mais progresso no processo evolutivo, devido aos novos indivíduos não possuírem *fitness* melhor que as soluções anteriores ou o valor da distância limite cair a zero, uma nova população com $N-1$ indivíduos é gerada, utilizando-se a melhor solução como modelo, onde 35% dos bits da melhor solução são aleatoriamente transferidos para os novos indivíduos gerados e o restante é aleatoriamente escolhido.

Para avaliar as soluções de cada processo evolutivo, utilizou-se o algoritmo RIPPER, sendo o *fitness* do indivíduo igual ao valor da medida-F obtida em um conjunto de teste independente do conjunto utilizado para treinamento. Ao final, as melhores soluções (maiores medidas-F), de cada um dos k processos evolutivos, são agrupadas para formar um conjunto com um número reduzido de amostras negativas, que juntamente com as amostras positivas irão compor o conjunto de treinamento (TS). O conjunto de treinamento é então utilizado para treinar o classificador final, sendo o desempenho deste medido em um conjunto de teste independente dos demais utilizados.

Algoritmos evolutivos podem ser bastante sensíveis ao tipo de codificação utilizada. Para o problema de seleção de amostras, geralmente uma codificação binária é adotada. Assim, para conjuntos contendo grandes quantidades de amostras, os indivíduos serão representados por cadeias de bits de mesma ordem, o que pode dificultar a convergência do algoritmo, levando a solução subótimas. A estratificação, como proposta por CANO *et al.* (2007), procura aliviar este comportamento. Desde que amostras sejam independentes umas das outras, esta divisão não implica em perda de informação.

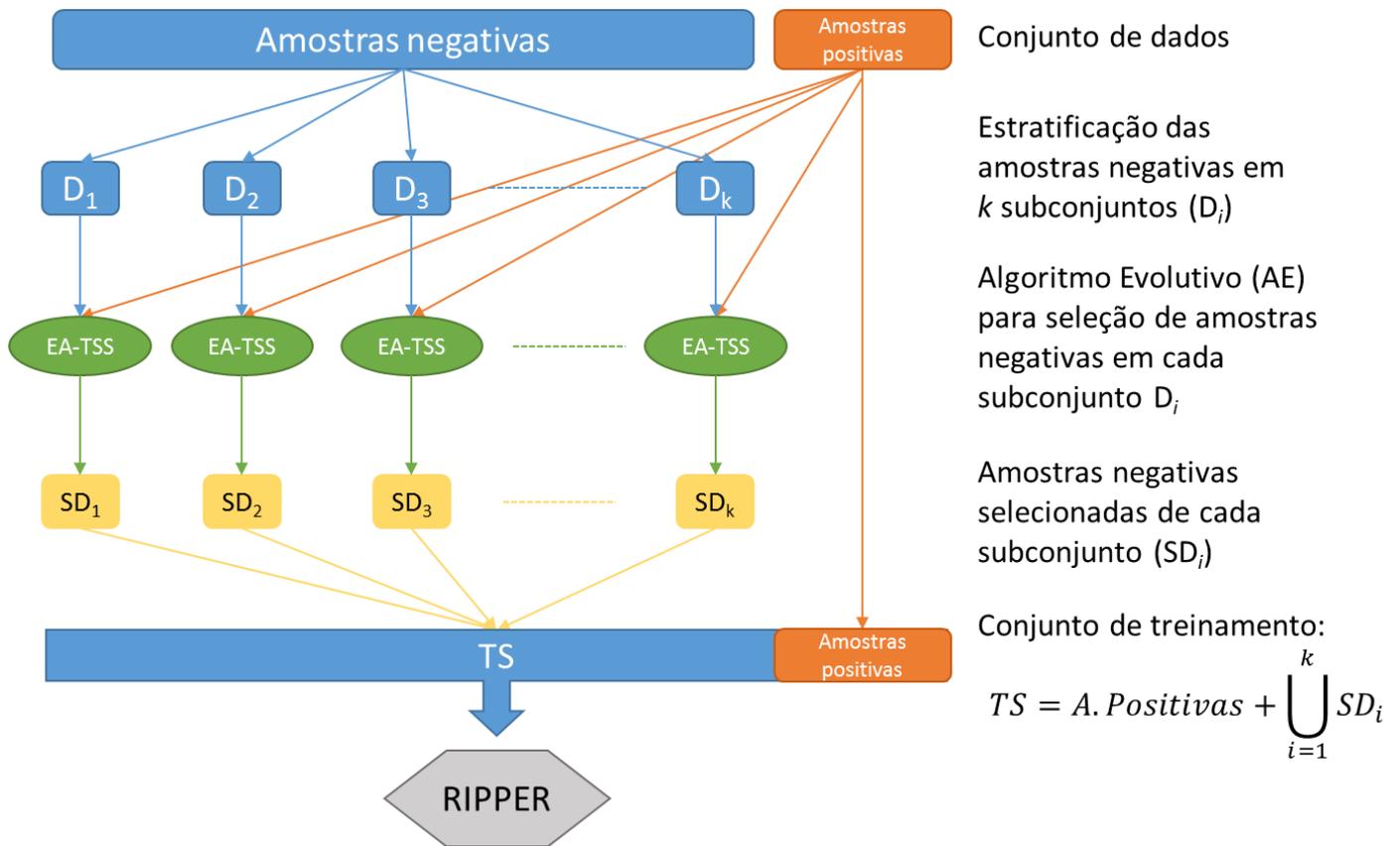


Figura 5-5: Proposta evolutiva para a seleção de conjuntos de treinamento balanceados. Modificação da abordagem proposta em CANO et al. (2007). AE: algoritmo evolutivo; TS: training set (conjunto de treinamento).

Por outro lado, métodos especializados procuram realizar alterações nos algoritmos de aprendizado como tentativa de superar os desafios advindos do desbalanceamento entre as classes. Estes diferem dos métodos anteriores, pois, não realizam pré-processamento dos dados a fim de aliviar o desbalanceamento. Pelo contrário, utilizam todo o conjunto de dados disponível, tratando o desbalanceamento simultaneamente com o aprendizado. Entre eles, encontram-se técnicas baseadas em *clusterização* (Wu, et al., 2010), *boosting* e *bagging* (Galar, et al., 2012; Błaszczyszki, et al., 2013), meta-algoritmos (Basheer & Hamid, 2012; Zięba, et al., 2014) e indutores de regras especializados (Napierala & Stefanowski, 2012). Técnicas utilizando *boosting* e *bagging* consistem em construir *ensembles* de classificadores-base treinados em diferentes subconjuntos dos dados ou diferentes pesos associados a cada amostra, de forma a fornecerem um classificador mais diversificado, composto pela união das saídas de cada um dos classificadores-base. Uma revisão bastante completa destes métodos foi realizada por GALAR et al. (2012), onde os autores realizaram diversos experimentos utilizando uma vasta diversidade de métodos especializados para problemas desbalanceados, incluindo *boosting* e *bagging*. Os autores reportam que os métodos SMOTEBagging (Wang & Yao, 2009), RUSBoost (Seiffert, et al., 2009) e UnderBagging (Barandela, et al., 2003) foram os que resultaram em melhor desempenho,

segundo critérios de taxas de classificação, tempo de execução e complexidade de implementação dos algoritmos.

O método SMOTEBagging incorpora o método de pré-processamento à técnica de *bagging* (similarmente o algoritmo SMOTEBoost (Chawla, et al., 2003), que incorpora o SMOTE à técnica de *boosting*). *Bagging* (Breiman, 1996) refere-se ao conceito de *bootstrap aggregating*, utilizado para construir *ensembles* de classificadores. Um *ensemble* é formado pela união de vários classificadores, formando um tipo de comitê, onde a saída de cada classificador é, então, unificada para se obter uma saída global do *ensemble*. Esta unificação pode ser feita de diversas formas, incluindo por exemplo um esquema de votação, onde o *ensemble* classifica uma amostra de acordo com a maioria dos votos (classificação de cada classificador no *ensemble*). No caso de *Bagging*, amostras são selecionadas aleatoriamente (com reposição) do conjunto de treinamento e utilizadas para treinar classificadores-base que serão utilizados para compor o esquema de *ensemble*. Dessa forma, os algoritmos SMOTEBagging e UnderBagging utilizam *bagging* para iterativamente amostrar (com reposição) um subconjunto dos dados de treinamento. No caso do SMOTEBagging, o subconjunto amostrado é pré-processado utilizando-se o algoritmo SMOTE para sobreamostragem da classe minoritária. Enquanto que no algoritmo UnderBagging é empregada a subamostragem aleatória da classe majoritária como pré-processamento. Os classificadores-base, treinados com conjuntos de dados balanceados, são utilizados para compor o *ensemble* de classificadores.

De forma similar, porém utilizando a técnica de *boosting*, o algoritmo RUSBoost utiliza subamostragem aleatória para criar os subconjuntos de dados que serão utilizados no treinamento de cada classificador base. De todos os algoritmos que fazem emprego da técnica de *boosting* (Schapire, 1990), sem dúvida o mais conhecido e utilizado é o algoritmo AdaBoost (Freund & Schapire, 1997). Diferentemente dos métodos de *bagging*, algoritmos de *boosting*, em especial o AdaBoost, não realizam amostragem dos dados, sendo o treinamento dos classificadores-base realizado utilizando-se todo o conjunto de treinamento. No entanto, o AdaBoost, a cada iteração, atualiza os pesos das amostras de acordo com a classificação encontrada pelo classificador-base, aumentando o peso das amostras erroneamente classificadas e reduzindo o peso de amostras corretamente classificadas para a próxima iteração. Assim, o classificador-base irá focar na classificação daquelas amostras com maior peso (erroneamente classificadas nas iterações passadas). O AdaBoost também atribui à cada classificador-base um peso associado ao seu desempenho geral. Para realizar a classificação de amostras, os pesos associados a cada classificador-base são utilizados para ponderar suas saídas, dando maior ênfase àqueles classificadores com maior desempenho durante o treinamento. A predição do classificador treinando com AdaBoost é, então, uma agregação ponderada das classificações realizadas pelos classificadores-base

presentes no *ensemble*. Para o caso de desbalanceamento, foi proposto o algoritmo RUSBoost (Seiffert, et al., 2009), que incorpora ao processo de *boosting* a subamostragem aleatória da classe majoritária. Dessa forma, classificadores-base são treinados em conjuntos de dados balanceados e com diferentes pesos para as amostras da classe majoritária e minoritária.

Recentemente, utilizando o mesmo princípio dos métodos RUSBoost e SMOTEBoost, GALAR *et al.* (Galar, et al., 2013) propuseram um método que utiliza uma abordagem evolutiva para realizar subamostragem da classe majoritária. Baseado nos algoritmos evolutivos para subamostragem (García & Herrera, 2009) e seleção de protótipos (García, et al., 2012), o método EUSBoost (*Evolutionary UnderSampling Boosting*) incorpora a subamostragem evolutiva ao AdaBoost, ao invés de utilizar subamostragem aleatória, como no RUSBoost.

Análogo ao método proposto em JAPKOWICZ (2003), WU *et al.* (2010) propuseram um método chamado COG (*Classification using lOcal clusterinG*) para decompor cada classe original do problema em subclasses, obtidas a partir do emprego de um algoritmo de clusterização a cada classe separadamente. Assim, em um problema em que o número total de classes é igual a c , e sendo o número de clusters para a i -ésima classe ($i=1\dots c$) igual a k , esta será subdividida em k subclasses, de forma que este processo seja aplicado a todas as classes. Com novos rótulos, as amostras são agrupadas e utilizadas para o treinamento de um algoritmo de classificação, sendo a saída do classificador convertida posteriormente para a classe original da amostra. Os autores recomendam o uso do algoritmo *K-means* (MacQueen, 1967), com um número de clusters k não maior que 4, e mostraram que o emprego do método eleva as taxas de desempenho dos classificadores treinados utilizando Máquinas de Vetores de Suporte (SVM) (Cortes & Vapnik, 1995) e o indutor de regras RIPPER (Cohen, 1995) para diversos conjuntos de dados.

Dada a diversidade de métodos propostos, é difícil saber qual método é o melhor para ser utilizado. Esta informação depende muito do problema abordado e das características dos dados. Dessa forma, a realização de testes faz-se necessária em cada caso.

O problema de classificação de resíduos de aminoácidos catalíticos enquadra-se na classe de problemas de classificação de casos raros, uma vez que nos conjuntos de cada sub-subclasse EC, e mesmo em toda a base de dados, o número de resíduos de aminoácidos catalíticos é geralmente inferior a 1% do total de amostras. Neste trabalho, procurou-se explorar o uso de alguns métodos de pré-processamento, seguindo trabalhos que mostraram evidências de superioridade destes métodos sobre outros. Apesar de existirem outras abordagens específicas, como por exemplo SVM's especializadas (Imam, et al., 2006; Veropoulos, et al., 1999) e Redes Neurais Artificiais (Adam, et al., 2010), optou-se por não utilizar estas abordagens neste trabalho, pois procura-se por modelos que possam fornecer uma interpretação direta de

seu mapeamento, em contraposição a métodos considerados como “caixas pretas” (*black box*).

5.3 Árvores de decisão

Árvore de decisão designa um modelo para representar decisões e possíveis consequências à medida que se caminha pelos nós da árvore (da raiz a uma das folhas). Em aprendizado de máquina, existem algoritmos que possibilitam a construção destas árvores de decisão a partir de um conjunto de dados fornecido para treinamento, criando assim um mapeamento entre as características de uma amostra e a respectiva classe desta amostra.

Uma árvore de decisão pode ser construída através da divisão de um conjunto de amostras em subconjuntos menores, baseando-se em um teste sobre a qualidade de um atributo em distinguir as amostras segundo suas classes, de forma que os subconjuntos derivados deste teste sejam mais homogêneos (exibam menor impureza) em relação à mistura de classes. Este processo é repetido para todos os nós da árvore (subconjuntos de amostras) até que, em um dado nó, todas as amostras tenham o mesmo valor para a variável de interesse. Este processo de construção de árvores de decisão, conhecido como *Top-Down Induction of Decision Tree* (TDIDT) (Quinlan, 1986) é um exemplo de algoritmo guloso, e é a forma mais utilizada para a construção de árvores de decisão. No entanto, não é a única estratégia possível para a construção destas árvores (Landeweerd, et al., 1983).

Existem diversos algoritmos na literatura para a construção de árvores de decisão. Entre os mais conhecidos estão o C4.5 (Quinlan, 1993; Quinlan, 1996), CART (*Classification and Regression Tress*) (Breiman, et al., 1984) e OC1 (Murthy, et al., 1993). No entanto, existem outros algoritmos como CHAID (Kass, 1980), MARS (Friedman, 1991), entre outros (veja KOTSIANTIS (2013) para uma revisão recente de novas propostas e modificações de algoritmos para a construção de árvores de decisão).

Árvores de decisão podem ser consideradas como univariadas ou multivariadas (ou oblíquas) dependendo do tipo de teste conduzido em cada ramificação da árvore. No caso univariado, apenas uma característica de cada vez é utilizada para compor os testes em cada ramificação. Por outro lado, árvores multivariadas podem considerar combinações lineares de múltiplas características ao mesmo tempo, e até mesmo combinações mais complexas, como o uso de *perceptrons* (*Perceptron Decision Trees*) (Bennett & Mangasarian, 1992; Bennett & Mangasarian, 1994a; Bennett & Mangasarian, 1994b; Breiman, et al., 1984; Broadley & Utgoff, 1995; Utgoff, 1989; Murthy, et al., 1994) e Redes Bayesianas (Kohavi, 1996).

Neste trabalho, restringiu-se o uso somente ao algoritmo C4.5, por construir árvores de decisão univariadas, uma vez que além destas árvores serem facilmente interpretadas, muitas vezes a combinação multivariada de características não possui um significado biológico claro. A Figura 5-6 ilustra um

esquema simples da diferença na separação entre duas classes realizadas por árvores de decisão univariadas e multivariadas. No caso de árvores de decisão univariadas, os testes para separação das amostras (ramificação) são realizados considerando-se somente um único atributo de cada vez. Ao passo que no caso multivariado é possível realizar combinações de múltiplos atributos em cada teste. Novamente, apesar de fornecerem árvores menores (menor número de nós), as combinações realizadas pelas árvores de decisão multivariadas podem não ter um significado natural. Por isso, árvores de decisão univariadas foram empregadas, mesmo que estas últimas possam gerar árvores maiores, aumentando a complexidade das regras.

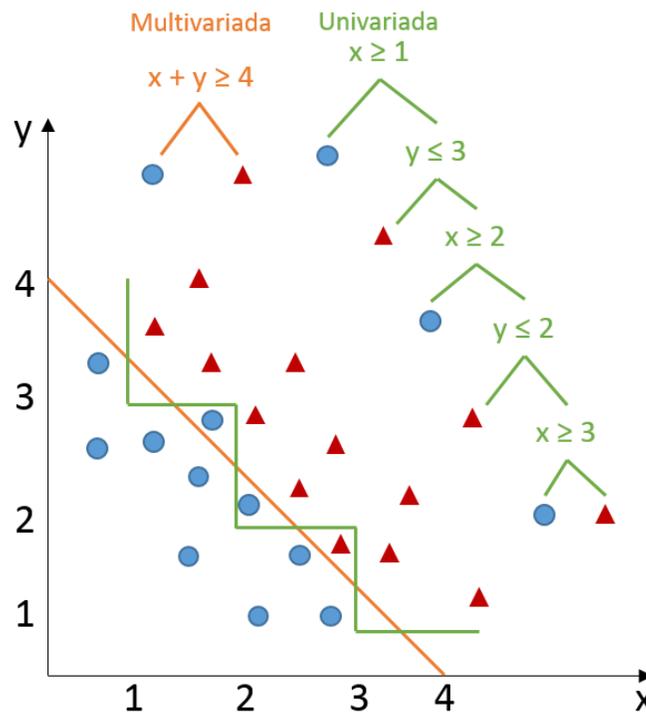


Figura 5-6: Comparação ilustrativa de Árvores de Decisão uni e multivariadas. Árvores multivariadas possuem menor número de nós, porém podem não oferecer significado contextual para as combinações de atributos. Árvores de decisão univariadas são maiores, porém facilmente interpretadas.

O C4.5 utiliza uma abordagem conhecida como *divide-and-conquer* (dividir e conquistar) para a construção das árvores de decisão. Nesta abordagem, as amostras são recursivamente separadas através de testes obtidos pela avaliação dos atributos, segundo algum critério adotado. No C4.5, o conjunto de amostras D é recursivamente subdividido em subconjuntos D_1, D_2, \dots, D_k segundo aplicação de um teste T com k possíveis resultados mutuamente exclusivos: T_1, T_2, \dots, T_k , de forma que D_i contenha as amostras que com o resultado T_i para o teste T . Assim, a cada ramificação, o algoritmo busca encontrar testes que minimizem a impureza num dado nó da árvore, levando ao final a um nó (terminal) que contenha apenas amostras de uma classe em particular. O critério adotado no C4.5 para avaliar a qualidade dos testes em

cada nova ramificação é a taxa de ganho (*gain ratio*), dada na forma:

$$GainRatio(D, T) = \frac{Gain(D, T)}{SplitInfo(D, T)} \quad \text{Equação 5-1}$$

onde $Gain(D, T)$ é o ganho de informação obtido pelo teste T com k possíveis resultados (valores para o atributo considerado no teste) dado pela Equação 5-2, e $Info(D)$ (Equação 5-4) é a medida de entropia (incerteza) no conjunto de amostras D , com relação às suas classes. Assim, em um nó da árvore onde todos as amostras pertençam a uma mesma classe, a entropia será zero, mostrando nenhuma desordem (incerteza) sobre os valores das classes neste nó. O denominador $SplitInfo(D, T)$ refere-se à distribuição das amostras para o teste T , resultando na normalização do ganho (Equação 5-3). A taxa de ganho é calculada para todos os possíveis testes e, dentre aqueles maiores do que o ganho médio, o que fornece uma ramificação com máxima taxa de ganho é escolhido.

$$Gain(D, T) = Info(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} \log_2 Info(D_i) \quad \text{Equação 5-2}$$

$$SplitInfo(D, T) = - \sum_{i=1}^k \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|} \quad \text{Equação 5-3}$$

$$Info(D) = - \sum_{j=1}^c p(D, j) \log_2 p(D, j) \quad \text{Equação 5-4}$$

Originalmente projetado para aceitar somente atributos discretos (categóricos ordinais e não-ordinais) o algoritmo C4.5 foi então expandido para também aceitar atributos contínuos, ampliando assim seu uso na área de aprendizado de máquina (Quinlan, 1996). Para evitar o sobreajuste das árvores de decisão induzidas, algoritmos de construção de árvores de decisão geralmente empregam alguma fase de poda (*prune*), onde uma subárvore inteira é substituída por uma folha, a fim de simplificar o modelo criado e garantir uma maior capacidade de generalização do modelo. No C4.5, após a construção de uma árvore de decisão (possivelmente sobreajustada) o algoritmo realiza a poda da árvore, buscando reduzir a taxa de erro na predição de amostras. Para isto, uma estimativa de erro é calculada para as folhas e subárvores a partir do próprio conjunto de treinamento, utilizando a técnica de poda estatística. Assim, uma folha cobrindo N amostras, sendo E incorretamente classificadas, terá uma taxa de erro estimada dada pela valor da distribuição binomial $B_{CF}(N, E) \sim B(n, p)$, considerando o intervalo de confiança CF , fornecido como parâmetro do algoritmo. Os nós terminais (folhas), portanto, terão um número de amostras incorretamente classificadas dada por $N \times B_{CF}(N, E)$, sendo o número de erros associado a subárvores dado pela soma dos erros em cada uma de suas ramificações. Caso o número de amostras incorretamente classificadas ($N \times B_{CF}(N, E)$) em uma subárvore seja maior do que caso esta fosse

substituída por uma folha, o algoritmo realiza a poda desta subárvore.

5.4 Indução de regras

Apesar de ser possível a extração de regras a partir de árvores de decisão, existem algoritmos especializados em encontrar estas regras sem a construção de modelos auxiliares (e.g. árvore de decisão), conhecidos como indutores de regras. Estas regras podem ser uma representação completa de um modelo científico ou simplesmente uma representação de padrões nos dados. As regras podem ainda apresentar uma ordem pré-definida (*ordered rules*) ou simplesmente nenhuma ordem (*unordered rules*). A diferença é que regras ordenadas precisam ser aplicadas sempre na mesma ordem com que foram encontradas, ou seja, para um conjunto de regras ordenadas contendo R regras, a regra R_{i+1} somente será aplicada a uma amostra caso a regra R_i não cubra esta amostra. No caso sem uma ordem definida, as regras podem ser aplicadas em qualquer ordem. A construção de conjuntos de regras ordenados e não ordenados é obtida a partir de diferentes técnicas de aprendizado.

Dentre os algoritmos mais estudados e utilizados para indução de regras encontra-se o RIPPER (Cohen, 1995). Acrônimo para *Repeated Incremental Pruning to Produce Error Reduction*, o algoritmo utiliza a abordagem *separate-and-conquer* para encontrar um conjunto de regras ordenadas. Baseado no algoritmo IREP* (Fürnkranz & Widmer, 1994), o algoritmo aplica iterativamente a técnica IREP* para obter um conjunto de regras até que todas as amostras tenham sido cobertas. O RIPPER realiza busca por regras partindo-se da classe com menor número de amostras para a classe com maior número de amostras, sendo que para a última classe sempre é gerada uma única regra vazia (sem antecedente) que classifica todas as amostras restantes como pertencentes a esta classe (regra padrão). Assim, num caso de classificação binária, o RIPPER irá construir regras somente para a classe minoritária (positiva), sendo as amostras não cobertas por nenhuma das regras tidas como pertencentes à classe majoritária (negativa). Partindo-se de uma regra geral (sem antecedentes) para cada classe, a cada iteração o algoritmo busca regras especializadas que cubram o maior número possível de amostras da classe e excluam as amostras de outras classes. Por causa desta busca por regras separadamente para cada, classe o algoritmo se enquadra na categoria de algoritmo *separate-and-conquer*.

Uma vantagem desta abordagem, no caso de desbalanceamento entre classes, é que o algoritmo não irá desprezar as amostras da classe minoritária, a menos que estas realmente não resultem em nenhum ganho para a classificação. No entanto, como nos outros métodos, não há garantias de que o algoritmo irá ter um bom desempenho nestes casos, inclusive onde ainda pode haver sobreposição entre classes ou pequenos conjuntos disjuntos para a classe minoritária.

Apesar de produzir regras bastante simples e facilmente interpretáveis (Rückert & De Raedt, 2008),

e em alguns casos mostrar desempenho ligeiramente superior ao C4.5 (Stefanowski, 2013), o algoritmo é bastante sensível a ruídos nos dados e indutores de regra em geral tendem a possuir baixa capacidade de generalização.

Inicialmente, o algoritmo RIPPER divide o conjunto de amostras em dois subconjuntos, sendo um deles usado para a indução das regras (*grow set*) e outro para realizar a poda das regras (*prune set*), como tentativa de evitar sobreajuste (*overfitting*) das regras. A partir do conjunto de indução das regras (*grow set*), antecedentes são escolhidos de forma a maximizarem o ganho de informação de FOIL (*First Order Inductive Learner*) (Quinlan, 1990) (Equação 5-5).

$$Info_{FOIL}(R_i, R_{i-1}) = p_i * \left(\log_2 \frac{p_i}{p_i+n_i} - \log_2 \frac{p_{i-1}}{p_{i-1}+n_{i-1}} \right) \quad \text{Equação 5-5}$$

onde p_i e n_i são os números de amostras positivas e negativas, respectivamente, cobertas pela regra R_i . Sendo R_i a regra estendida a partir de R_{i-1} via adição de um novo antecedente, com p_{i-1} e n_{i-1} amostras positivas e negativas, respectivamente.

Toda nova regra gerada é então podada pelo algoritmo utilizando o mesmo procedimento que o algoritmo IREP* (Furnkranz & Widmer, 1994), mas utiliza-se uma métrica diferente para remover os últimos antecedentes da regra, de forma que estes maximizem a Equação 5-6.

$$v = \frac{TP - FP}{TP + FP} \quad \text{Equação 5-6}$$

O algoritmo realiza ainda diversas iterações de otimização do conjunto de regras, onde o mesmo procedimento de aprendizado e poda das regras é repetido alterando-se os conjuntos de treinamento e poda. Para cada regra no conjunto de regras $RS = \{R_1, R_2, \dots, R_k\}$, duas novas regras são geradas: uma regra de substituição e outra de revisão. Enquanto que a regra de substituição inicialmente está vazia, a regra de revisão é uma cópia de regra R_i . Ambas as regras são então submetidas ao processo de indução e poda. Como os conjuntos de treinamento e poda diferem daqueles utilizados anteriormente, a regra de substituição poderá ser completamente diferente da regra anterior e, ainda uma versão melhorada desta. O mesmo ocorre para a regra de revisão, em que partindo-se da regra anterior, novos antecedentes serão adicionados aumentando a especificidade da regra. No entanto, ao ser podada, a regra de revisão, pode resultar em uma regra melhor do que R_i . A partir das novas regras geradas, novos conjuntos de regras temporários são construídos. O RIPPER, então, realiza o cálculo da MDL (*Minimum Description Length*) (Rissanen, 1978; Wallace & Boulton, 1968) para cada um dos três conjuntos de regras (RS e os dois conjuntos temporários criados a partir das regras de substituição e revisão). A MDL é uma métrica muito utilizada e importante na área de teoria da informação, podendo ser escrita como mostrado na Equação

5-7. A DL (*Description Length*) de um conjunto de regras pode ser entendida como o somatório das DL's de todas as regras presentes no conjunto, bem as DL's das amostras não cobertas por nenhuma regra neste conjunto. Dessa forma, dentre estas três regras, aquela que ao ser adicionada ao conjunto de regras resulta em uma menor MDL, de forma que o conjunto de regras possua regras mais simples (menor número de antecedentes) e cubra o maior número de amostras, é escolhida para compor o conjunto final de regras RS*. Este processo é repetido até que todas as regras de RS tenham sido otimizadas.

$$MDL(H) = I(H) - I(E|H) \quad \text{Equação 5-7}$$

onde $I(H)$ representa a quantidade de informação necessária para transmissão da hipótese H , e $I(E|H)$ a quantidade de informação necessária para transmitir um conjunto de exemplos E , no qual a hipótese H foi construída.

5.5 Seleção de Atributos

Devido ao grande número de atributos (descritores estruturais de proteínas) a escolha do melhor atributo a cada iteração dos algoritmos descritos anteriormente apresenta diversos desafios estatísticos inerentes à análise de dados de alta dimensão (Johnstone & Titterington, 2009), além de tendência ao sobreajuste dos modelos treinados.

Métodos de seleção de atributos procuram selecionar um subconjunto dos atributos originais, de acordo com algum critério adotado de forma que este subconjunto leve a um melhor desempenho dos modelos treinados, além de reduzir o custo computacional do treinamento e aumentar a interpretabilidade dos classificadores. Os principais tipos de métodos de seleção de atributos incluem filtros e *wrappers*, além de métodos incorporados diretamente nos algoritmos de treinamento.

Filtros representam a categoria de pré-processamento. Assim, o processo de seleção de atributos não influencia no processo de treinamento do classificador, e vice-versa. São baseados em métricas que exploram características diretamente extraídas dos dados, como distâncias, consistências, dependência, informação e correlação. Entre os mais representativos, encontram-se métodos baseados ou derivados de: ganho de informação (Peng, et al., 2005), Fisher score (Duda, et al., 2001) e Relief (Sikonja & Kononenko, 2003).

Por outro lado, *wrappers* utilizam o desempenho dos algoritmos de aprendizado para determinar a qualidade de um subconjunto de atributos selecionados. Estes métodos geralmente possuem alta complexidade computacional, principalmente quando o número de atributos é muito alto. No entanto, por utilizarem informações diretamente obtidas dos classificadores, estes métodos tendem a proporcionar melhores resultados do que abordagens usando filtros (Tang, et al., 2014).

Em outra categoria, a seleção de atributos é incorporada diretamente aos algoritmos de aprendizado, sendo que, simultaneamente ao ajuste do modelo, ocorre a seleção de atributos (Quinlan, 1993; Deng & Runger, 2012; Tibshirani, 1996). Algoritmos de indução de regras e árvores de decisão, como RIPPER e C4.5, encaixam-se nesta categoria. A vantagem é que estes métodos conseguem atingir desempenho equiparável aos métodos que utilizam *wrappers* porém com tempo de execução similar aos filtros.

Capítulo 6 - Metodologia a ser adotada com base em ferramentas de aprendizado de máquina

Nas subseções, presentes nesta seção, será apresentada a metodologia utilizada para realização do estudo abordado neste trabalho. Inicialmente, foi abordada a caracterização dos resíduos de aminoácidos catalíticos das diversas sub-subclasses escolhidas, através de regras induzidas a partir de um conjunto de amostras e respectivos descritores estruturais de proteínas sem considerar descritores de conservação da estrutura primária das enzimas (seção 6.1). Adicionalmente, foi realizado o estudo do impacto do número de regras e seus respectivos tamanhos (número de antecedentes) na caracterização dos CSR's via algoritmo multiobjetivo (MOEA-RIPPER).

As capacidades preditivas de regras de classificação foram avaliadas seguindo as abordagens apresentadas nas seções subsequentes (6.2, 6.3, 6.4, 6.5 e 6.6). Inicialmente, procurou-se avaliar o desempenho dos métodos classificadores utilizando as proporções originais entre as classes de resíduos de aminoácidos. Dados os desafios inerentes à classificação em dados desbalanceados, avaliou-se também o impacto de técnicas de sub e sobreamostragem no desempenho dos classificadores. No caso da sobreamostragem via introdução de amostras sintéticas, foi realizada uma pré-seleção de atributos, a fim de reduzir o número de atributos e viabilizar o emprego da técnica SMOTE.

De acordo com estudos encontrados na literatura, técnicas de *ensemble* de classificadores especializadas em problemas desbalanceados mostraram contribuir para o aumento do desempenho das classificações. Por este motivo, avaliou-se também o emprego das técnicas RUSBoost, SMOTEBoost e SMOTEBagging ao problema de classificação de resíduos de aminoácidos catalíticos. Essas abordagens empregam algum tipo de sub ou sobreamostragem durante a construção dos *ensembles* como tentativa de superar os desafios inerentes a classificação de dados desbalanceados.

A introdução de descritores de conservação da estrutura primária das enzimas como conjunto de atributos foi realizada a fim de avaliar o impacto desses descritores. Sendo importantes para classificação de CSR's e largamente utilizados na literatura, no entanto, apresentando as desvantagens já discutidas, comparações entre o desempenho dos classificadores treinados com e sem uso de descritores de conservação foram realizadas.

Por último (seção 6.7), é descrito o método utilizado para construção de um classificador genérico para resíduos de aminoácido catalíticos, ou seja, sem separação das enzimas por sub-subclasses. Neste caso, tem-se o intuito de obter uma comparação com alguns métodos disponíveis na literatura e apresentados no capítulo Capítulo 2 e, avaliar as capacidades preditivas dos descritores presentes no Blue Star STING comparados àqueles utilizados por outros métodos.

6.1 Caracterização de resíduos catalíticos

Motivados por estudos realizados utilizando-se o ¹PD, onde foi possível obter regras descrevendo o nanoambiente dos resíduos de aminoácidos catalíticos para várias famílias proteicas (ver Apêndice D). Inicialmente, buscou-se descrever o nanoambiente de enzimas responsáveis por catalisarem a mesma reação química, utilizando regras construídas a partir dos descritores estruturais do Blue Star STING. Considerando-se, portanto, conjuntos de enzimas com similaridade sequencial máxima de 95%, separadas pelo terceiro nível da hierarquia EC (sub-subclasse), regras foram geradas utilizando-se o algoritmo RIPPER (JRip disponível no *software* Weka (versão 3.7.11) (Hall, et al., 2009)). Para gerar as regras, foram consideradas todas as enzimas de um mesmo subconjunto (mesma sub-subclasse EC), sem uso da fase de poda do algoritmo. Este experimento tem como objetivo fornecer uma descrição do nanoambiente dos resíduos de aminoácidos catalíticos que compõem as enzimas de uma mesma sub-subclasse. A partir das regras obtidas, pode-se estudar quais CSR's possuem características similares que permitem a estes serem agrupados e definidos por uma única regra, envolvendo suas propriedades físicas, químicas e estruturais. Dada a grande quantidade de descritores e as possíveis combinações destes para a construção das regras, encontrar a regra que seja mínima no número de descritores e que maximize o número de resíduos selecionados é um problema combinatório, em que uma abordagem exaustiva é totalmente inviável considerando-se o tempo computacional. Assim, o uso de heurísticas e métodos de busca apropriados para lidar com estas situações são geralmente empregados nestes casos. O algoritmo RIPPER utiliza diversas técnicas de busca local para encontrar uma solução que obedeça essas restrições. No entanto, devido a seu caráter local, não há garantias de que as regras geradas sejam ótimas, no sentido de fornecerem menor número de descritores e maior suporte.

Devido aos objetivos conflitantes (regra mínima vs. máximo suporte), técnicas de otimização multiobjectivo apresentam-se como interessantes candidatas a serem empregadas às buscas, uma vez que, são capazes de fornecer diferentes soluções com diferentes compromissos entre esses dois objetivos conflitantes. Com base nisto, utilizou-se o algoritmo evolutivo multiobjectivo NSGA-II (Deb, et al., 2002) para buscar por soluções (conjuntos de regras) geradas pelo RIPPER de forma a atenderem estes dois objetivos. Usando uma codificação binária representando a escolha de um subconjunto de descritores, onde o valor 1 indica a inclusão do descritor e o valor 0 sua remoção, um esquema de *wrapper* foi empregado para avaliar os indivíduos, segundo o número de regras geradas pelo RIPPER e o desempenho deste conjunto ao ser realizada a redução de dimensionalidade indicada na codificação. A avaliação do desempenho das regras foi mensurada através da medida-F (*F-measure*), por apresentar um compromisso entre sensibilidade e precisão, levando em conta o desbalanceamento.

Ainda que os resultados obtidos por esses processos não forneçam garantias de que as regras encontradas serão capazes de manter sua aplicação a novas enzimas não utilizadas para gerá-las, tem-se um conjunto de regras que podem ser utilizadas para estudos das enzimas envolvidas e suas propriedades estruturais. Entender o mecanismo catalítico das enzimas envolve, além de outros fatores, o estudo destas propriedades que podem, juntamente com as regras, fornecerem explicações sobre o papel e o funcionamento destas enzimas.

6.2 Predição de resíduos catalíticos e subamostragem aleatória

Nesta etapa, procurou-se a construção de modelos classificadores que pudessem ser empregados para a predição de resíduos de aminoácidos catalíticos em outras enzimas. Ainda que seja difícil de avaliar e garantir o mesmo desempenho obtido no treinamento dos classificadores, existem formas de se estimar a capacidade destes classificadores em generalizar a hipótese construída. Uma estimativa para erro e desempenho pode ser obtida realizando-se uma validação cruzada durante o treinamento dos classificadores. Uma validação cruzada com k pastas (*k-fold cross validation*) consiste em dividir aleatoriamente o conjunto de dados em k subconjuntos (pastas) balanceados e treinar igual número de classificadores, tendo como conjunto de validação um dos k subconjuntos, de forma que cada subconjunto seja utilizado uma única vez como validação. O desempenho obtido pelos k classificadores no conjunto de validação pode ser então utilizado para avaliar a capacidade de generalização, ao serem empregados em novos conjuntos de dados. Valores de média e desvio padrão fornecem indicativos do desempenho dos classificadores e dependência destes ao conjunto de treinamento, sendo estes valores comumente reportados como resultado da validação cruzada. Neste trabalho, empregou-se validação cruzada 2x5, ou seja, uma validação cruzada com 5 pastas ($k = 5$) é repetida duas vezes distribuindo-se as amostras aleatoriamente em cada subconjunto, sendo reportados os valores médios e desvio padrões para os 10 classificadores treinados.

Inicialmente, a construção dos classificadores para a predição de resíduos catalíticos foi realizada utilizando-se os algoritmos RIPPER e C4.5 (JRip e J48 disponibilizados pelo *software* Weka (Hall, et al., 2009)) sem pré-processamento dos dados, ou seja, todas as amostras negativas e positivas foram consideradas para a validação cruzada 2x5. Posteriormente, avaliou-se também o impacto da subamostragem aleatória da classe negativa no desempenho dos classificadores, considerando-se diferentes proporções entre amostras positivas e negativas nos conjuntos de treinamento. As proporções escolhidas foram definidas com base em experimentos prévios realizados, garantindo redução de amostras e variação no desempenho, e obtiveram as proporções de 1 CSR para cada 1 não CSR (1:1), 1 CSR para cada 2 não CSR (1:2), 1 CSR para cada 6 não CSR (1:6), 1 CSR para cada 10 não CSR (1:10),

1 CSR para cada 25 não CSR (1:25) e 1 CSR para cada 50 não CSR (1:50).

6.3 Seleção de atributos e redução de dimensionalidade

Além do desbalanceamento, a alta dimensionalidade dos dados pode trazer dificuldades para os métodos de aprendizado de máquina, principalmente para métodos que realizam uma seleção interna dos atributos, como no caso do RIPPER e C4.5. Pelo fato destes algoritmos utilizarem a busca pelo melhor atributo e seu respectivo valor iterativamente, tendo-se em vista somente a maximização da métrica adotada de forma local, estes são susceptíveis a encontrar soluções ótimas locais. Neste processo, atributos podem apresentar valores próximos para a métrica considerada, mas devido ao caráter guloso destes algoritmos, somente uma delas (a de maior valor) será considerada. No entanto, é possível que outros atributos retornem melhores soluções, uma vez que não fornecem valores tão distantes do atributo escolhido. Assim, a seleção de atributos como pré-processamento pode trazer ganhos a predição, além de reduzir o tempo computacional necessário para o treinamento dos classificadores.

Dentre as diversas técnicas para seleção de atributos e redução de dimensionalidade, é difícil saber qual delas resultará em ganhos expressivos. No entanto, a utilização de técnicas que envolve a avaliação individual dos atributos pouco pode influenciar, devido a semelhanças com as técnicas utilizadas internamente pelos algoritmos RIPPER e C4.5. Uma técnica muito empregada para seleção de atributos, fornecendo boa redução da dimensionalidade e manutenção ou aumento do desempenho, é a busca sequencial. Assim optou-se por utilizar a busca sequencial incremental (do inglês *forward feature selection*). Partindo-se de um conjunto inicialmente vazio de atributos, a busca é feita de forma a encontrar o atributo que, ao ser adicionado ao conjunto de atributos selecionados, resulta em um maior ganho de desempenho. A busca sequencial incremental, ainda que não seja uma busca global pelo melhor subconjunto de atributos, é computacionalmente mais barata que o método decremental, também não global, onde o processo inverso é aplicado, ou seja, parte-se de um conjunto inicial contendo todos os atributos. A seleção de atributos e o posterior treinamento dos classificadores foram realizadas dividindo-se o conjunto de dados em dois subconjuntos: 30% das amostras foram utilizadas para a seleção de atributos, de forma que foi escolhido o subconjunto que resultou em um maior valor para medida-F, quando aplicado ao treinamento do RIPPER. As amostras restantes (70%) foram utilizadas para treinar os classificadores e realizar a validação cruzada, após a redução de dimensionalidade.

6.4 Sobreamostragem da classe minoritária

A sobreamostragem da classe minoritária é também empregada em problemas desbalanceados, como tentativa de reduzir o desbalanceamento e aumentar o desempenho dos classificadores. Técnicas

de sobreamostragem aleatória consistem na replicação das amostras da classe minoritária, sem a introdução de “novas” informações ao processo de aprendizado. Para avaliar o impacto da sobreamostragem aleatória, utilizaram-se diferentes taxas de amostragem de forma a obter as proporções de 1:1, 1:2, 1:6, 1:10, 1:25 e 1:50 entre as classes. Novamente, o algoritmo RIPPER foi utilizado para a indução de regras e a avaliação do desempenho foi medida através da medida-F.

Outras técnicas de sobreamostragem envolvem a determinação de uma vizinhança das amostras, como é o caso do SMOTE. Para criar novas amostras da classe minoritária, são considerados os k vizinhos mais próximos de uma amostra também da classe minoritária. A nova amostra é então criada em uma posição intermediária entre as amostras vizinhas. No entanto, em espaços de alta dimensão, esta noção de vizinhança pode se tornar ineficaz (Marimont & Shapiro, 1979; Chávez, et al., 2001). À medida que a dimensionalidade aumenta, o volume do espaço amostral aumenta tão rapidamente que os dados disponíveis tornam-se esparsos. A busca por vizinhos de uma amostra geralmente concentra-se em detectar áreas onde as amostras formam grupos com propriedades similares. Quando o número de atributos é alto, todas as amostras tornam-se esparsas e dissimilares em muitos sentidos, o que leva a busca por vizinhos a ser menos eficiente. Assim, a eficiência destas técnicas é comprometida e as novas amostras criadas tendem a gerar um espaço de busca ainda mais problemático, dificultando o aprendizado de regras. No entanto, a redução de dimensionalidade através dos métodos discutidos na seção anterior pode trazer ganhos expressivos, quando combinados a tais técnicas.

Combinando redução de dimensionalidade e sobreamostragem utilizando o algoritmo SMOTE, classificadores foram treinados. Com a dimensionalidade reduzida, o algoritmo SMOTE torna-se mais eficaz e produz amostras sintéticas que podem auxiliar o aprendizado (Blagus & Lusa, 2013).

6.5 Comitês de classificadores – Ensembles

Comitês de classificadores, conhecidos também como *ensembles* de classificadores, são utilizados para agregar predições de diversos classificadores e fornecer melhorias nas predições. *Ensembles* podem ser compostos de classificadores de tipos diversos ou ainda treinados utilizando-se porções diversas dos dados. Para o caso de problemas desbalanceados, os *ensembles* são construídos utilizando-se classificadores treinados em subconjuntos balanceados. Explorou-se três algoritmos para a construção de *ensembles*, discutidos na seção 5.2, sendo um para subamostragem da classe majoritária (RUSBoost) e dois para sobreamostragem da classe minoritária (SMOTEBoost e SMOTEBagging). Todos os *ensembles* são compostos de 10 classificadores, e nos casos em que utilizou-se SMOTE para sobreamostragem da classe minoritária, a seleção de atributos foi previamente realizada para evitar problemas no cálculo de distâncias Euclidianas em espaços de alta dimensão.

À medida que o número de classificadores no *ensemble* aumenta, tem-se uma redução da interpretabilidade dos modelos. Uma vez que estes são agregados segundo seus pesos, a interpretação das regras geradas torna-se mais difícil e menos determinística, ao ponto que mais de uma regra pode ser utilizada para classificar uma amostra, resultando em previsões ambíguas.

6.6 Incorporando Descritores de Conservação

Descritores de conservação de estrutura primária configuram-se como um dos mais poderosos descritores para a predição de resíduos de aminoácidos catalíticos (Barlett, et al., 2002). Devido à importância dos CSR's, estes possuem grande tendência a serem preservados em enzimas similares, e por isso, são ditos mais conservados do que os demais resíduos. No entanto, estas propriedades são derivadas a partir de alinhamentos sequenciais múltiplos e envolvem características compartilhadas por um conjunto de enzimas, e não de uma enzima em particular, sendo portanto impossível obter estas informações para enzimas que não possuem sequências homólogas conhecidas, chamadas de órfãs.

Mesmo assim, estes descritores são amplamente utilizados e sua introdução foi realizada para uma comparação com os classificadores treinados sem descritores de conservação. Através desta comparação, pode-se evidenciar se somente os descritores estruturais são capazes de produzir resultados iguais ou similares aos obtidos com a adição de descritores de conservação, possibilitando assim realizar a predição de resíduos de aminoácidos catalíticos também em enzimas órfãs, uma vez que não é necessária a presença de homólogos para a derivação dos descritores estruturais, bastando se ter a estrutura resolvida da enzima em questão.

Dessa forma, os classificadores empregados na seção 6.3 foram novamente treinados considerando-se também o uso de descritores de conservação, além de descritores estruturais. Da mesma maneira, foi realizada a seleção prévia dos atributos utilizando-se o método de busca sequencial em 30% das amostras, com o restante das amostras utilizado para treinamento e validação dos modelos criados. Utilizando validação cruzada 2x5, estimou-se o desempenho dos classificadores e uma comparação com os classificadores anteriormente construídos foi realizada, visando compreender melhor o impacto de descritores de conservação na predição de resíduos de aminoácidos catalíticos.

6.7 Modelos classificadores gerais

Uma comparação entre os classificadores construídos a partir da separação das enzimas segundo sub-subclasses EC com outros métodos disponíveis na literatura é impraticável. Devido à impossibilidade de construção de classificadores para todas as sub-subclasses EC, dada a falta de dados e as diferentes distribuições para cada sub-subclasse, uma comparação direta fica restrita somente aos casos em que foi

possível obter classificadores. No entanto, para avaliar as capacidades preditivas dos descritores estruturais de proteínas disponíveis no Blue Star STING, classificadores gerais foram construídos.

Utilizando todas as enzimas não redundantes em 40% de identidade sequencial, nenhuma separação por números EC foi realizada. Estes dados foram então utilizados para treinamento e validação de Máquinas de Vetores Suporte (SVM). O mesmo procedimento descrito em outros trabalhos foi empregado durante as comparações, utilizando os conjuntos de enzimas fornecidos por cada método. Realizou-se a comparação com três métodos de predição de CSR's que utilizam informações estruturais (EXIA, POOL e o classificador proposto por CILIA & PASSERINI (2010)), e um método que utiliza informações provenientes somente das sequências de aminoácidos das enzimas (CRPred), descritos nas seções 2.2 e 2.3.

Capítulo 7 - Resultados e Discussões

A aplicação da metodologia discutida no 0 levou à compreensão dos desafios e dificuldades apresentados durante o treinamento dos modelos classificadores. Também foi possível um estudo amplo do impacto do número de amostras de resíduos de aminoácidos catalíticos na qualidade dos classificadores gerados e na composição das regras segundo tamanho (número de antecedentes) e especialização. Nas seções seguintes, são apresentados os resultados, bem como uma discussão destes resultados, a fim de elucidar e contribuir para um melhor entendimento do comportamento dos resíduos de aminoácidos catalíticos quando abordados em um problema de classificação em aprendizado de máquina.

Na Seção 7.1, é apresentado o método “manual” de aplicação de filtros utilizando o *Java Protein Dossier* (JPD) para a criação de regras que selecionem somente os resíduos de aminoácidos catalíticos para algumas poucas enzimas. Este processo manual foi uma das motivações para elaboração e construção da hipótese principal deste trabalho. Na Seção 7.2, é introduzida a ferramenta implementada neste trabalho para indução de regras automaticamente a partir do JPD.

Na Seção 7.3, uma análise estatística dos dados comparando as duas classes de resíduos estudadas foi realizada a fim para levantamento de diferenças e características próprias e mais discriminantes dos resíduos de aminoácidos catalíticos.

Na Seção 7.4, o algoritmo RIPPER foi utilizado com o intuito de obter regras para todo o conjunto de amostras, como forma de proporcionar um entendimento melhor do comportamento dos conjuntos de regras encontrados pelo algoritmo e o impacto de alguns parâmetros do algoritmo.

Nas Seções 7.5, 7.6, 7.7, 7.8 e 7.9, é discutido o resultado dos modelos classificadores gerados utilizando-se então uma abordagem de validação cruzada para estimativa do desempenho destes modelos. Por fim, na Seção 7.10, é discutido melhor o impacto do desbalanceamento e os desafios encontrados durante a construção dos modelos classificadores. Como forma de amenizar o impacto do número de amostras positivas (catalíticos) na construção dos classificadores, na Seção 7.11, é apresentado o resultado do processo de transferência de anotações e o impacto disto nos conjuntos de dados. Seguindo então outros métodos de predição de CSR's encontrados na literatura, na Seção 7.12, são apresentados e discutidos os resultados obtidos com a construção de um modelo classificador genérico, incluindo sua comparação com outros métodos encontrados na literatura.

7.1 Seleção de resíduos catalíticos utilizando *Java Protein Dossier* (JPD)

Descrito na Seção 3.4, o módulo *Java Protein Dossier* presente na plataforma Blue Star STING pode

ser utilizado para visualizar os descritores físico-químicos e estruturais de proteínas. Além disto, o módulo ^{Java} Protein Dossier é uma ferramenta muito poderosa para selecionar resíduos que atendam condições específicas impostas aos descritores físico-químicos e estruturais através de seus valores, com uso da ferramenta para aplicação de filtros (*JPD Select*).

Dessa forma, pode-se utilizar o ^JPD para construir filtros que selecionam apenas os resíduos de aminoácidos catalíticos de diversas enzimas. Estes filtros podem então ser convertidos em regras na forma:

$$\text{Se } A_1 \text{ AND } A_2 \text{ AND } \dots \text{ AND } A_k \rightarrow C \quad \text{Equação 7-1}$$

onde A_k representa um antecedente da regra, composto por um descritor e seu respectivo valor (ou intervalo de valores). Por exemplo, pode-se ter *número de contatos hidrofóbicos* ≥ 5.0 e ainda *potencial eletrostático no LHA* ≤ -100 . O conseqüente da regra (C) representa uma das classes. Como se está interessado em somente obter regras para os resíduos catalíticos, tem-se neste caso $C = \text{CSR}$.

Dessa forma, podem-se construir diversos filtros e aplicá-los às estruturas membros de um mesmo nível hierárquico na nomenclatura EC, visando obter uma regra que seja válida para vários membros desta mesma classe EC, definindo assim um padrão do nanoambiente dos resíduos de aminoácidos catalíticos para todo um conjunto de enzimas que catalisam uma mesma reação química. Utilizando a ferramenta *JPD Select* é possível encontrar regras que selecionam diversos outros resíduos de interesse, como *ligand pocket residues* e resíduos formadores de interface. Essa poderosa ferramenta permite obter uma descrição do nanoambiente, através de propriedades físico-químicas e estruturais, dos resíduos de aminoácidos importantes para a funcionalidade das proteínas de diversas regiões da molécula.

O módulo ^JPD foi um dos grandes motivadores deste trabalho, pois utilizando tais filtros pôde-se constatar a existência de padrões nos resíduos catalíticos para diversas famílias de enzimas, como será ilustrado mais adiante. No entanto, é bastante difícil obter regras, através de tentativa e erro, que cobrem os resíduos catalíticos de diversas enzimas. Basicamente, o método “manual” (sem considerar métodos computacionais de reconhecimento de padrões), consiste em obter regras que satisfaçam a condição de selecionar apenas os CSR’s para uma estrutura e, então, aplicá-la a outros membros da mesma família e verificar sua validade. Caso a regra não seja válida para outros membros, uma nova regra deve ser encontrada por tentativa e erro, com base na suposição de que tal regra exista.

Para encontrar as regras, utiliza-se outra ferramenta do ^JPD que retorna os valores máximos e mínimos de cada descritor. Dessa forma, comparam-se os valores de máximo e mínimo de cada descritor, entre os resíduos de aminoácidos catalíticos e não catalíticos, e escolhe-se de forma empírica (por tentativa e erro) aqueles que supostamente são mais discriminantes para selecionar apenas os CSR’s.

Repetindo este processo até que somente os resíduos de aminoácidos catalíticos sejam selecionados pelo filtro, obtém-se uma regra que pode ser aplicada a outros membros da família, de forma a verificar seu poder de cobertura.

É evidente que este processo manual está sujeito a muita subjetividade e é inviável de se realizar para todos os membros da mesma sub-subclasse EC. Assim, a utilização de métodos computacionais para extração de regras que cobrem o máximo possível de membros de uma sub-subclasse foi o principal objetivo deste trabalho. A seguir, apresentam-se alguns exemplos das regras encontradas, utilizando este processo manual, para as enzimas que catalisam a reação de hidrólise de proteínas semelhantes à tripsina (*trypsin-like*) (EC 3.4.21.81 – Estreptogrisina B); para as Poligalacturonases (EC. 3.2.1.15), enzimas que catalisam a reação de hidrólise de polissacarídeos presentes na parede celular de plantas, através da quebra de ligações glicosídicas que ligam resíduos de ácido galacturônico; e para as Metaloendopeptidases, que catalisam a proteólise de ligações peptídicas de aminoácidos não terminais, envolvendo um metal em seu mecanismo catalítico, tipicamente Zinco ou Cobalto (EC 3.4.24).

Os três exemplos ilustrados nas Figuras 7-1, 7-2 e 7-3 mostram diferentes abordagens e abrangências das regras, sendo obtidas regras para caracterizar enzimas com um mesmo número EC (3.4.21.81); enzimas com dois números EC diferentes, mas que catalisam a mesma reação para substratos diferentes (3.2.1.15 e 3.2.1.171); e enzimas que catalisam uma reação em geral, sem considerar a especificidade do substrato (EC 3.4.24), respectivamente.

Nas Tabelas 7-1, 7-2 e 7-3 são mostrados os alinhamentos múltiplos sequenciais e estruturais para as enzimas ilustradas nas Figuras 7-1, 7-2 e 7-3, respectivamente, bem como a raiz do erro quadrático médio (RMSD do inglês *Root-Mean-Square Deviation*) das posições atômicas e a medida de *strict core* (sobreposição das cadeias principais) obtidos durante o alinhamento múltiplo estrutural da cadeia principal, utilizando o *software* MAMMOTH (Lupyan, et al., 2005).

Para as Estreptogrisinas B (Figura 7-1) são apresentadas cinco estruturas da mesma enzima de *Streptomyces griseus*. Por se tratarem de estruturas de uma mesma enzima, a regra encontrada facilmente cobre os seus CSR's, uma vez que os descritores físico-químicos e estruturais que descrevem o nanoambiente de seus resíduos de aminoácidos apresentam valores muito similares, se não idênticos. Já no caso das Poligalacturonases (Figura 7-2) três estruturas, obtidas através de cristalografia por difração de raios-X, de enzimas provenientes de organismos diferentes: *Fusarium verticillioides* (PDB: 1HG8, resolução: 1.73Å), *Chondrostereum purpureum* (PDB: 1K5C, resolução: 0.96Å) e *Pectobacterium carotovorum subsp. carotovorum* (PDB: 1BHE, resolução: 1.90Å) são apresentadas. Neste caso, observa-se uma baixa similaridade sequencial (máxima de ~40%), ainda que a similaridade estrutural seja elevada (RMSD = 0.85Å e cadeias principais com mais de 80% de sobreposição). Mesmo assim as regras são

capazes de selecionar os resíduos de aminoácidos catalíticos das três enzimas apresentadas, ilustrando a dependência das propriedades estruturais para manutenção de um nanoambiente favorável aos CSR's. O caso das Metaloproteases (Figura 7-3) cinco estruturas, obtidas através de cristalografia por difração de raios-X, de cinco enzimas de diferentes organismos são consideradas: *Crotalus adamanteus* (PDB: 2AIG, resolução: 2.60Å), *Bothrops asper* (PDB: 1ND1, resolução: 1.93Å), *Crotalus atrox* (PDB: 2DW2, resolução: 2.70Å), *Homo sapiens* (PDB's: 1R54 e 1A85, respectivamente com resoluções: 1.85Å e 2.00Å). Esse é um intermediário entre os outros dois casos, pois, com uma similaridade sequencial moderada entre as enzimas (~71%) e, uma ligeira diferença no RMSD em relação ao apresentado pelas Poligalacturonases, a eficácia das regras demonstra a princípio uma relação das regras com os fatores estruturais, muito maior do que com os fatores sequenciais. Ou seja, os fatores que definem os CSR's estão relacionados em um primeiro momento com as conformações tridimensionais das enzimas, em maior grau do que com as sequências de aminoácidos que estas apresentam.

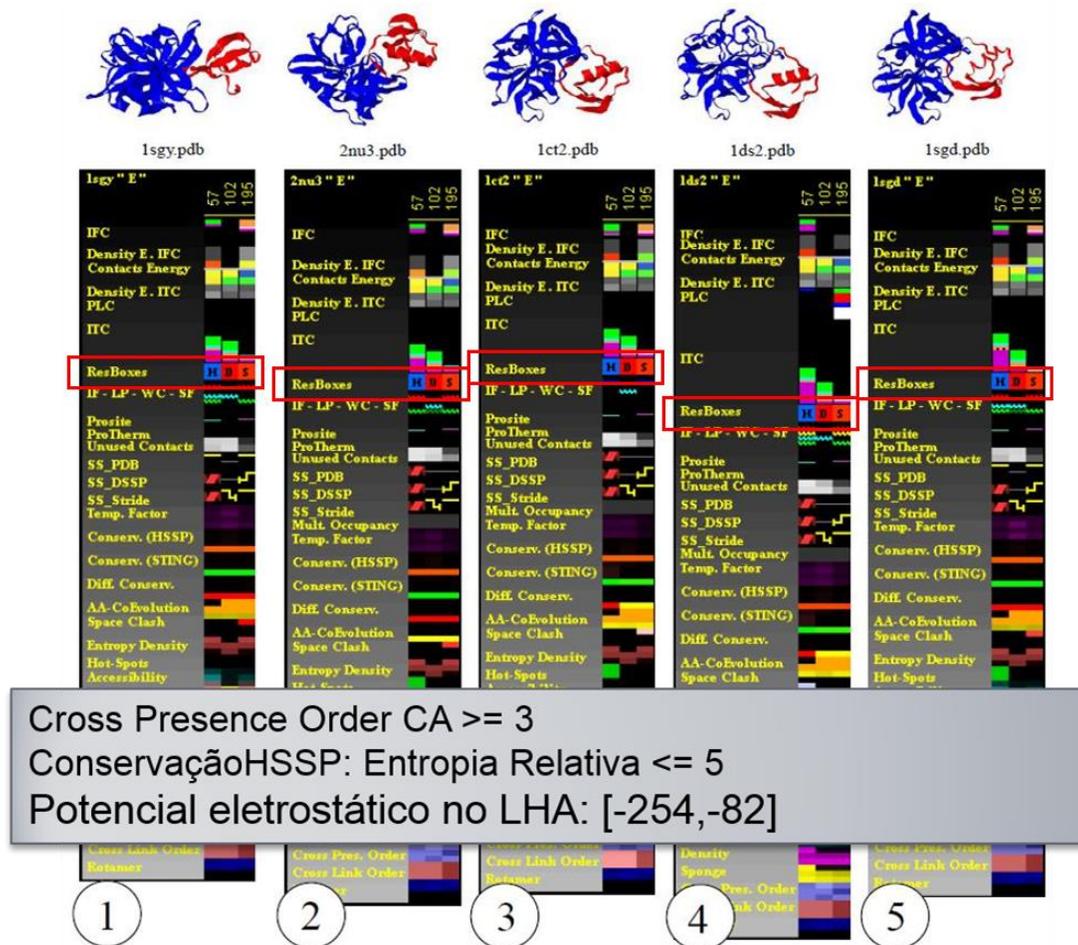


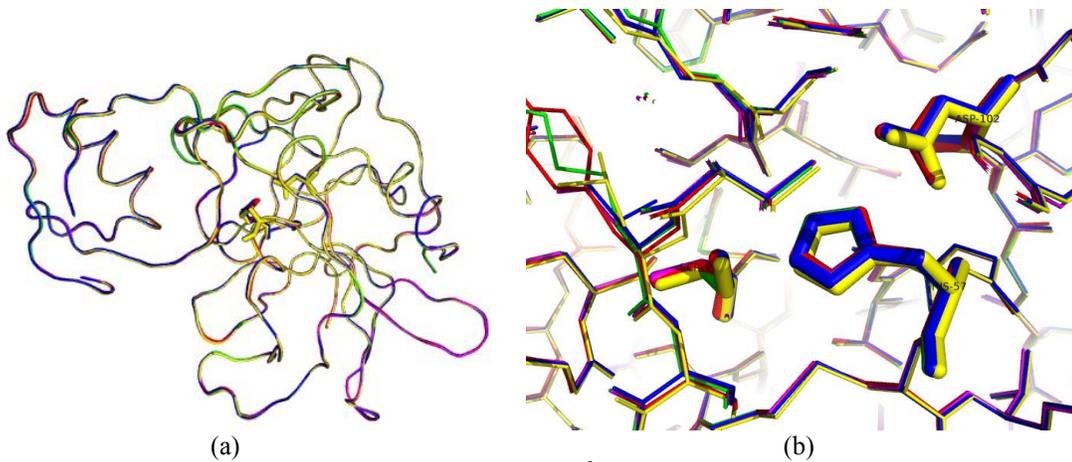
Figura 7-1: Regra encontrada utilizando o JPD para cinco estruturas da enzima Estreptogrisina B de *Streptomyces griseus* (EC: 3.4.21.81). A regra encontrada utilizando o JPD (quadro sobrepondo a imagem do JPD , em cinza), compostas de três antecedentes (uma em cada linha), seleciona apenas os resíduos catalíticos para as cinco estruturas mostradas na figura. (PDB: 1SGY:E, 2NU3:E, 1CT2:E, 1DS2:E e 1SGD:E). Em destaque (quadro vermelho) os códigos dos resíduos selecionados pelas regras no JPD .

Tabela 7-1: Alinhamentos múltiplos sequencial e estrutural para as enzimas da Figura 7-1. a) alinhamento estrutural com alta sobreposição entre as estruturas; b) em destaque (sticks) CSR's sobrepostos: HIS, ASP e SER.

```

1SGY_E ISGGDAIYSSTGRCSLGFNVRSGSTYYFLTAGHCTDGATTWWANSARTTV 50
2NU3_E ISGGDAIYSSTGRCSLGFNVRSGSTYYFLTAGHCTDGATTWWANSARTTV 50
1SGD_E ISGGDAIYSSTGRCSLGFNVRSGSTYYFLTAGHCTDGATTWWANSARTTV 50
1CT2_E ISGGDAIYSSTGRCSLGFNVRSGSTYYFLTAGHCTDGATTWWANSARTTV 50
1DS2_E ISGGDAIYSSTGRCSLGFNVRSGSTYYFLTAGHCTDGATTWWANSARTTV 50
*****
1SGY_E LGTTSGSSFPNNDYGIVRYTNTTIPKDGTVGGQDITSAANATVGMVTRR 100
2NU3_E LGTTSGSSFPNNDYGIVRYTNTTIPKDGTVGGQDITSAANATVGMVTRR 100
1SGD_E LGTTSGSSFPNNDYGIVRYTNTTIPKDGTVGGQDITSAANATVGMVTRR 100
1CT2_E LGTTSGSSFPNNDYGIVRYTNTTIPKDGTVGGQDITSAANATVGMVTRR 100
1DS2_E LGTTSGSSFPNNDYGIVRYTNTTIPKDGTVGGQDITSAANATVGMVTRR 100
*****
1SGY_E GSTTGTHSGSVTALNATVNYGGDVVYGMIRTNVCAEPGDSGGPLYSGTR 150
2NU3_E GSTTGTHSGSVTALNATVNYGGDVVYGMIRTNVCAEPGDSGGPLYSGTR 150
1SGD_E GSTTGTHSGSVTALNATVNYGGDVVYGMIRTNVCAEPGDSGGPLYSGTR 150
1CT2_E GSTTGTHSGSVTALNATVNYGGDVVYGMIRTNVCAEPGDSGGPLYSGTR 150
1DS2_E GSTTGTHSGSVTALNATVNYGGDVVYGMIRTNVCAEPGDSGGPLYSGTR 150
*****
1SGY_E AIGLTSGGSGNCSSGGTFFQPVTEALSAYGVS VY 185
2NU3_E AIGLTSGGSGNCSSGGTFFQPVTEALSAYGVS VY 185
1SGD_E AIGLTSGGSGNCSSGGTFFQPVTEALSAYGVS VY 185
1CT2_E AIGLTSGGSGNCSSGGTFFQPVTEALVAYGVS VY 185
1DS2_E AIGLTSGGSGNCSSGGTFFQPVTEALVAYGVS VY 185
*****

```



RMSD = 0.07Å
 STRICT CORE = 100%
 Similaridade sequencial máxima = 100%



Energia de contatos não usados ≥ 997
 Num. de contatos carregados repulsivos ≥ 2
 Energia total de contatos ≥ 95
 Densidade energética de contatos ($C\alpha$ raio=3Å) ≥ 47

Figura 7-2: Regra encontrada utilizando o ^JPD (quadro em cinza) para enzimas Poligalacturonase (EC: 3.2.1.15) de *Fusarium verticillioides* (PDB: 1HG8:A), *Chondrostereum purpureum* (PDB: 1K5C:A) e *Pectobacterium carotovorum* subsp. *carotovorum* (PDB: 1BHE:A). Regra seleciona o grupo de resíduos de aminoácidos catalíticos formado por três ácidos aspárticos e uma histidina nas três enzimas.

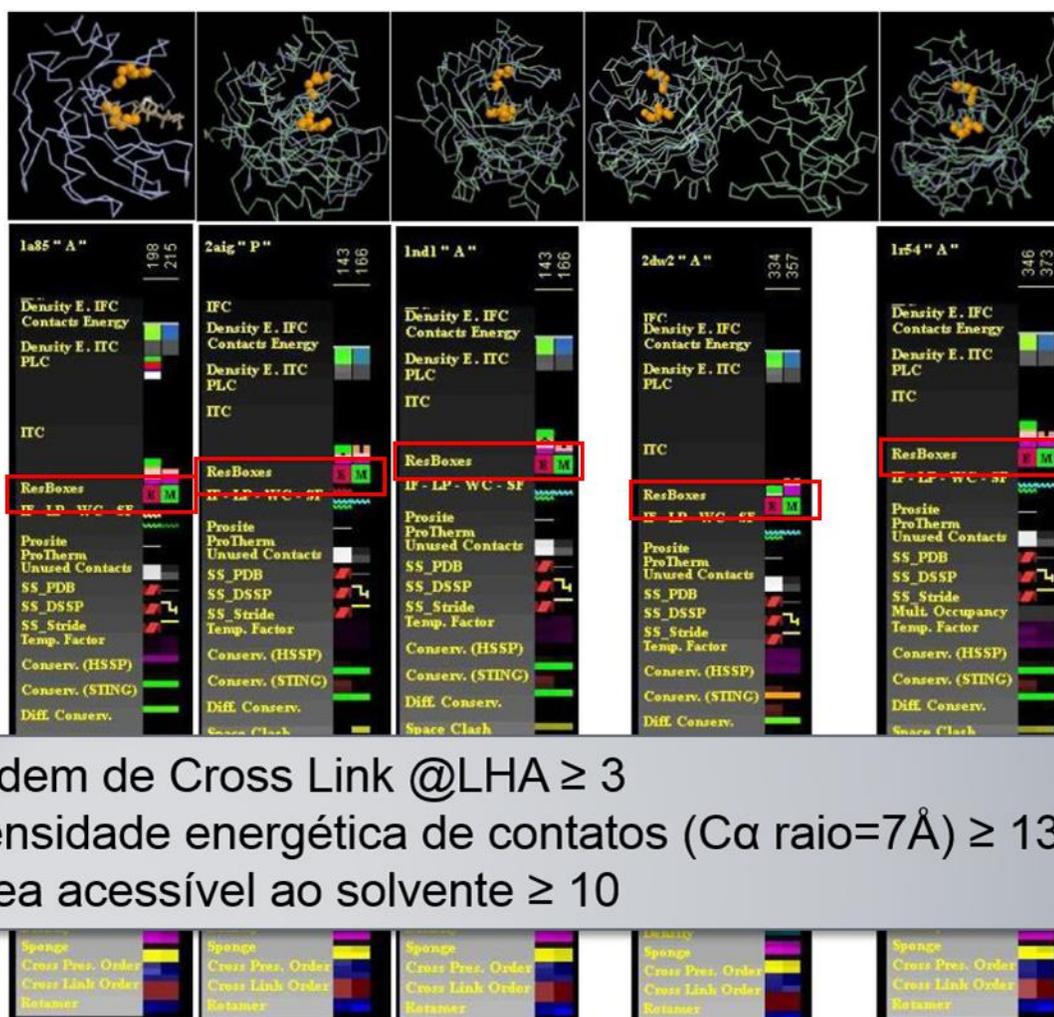
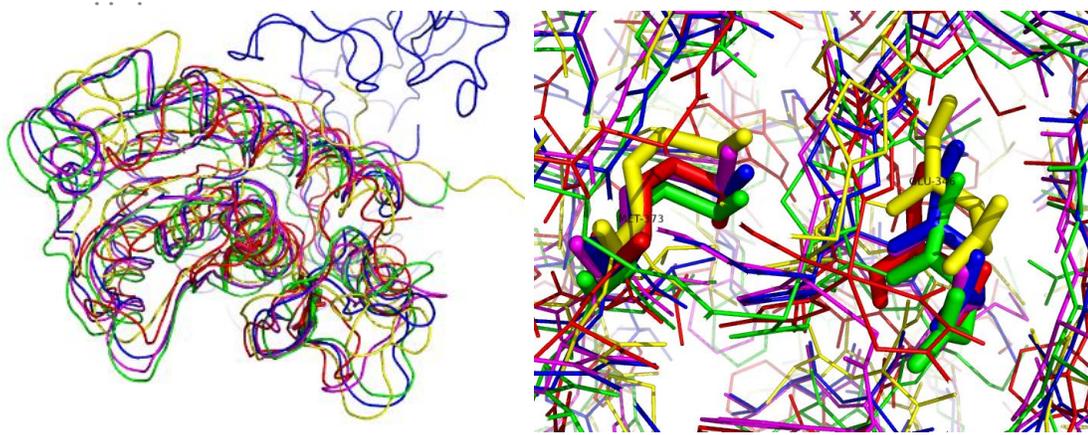


Figura 7-3: Regra encontrada utilizando o JPD para cinco Metaloproteases. A regra seleciona a díade catalítica composta pelos ácido glutâmico e metionina para as enzimas com códigos no PDB: 1a85A (EC 3.4.24.34), 2aigP (EC 3.4.24.46), 1nd1A (EC 3.4.24), 2dw2A (EC 3.4.24) e 1r54A (EC 3.4.24).

Tabela 7-3: Alinhamentos múltiplos sequencial e estrutural para as enzimas da Figura 7-3. a) alinhamento estrutural com boa sobreposição entre as estruturas, com pequenas divergências em regiões mais flexíveis (laços); b) em destaque (sticks) CSR's sobrepostos: GLU e MET, apresentam relativa diferença em suas posições.

```

2AIG_P --XNLPQRYIELVVVADRRVFMKYNSDLNIIIRTVHEIVNIINEFYRSLN 48
1ND1_A --ERFSPRYIELAVVADHGIFTKYNSNLNITRTRVHEMLNTVNGFYRSVD 48
2DW2_A HQKYNPFRFVELVLVVDKAMVTKNGDLDKIKTRMYEIVNTVNEIYRYMY 50
1R54_A -EARRTRKYLELYIVADHTLFLTRHRNLQHTKQRLLEVANYVDQLLRITLD 49
1A85_A -----NPKWERTN-LTYRIRNYTPQLSEAEV 25
      . . . . .
      . . . . .
2AIG_P IRVSLTDLLEIWSGQDFITIQSSSSNTLNSFGEWRRVLLTRKRHDNAQLL 98
1ND1_A VHAPLANLEVWSKQDLIKVQKDSSKTLKSGEWRRDLLPRISHDHAQLL 98
2DW2_A IHVALVGLEIWSNEDKITVKPEAGYTLNAFGEWRKTDLLTRKKHDNAQLL 100
1R54_A IQVALTGLEVWTERDRSRVTQDANATLWAFLOWR-RGLWAQRPHDSAQLL 98
1A85_A ERAIKDAFELWSVASPLIFTRISQGEADINIAFYQRDHGDNSPFDGPNGI 75
      . . . . .
      . . . . .
2AIG_P TAINFEGKIIIGKAYTSSMCNPRSSVGIVKDHSPINLLVAVTMAHELGHNL 148
1ND1_A TAVVFDGNTIGRAYTGGMCDPRHSVGVVVDHDSKNNLWVAVTMAHELGHNL 148
2DW2_A TAIDLD-RVIGLAYVGSMPKPKRSTGI IQDYSEINLVAVIMAHEMGHNL 149
1R54_A TGRAFQGATVGLAPVEGMCRAESSGGVSTDHSELPIGAAATMAHEIGHSL 148
1A85_A LAHAFQ---PGQGIGG---DAHFDAEETWNTSANYNLFVAAHEFGHSL 119
      . . . . .
      . . . . .
2AIG_P GMEHDGKDCLRGASL----CIMRPGLTPGRSYEFSDDSMGYQKFLNQYK 194
1ND1_A GIDHDTGSCSCGAKS----CIMASVLSKVLSEYFSDCSQNYETYLTNHN 194
2DW2_A GINHDSGYCSCGDYA----CIMRPEISPEPSTFFSNCSYFECWDFIMNHN 195
1R54_A GLSHDPDGCCVEAAAEESGGCVMAAATGHPFPRVFSACSRRLRAFFRKG 198
1A85_A GLAHSSD-----PGALMPNYAFRETSNYSLPPQDDIDG- 152
      * : *
      . . . . .
2AIG_P PQCILNKP----- 202
1ND1_A PQCILNKP----- 202
2DW2_A PECILNEPLGTDIISPPVCGNELLEVGEECDGTPENCQNECCDAATCKL 245
1R54_A GACLSNAPSGHHHHHH----- 214
1A85_A IQAIYG----- 158
      . . .
  
```



(a) (b)

RMSD = 0.95
 STRICT CORE = 71.52%
 Similaridade sequencial máxima: 71.8%

Como forma de ilustrar a capacidade e importância do método proposto utilizando o ¹PD é possível obter a regra da Tabela 7-4, a qual seleciona somente os resíduos catalíticos da enzima HIV-1 Integrase (PDB: 1biu). HIV-1 Integrase é uma enzima essencial no ciclo de vida do vírus HIV, responsável pela catálise da inserção do genoma viral no cromossomo da célula hospedeira, o que fornece um atrativo alvo para projeto de novas drogas antivirais.

Tabela 7-4: Regra descrevendo o nanoambiente dos resíduos catalíticos para a HIV-1 Integrase.

Pocket	Volume (#volumes de água) > 0
Potencial Eletrostático	Médio: [-300, -5]
Acessibilidade	Isolação: ≤ 28

É interessante, neste ponto, comentar o significado destes descritores que podem caracterizar o nanoambiente criado pelos resíduos de aminoácidos pertencentes ao conjunto de resíduos catalíticos da HIV-1 Integrase. Fica claro que se trata de um ambiente com potencial eletrostático negativo (<-5 kT/J/mol, atraente para íons positivos), de um ambiente que geometricamente apresenta uma invaginação na superfície da proteína (*pocket* com volume > 0), adequado para acolher um íon positivo, por exemplo. Em adição, estes resíduos estão parcialmente acessíveis ao solvente (área acessível ao solvente ≤ 28%), uma vez que estes se encontram em um *pocket* na superfície da molécula.

Com esta descrição do nanoambiente dos resíduos catalíticos da HIV-1 Integrase, fica fácil entender o sentido dos descritores envolvidos na seleção dos CSR's, uma vez que experimentos relatados e descritos na literatura demonstram que esta enzima está abrigando um íon positivo de magnésio (Mg²⁺) (Goldgur, et al., 1998).

Apesar das estruturas apresentadas nos exemplos acima possuírem alta similaridade sequencial e estrutural, estes experimentos simples motivam a procura por casos similares (regras simples que podem ser usadas para a descrição do nanoambiente dos CSR's, que permitam uma interpretação fácil do seu significado biológico) em outras enzimas.

No Apêndice D, encontra-se uma tabela com os resultados do processo manual aplicado para diversas enzimas utilizando-se o ^JPD, obtidos durante análises nos anos de 2008 e 2009. O objetivo é exemplificar de forma simples a prova de conceito que foi usada para iniciar este trabalho.

7.2 Extensão do ^JPD para a busca de regras

Baseando-se nos resultados obtidos pelo processo manual de extração de regras, uma extensão para o módulo ^JPD foi criada para gerar automaticamente regras para um conjunto de resíduos selecionados. Os resíduos selecionados através da interface do ^JPD são rotulados como sendo da classe positiva e os demais resíduos (não selecionados) como pertencentes à classe negativa. Realiza-se a escolha de quais descritores serão utilizados como entrada para o algoritmo RIPPER (sem fase de poda), sendo possível inclusive omitir o uso de descritores de conservação. O conjunto de regras retornado pelo algoritmo é aplicado à proteína e o resultado é o mesmo de quando se utiliza o processo manual descrito na seção anterior. No entanto, esta extensão não requer nenhuma intervenção do usuário (exceto escolha dos

resíduos rotulados como positivos) e permite obter regras simples que podem ser utilizadas para descrever o nanoambiente dos resíduos de interesse. Esta ferramenta não se limita somente aos CSR's, pois qualquer resíduo da proteína pode ser escolhido como pertencente à classe positiva e, então, regras são criadas para selecionar estes resíduos, como por exemplo resíduos de interface, resíduos que possuem interações com um ligante, etc. A Figura 7-4 ilustra o uso desta ferramenta para a enzima Poligalacturonase de *Erwinia carotovora subsp. carotovora*, enzima ligada ao processo de nutrição da bactéria através da catalisação da reação de quebra do ácido galacturônico, presente nas paredes celulares das plantas, conseqüentemente com excreção de enzimas digestivas na célula hospedeira. As regras sugerem um nanoambiente catalítico com potencial eletrostático bastante negativo e com alta densidade de energia de contatos, devido majoritariamente a ligações de hidrogênio. Uma busca na literatura evidencia a alta eletronegatividade de seu sítio ativo e grande número de ligações de hidrogênio entre resíduos de aminoácidos do sítio ativo (Pickersgill, et al., 1998).

Ainda que este processo não requeira intervenção do usuário, não há garantias de que as regras obtidas servirão para descrever e selecionar resíduos semelhantes em outras proteínas, uma vez que somente a proteína carregada no ^JPD é utilizada para a criação das regras. Porém, a ferramenta auxilia no estudo e entendimento das propriedades da molécula utilizada e as propriedades de seus resíduos de aminoácido, de acordo com o interesse do usuário.

Para facilitar a busca por regras para CSR's, o módulo foi integrado com a base de dados obtida do *Catalytic Site Atlas* (versão 2.0.12), para automaticamente identificar os resíduos de aminoácidos catalíticos de enzimas presentes no CSA, além de fornecer informações sobre as evidências da anotação e mecanismo catalítico.

7.3 Análise de dados de resíduos de aminoácidos catalíticos

Para facilitar a análise, utilizaram-se gráficos das curvas da Função de Distribuição empírica Acumulada (FDA empírica) de cada descritor. Comparando-se duas FDA's, uma para cada classe, é possível analisar se existem diferenças consideráveis entre as duas distribuições. O teste estatístico de Kolmogorov-Smirnov (Chakravarti, et al., 1967) baseia-se nas FDA's empíricas e, através da máxima distância entre as curvas de dois conjuntos de amostras, o método avalia se estas provêm de diferentes populações. Uma função distribuição acumulada (FDA) descreve a distribuição da probabilidade de uma variável aleatória X , ou seja, para cada possível valor x que a variável aleatória X pode assumir sua FDA é dada por:

$$F(x) = P(X \leq x) \quad \text{Equação 7-2}$$

onde F é a probabilidade de a variável aleatória X assumir um valor inferior ou igual a x . No entanto, a

função de distribuição acumulada geralmente não é conhecida a priori, sendo então estimada através da função de distribuição acumulada empírica pelo uso da equação:

$$\hat{F}(x) = \frac{n^{\circ} \text{ de amostras } \leq x}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \leq x\} \quad \text{Equação 7-3}$$

onde x_i ($i = 1 \dots n$) é a variável aleatória da i -ésima amostra no conjunto com n amostras, e o operador $\mathbf{1}\{E\}$ é um indicador do evento E , ou seja, $x_i \leq x$.

Assim, através da FDA empírica de um descritor para um conjunto de amostras é possível obter a probabilidade das amostras possuírem valores inferiores ou superiores a certo valor de referência. Ainda, através da comparação de duas FDA's empíricas de dois conjuntos de amostras é possível obter o quão diferente são esses conjuntos e, então avaliar se um dado descritor é um bom candidato a oferecer uma separação entre as duas classes de resíduos de aminoácidos (CSR e não-CSR).

Em acordo com outros trabalhos (Barlett, et al., 2002), verificou-se que CSR's geralmente estão localizados dentro de *pockets* na superfícies das moléculas (76% dos CSR's apresentam valores maiores que zero), sendo que os volumes desses *pockets* variam de 96 \AA^3 (mínimo) até 3140 \AA^3 (máximo), enquanto que pouco menos de 34% dos demais resíduos (não catalíticos) encontram-se nessas regiões, como pode ser visto pela gráfico da FDA empírica apresentada na Figura 7-5. Da mesma forma, mais de 93% dos CSR's apresentam área acessível ao solvente maior que zero, indicando presença na superfície das moléculas, uma vez que tais resíduos necessitam estar expostos para a interação com o substrato. Apesar de grande parte estar exposta ao solvente, o mesmo ocorre para os demais resíduos, onde 85% destes também encontram-se na superfície das moléculas. Assim, há fortes indícios de que este descritor não se configura como um bom candidato a promover uma separação entre as classes de resíduos com suficiente acurácia.

Resíduos de aminoácido catalíticos possuem ainda valores para centralidade de proximidade (*closeness*) e distância ao centro de massa da molécula bastante característicos, o que está de acordo com resultados apontados por FAJARDO & FISER (2013). No geral, descritores obtidos a partir da representação das cadeias proteicas em forma de grafos não direcionados resultaram em uma boa separação entre as classes (Figura 7-6) indicando que as propriedades locais, espaciais e os contatos estabelecidos são importantes para descrição e caracterização dos resíduos funcionais incluindo CSR's.

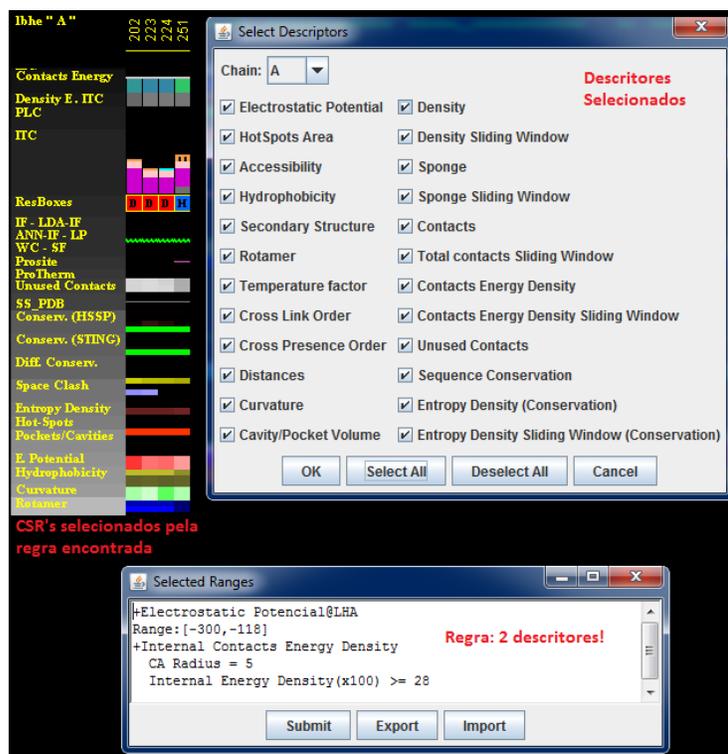


Figura 7-4: Extensão do *JPD* para encontrar regras para resíduos de aminoácido selecionados pelo usuário. Na figura, são selecionados os resíduos de aminoácidos catalíticos para a enzima Poligalacturonase de *Erwinia carotovora subsp. Carotovora* (PDB: 1bhe:A) utilizando apenas dois descritores estruturais de proteínas: potencial eletrostático no LHA entre -300 e -118, e densidade de energia de contatos internos no CA (raio=5Å) ≥ 28 .

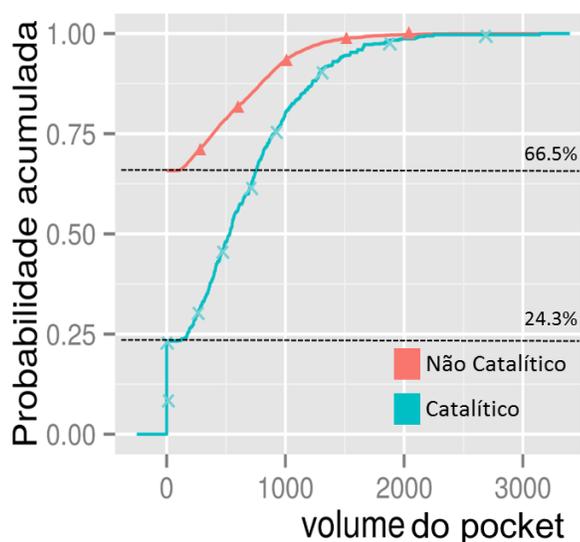


Figura 7-5: FDA's empíricas para o descritor de volume de pocket das classes de resíduos catalíticos e não catalíticos. Volume = 0 indica que o resíduo não se encontra em nenhum pocket. Pelas funções, percebe-se que menos de 25% dos CSR's possuem volume de pocket = 0, enquanto que pouco mais de 66% dos não catalíticos possuem volume de pocket = 0, indicando a preferência de localização dos CSR's nessas regiões da superfície das moléculas.

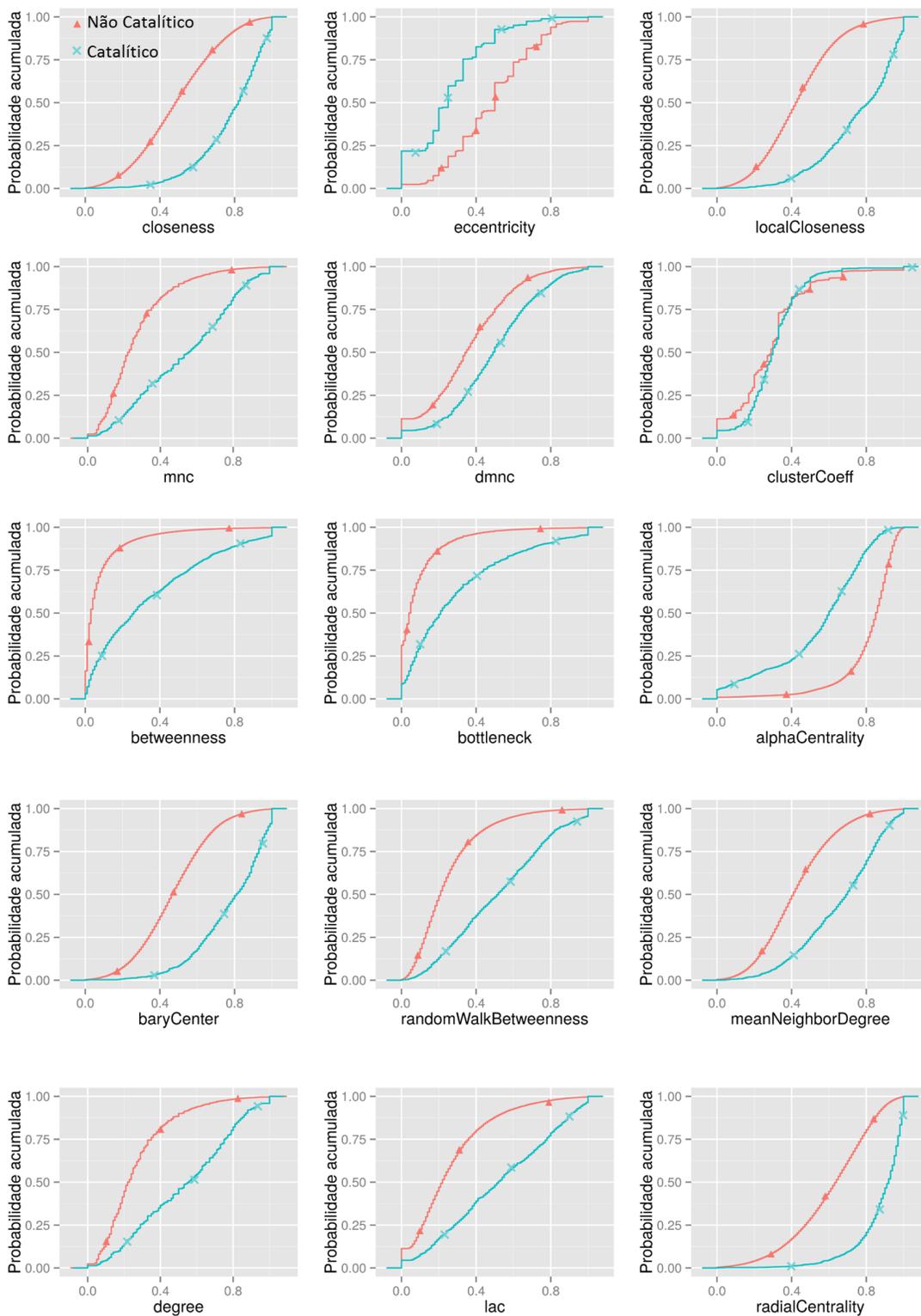
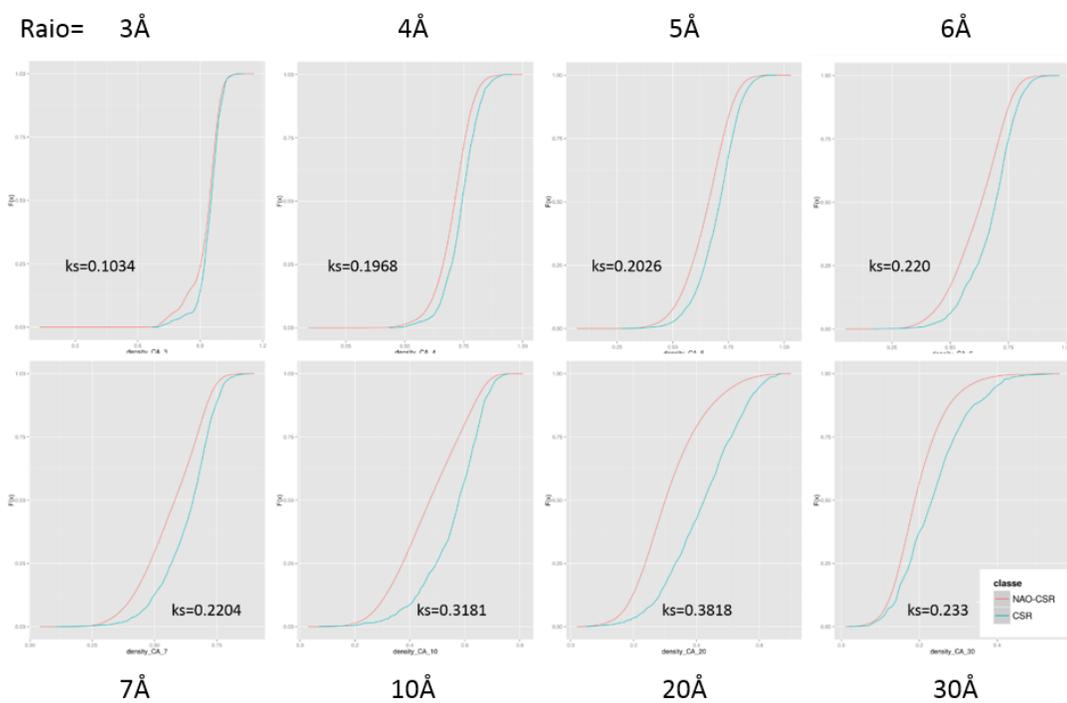


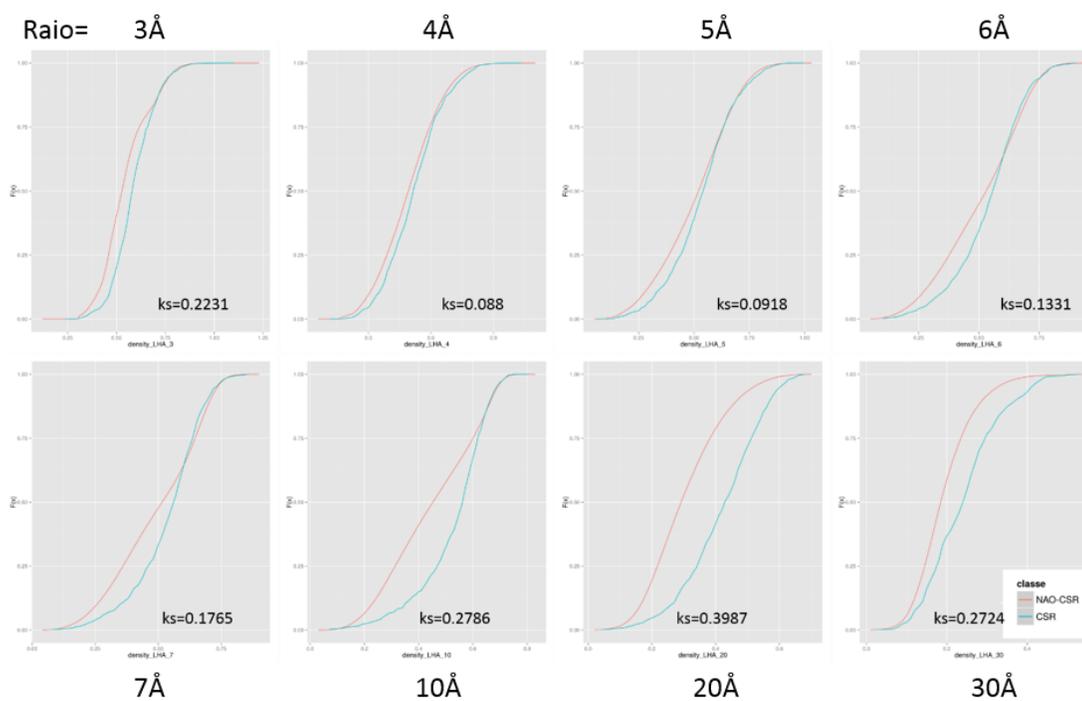
Figura 7-6: FDA's dos descritores obtidos a partir de grafos apresentam significantes diferenças entre as distribuições para os resíduos catalíticos e demais resíduos. Linhas vermelhas: distribuições dos resíduos catalíticos; linhas azuis: distribuições dos resíduos não catalíticos. De cima para baixo, da esquerda para direita encontram-se as FDA's dos descritores: centralidade de proximidade (closeness), excentricidade, centralidade de proximidade local (local closeness), MNC, DMNC, coeficiente de agrupamento, betweenness, gargalo (bottleneck), centralidade alfa, baricentro, betweenness de caminhos aleatórios, número médio de grau dos vizinhos, grau ou cardinalidade, LAC, centralidade radial.

Descritores de densidade e esponjicidade apresentaram, segundo análises gráficas e teste de Kolmogorov-Smirnov, poucas diferenças entre as duas distribuições. Nota-se, no entanto, que à medida que o raio da esfera sonda é aumentado a distância entre as FDA's para os CSR's e os demais resíduos também aumenta, ou seja, à medida que se considera no cálculo da densidade e esponjicidade vizinhanças maiores em torno dos resíduos, maiores são as diferenças entre as duas classes. Isto fica mais evidente quando se analisam os descritores de vizinhança calculados a partir da densidade e esponjicidade, onde as diferenças são ainda maiores, como no caso dos descritores WNA, onde se utiliza uma esfera sonda para a definição da vizinhança de raio igual a 15Å. Para analisar melhor este comportamento, realizou-se o recálculo destes descritores usando esferas sondas de raios 10, 20 e 30Å. Na Figura 7-8 e 7-8, são apresentadas as FDA's e distância máxima entre as duas distribuições (CSR e NÃO-CSR) para os descritores de densidade e esponjicidade. Percebe-se que, para os valores do raio das esferas sondas consideradas (3, 4, 5, 6, 7, 10, 20, e 30Å), a maior distância entre as curvas (estatística de Kolmogorov-Smirnov) é obtida quando o raio da esfera sonda é igual a 20Å, sendo que esta diferença é gradualmente aumentada a partir de $r=3\text{Å}$ até 20Å e, então, sofre uma queda para $r=30\text{Å}$. Para investigar este aumento gradual, seguido de uma queda para $r=30\text{Å}$, calculou-se os volumes das cadeias proteicas presentes na base de dados, através da soma dos volumes dos seus átomos (esferas de van der Waals), sendo assim, obtido um volume médio de aproximadamente de $48,083.22\text{Å}^3 (\pm 23,600.66 \text{Å}^3)$. As esferas sondas de raios igual a 20Å e 30Å possuem, respectivamente, volumes de $33,510.32\text{Å}^3$ e $113,097.34\text{Å}^3$. Dessa forma, a esfera de raio igual a 30Å possui volume suficiente para englobar todos os átomos de grande parte das cadeias na base de dados (cerca de 96% do total de cadeias nas bases de dados), criando um “nanoambiente” que é igual ou próximo de toda a cadeia proteica, como pode ser visto no gráfico da Figura 7-9.

Considerando ainda, as cadeias proteicas como esferas pode-se calcular o raio dessas esferas a partir dos volumes proteicos obtidos anteriormente. Assim, na base de dados encontra-se um valor de raio médio de $22\text{Å} (\pm 3.5\text{Å})$, mostrando que para esferas sonda de raio igual a 30Å a vizinhança de um resíduo de aminoácido e composta por todos os outros resíduos de aminoácidos da cadeia proteica, criando assim, “nanoambientes” indistinguíveis e similares para todos os resíduos de aminoácidos, uma vez que todos terão a mesma vizinhança. Por este motivo, observou-se tal queda na diferença entre as FDA's empíricas entre as classes de resíduos de aminoácido catalítico e não catalíticos.

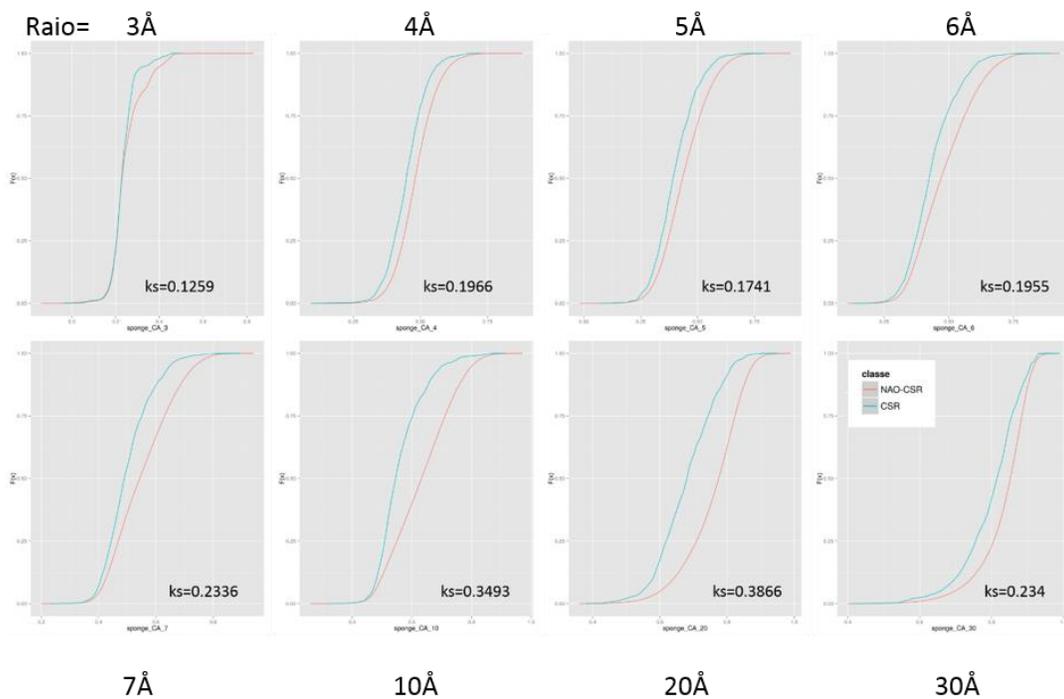


(a)

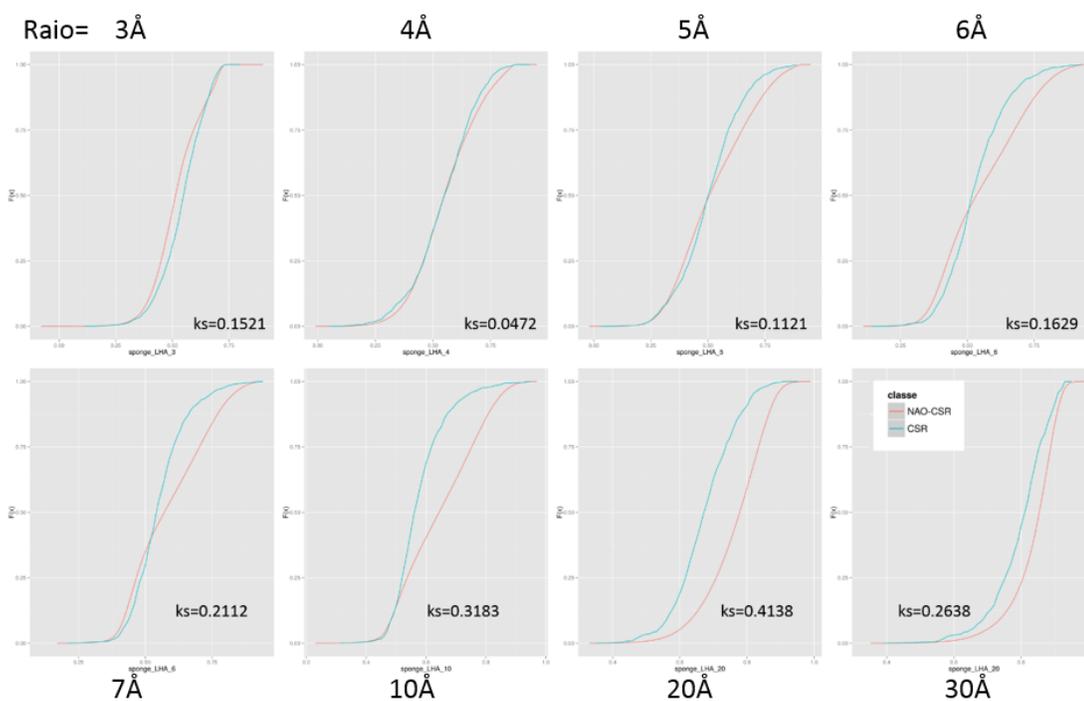


(b)

Figura 7-7: FDA's para descritores densidade, considerando esferas sondas de raios: 3, 4, 5, 6, 7, 10, 20 e 30Å, centradas nos átomos Ca (a) e LHA (b), com as respectivas distâncias máximas entre as duas curvas (estatística de Kolmogorov-Smirnov). Nota-se uma maior diferença entre as distribuições para raio da esfera sonda igual a 20Å (densidade: $ks=0.3818$ em (a) e $ks=0.3987$ em (b)).



(a)



(b)

Figura 7-8: FDA's para descritores esponjicidade, considerando esferas sondas de raios: 3, 4, 5, 6, 7, 10, 20 e 30Å, centradas nos átomos Ca (a) e LHA (b), com as respectivas distâncias máximas entre as duas curvas (estatística de Kolmogorov-Smirnov). Nota-se uma maior diferença entre as distribuições para raio da esfera sonda igual a 20Å: $ks=0.3866$ em (a) e $ks=0.4138$, em (b).

No caso de descritores de potencial eletrostático, as curvas FDA's são muito similares para as duas classes de resíduos consideradas, ou seja, não existem grandes diferenças entre o potencial eletrostático de resíduos catalíticos e os demais resíduos. (Figura 7-10). O potencial eletrostático no LHA, dentre os descritores de potencial eletrostático, é o único em que se pode notar uma ligeira diferença entre os potenciais dos resíduos de aminoácidos catalíticos e dos não catalíticos. Devido à forma que o potencial eletrostático é calculado para cada átomo da proteína, as cargas de todos os átomos da molécula são consideradas como um único sistema e, assim, ocorre uma interferência similar das cargas de todos os átomos a cada átomo da proteína, reduzindo as diferenças entre os potenciais dos resíduos catalíticos e não catalíticos, uma vez que, não se considera um nanoambiente no cálculo do potencial eletrostático. Considerando-se os potenciais dos nanoambientes de cada resíduo de aminoácido através de descritores de vizinhança obtidos através de grafos, nota-se que esta interferência é reduzida no caso dos descritores de potencial eletrostático no LHA, pois, neste caso considera-se o potencial de diferentes camadas (vizinhanças) ao redor de cada átomo LHA dos resíduos de aminoácido e, não o sistema como um todo. O gráfico da Figura 7-11 ilustra este comportamento para o descritor de vizinhança (GN) do potencial eletrostático no LHA, onde percebe-se uma maior diferenciação entre as FDA's empíricas dos CSR's e não-CSR's. Pelo gráfico nota-se que metade dos CSR's (50%) possuem potencial eletrostático na sua vizinhança menor que $-2,150 \text{ kT/J/mol}$, enquanto menos de 13% dos não-CSR possuem valores inferiores a este em suas vizinhanças.

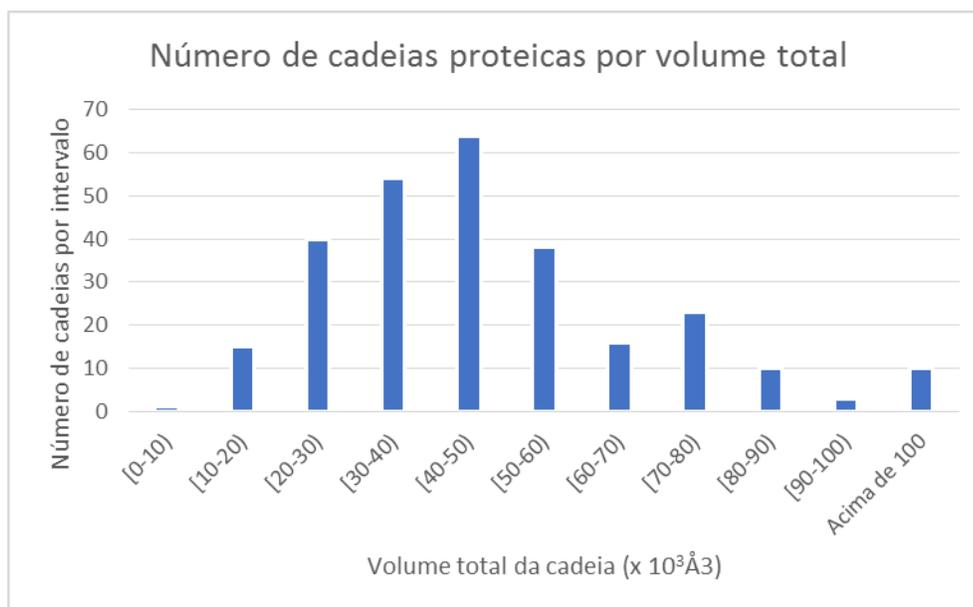


Figura 7-9: Número de cadeias por faixa de volume. Cadeias apresentam maior concentração para volume de $20,000 \text{ \AA}^3$ até $60,000 \text{ \AA}^3$, sendo a média dos volumes das cadeias igual a $48,000 \text{ \AA}^3$.

Considerando descritores de contatos estabelecidos, resíduos catalíticos mostraram possuir energias

totais para contatos internos superiores à maioria dos demais resíduos de aminoácido, onde 65% dos CSR's apresentam energias totais superiores a 35 kcal/mol, enquanto apenas 24% dos não catalíticos apresentam energias totais superiores a este valor. Analisando as distribuições para os 14 tipos diferentes de contatos, percebe-se que esta diferença nas energias totais entre resíduos de aminoácidos catalíticos e não catalíticos deve-se a um maior número de resíduos de aminoácidos catalíticos estabelecendo contatos carregados (atrativos ou repulsivos), ou seja, 64% dos CSR's estabelecem contatos deste tipo contra apenas 23% dos não catalíticos.

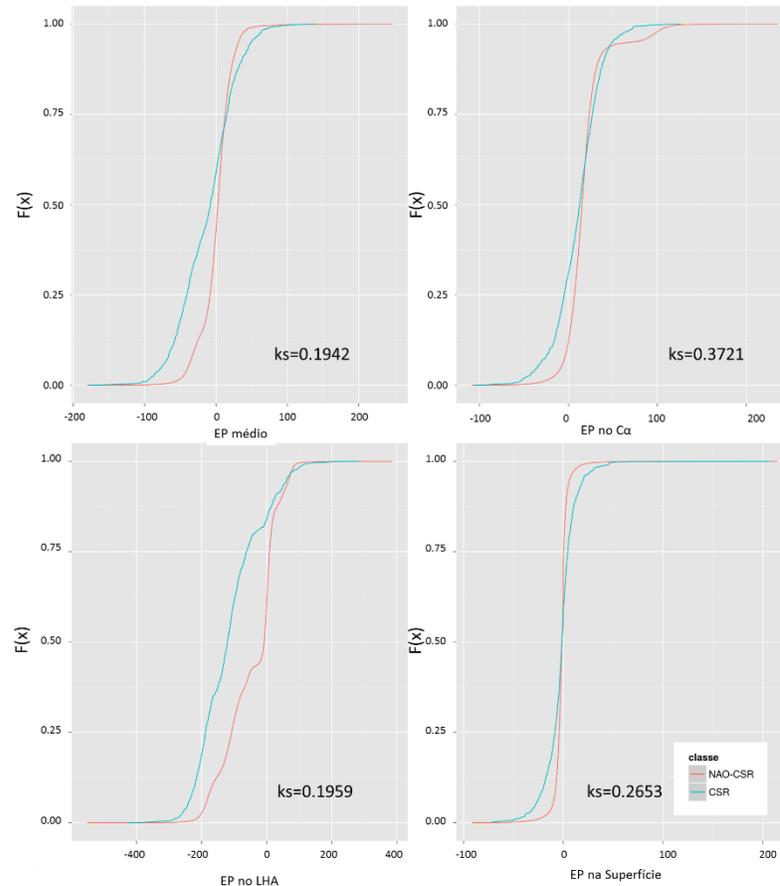


Figura 7-10: FDA's para descritores de potencial eletrostático com os respectivos resultados do teste de Kolmogorov-Smirnov. As curvas e os valores do teste não mostram grandes diferenças entre os dois conjuntos de amostras, sendo o potencial eletrostático no LHA o que possui maior diferença entre as classes de resíduos de aminoácido.

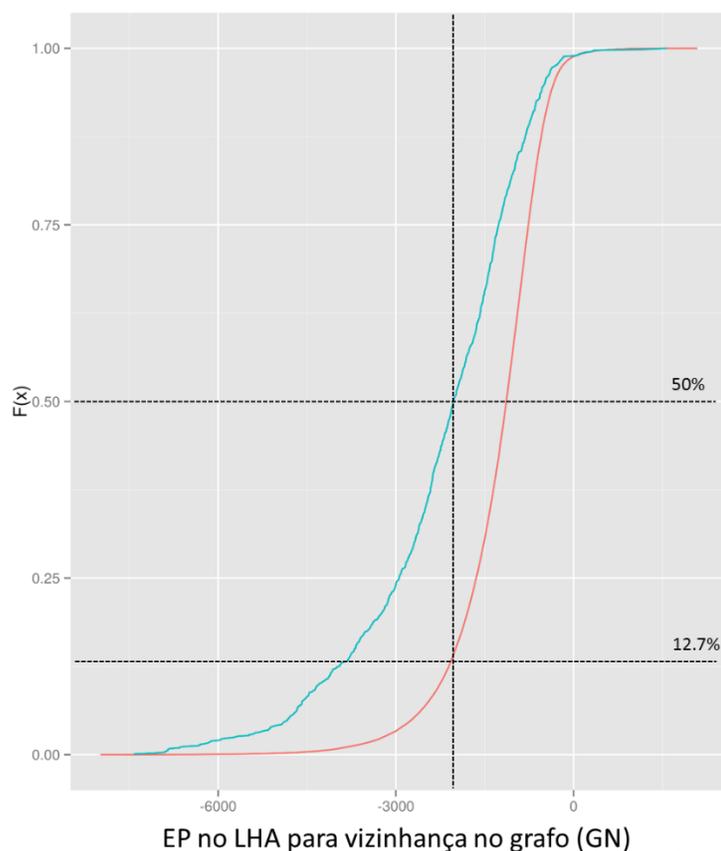


Figura 7-11: Potencial eletrostático no LHA calculado considerando vizinhança definida pela representação da proteína como grafo não direcionado (GN). Metade dos CSR's apresentam valores de EP inferior a $-2,150\text{kT/J/mol}$, enquanto apenas 12.7% dos não-CSR apresentaram tais valores.

Como é de se esperar, o inverso ocorre para o descritor de contatos não usados, onde os resíduos de aminoácido não catalíticos apresentam um maior potencial de contatos em relação aos catalíticos. O fato de um maior número de contatos carregados ser observado para resíduos de ambas as classes, em comparação com os outros 12 tipos de contatos, é devido às configurações das distâncias mínimas e máximas definidas durante o cálculo dos descritores. Para contatos carregados, tanto atrativos quanto repulsivos, a distância máxima é superior ao demais tipos (veja Tabela 4-3). Assim, durante os cálculos os resíduos de aminoácido terão uma tendência maior de possuírem um número de contatos carregados superior aos outros 12 tipos de contatos. O valor limite para a distância de estabelecimento de contatos carregados, estipulado no Blue Star STING, foi retirado do estudo realizado por LEE et al. (2002). Os autores estudaram a dependência entre a distância dos átomos e interações entre cargas em proteínas, e relatam que para distâncias acima de 10\AA são encontradas interações mais fracas entre cargas ($\Delta G \approx 0.1\text{kcal/mol}$), porém alertam que em certos casos o acúmulo destas interações fracas pode resultar em efeitos substanciais. Por este motivo, escolheu-se como distância máxima 12\AA . Este resultado propaga-

se para outros descritores que dependem de contatos em seus cálculos como o caso de ordem de *cross link*, mas não para densidade energética de contatos, uma vez que este último utiliza esferas sondas de raios menores que 12Å (máximo de 9Å), sendo os contatos resultantes de fracas interações excluídos dos cálculos.

Descritores de ordem de *cross link* e ordem de *cross presence* mostram uma tendência de resíduos de aminoácidos catalíticos por ordens mais elevadas do que às dos demais resíduos. No caso dos descritores calculados utilizando-se uma esfera sonda de raio igual 3.5Å, quase a totalidade dos resíduos de ambas as classes possuem ordens iguais a zero. Isto se deve ao fato de que, no pequeno volume considerado (esfera), não é possível encontrar resíduos vizinhos na sequência primária da proteína, considerando segmentos de tamanhos 15, 20 ou 30 aminoácidos. No entanto, para esferas de raios maiores, fica evidente esta tendência (Figura 7-12). Estes descritores estão ligados principalmente com a estabilidade da molécula, ou seja, descritores de ordem de *cross link* e *cross presence* quantificam a importância de resíduo para a manutenção desta estabilidade. O fato de resíduos catalíticos possuírem maior importância indica que os CSR's são, em maior parte, responsáveis por propiciarem uma região adequada para que sejam estabelecidas ligações com o substrato, uma vez que regiões pouco estáveis são menos propícias a fornecerem um ambiente para a ligação. Para entender melhor este comportamento, estudaram-se estes descritores utilizando outros raios e tamanhos de segmentos. Novos descritores foram calculados considerando raios de 10, 20 e 30Å e segmentos de tamanhos 30, 40 e 50 aminoácidos. Pelos histogramas da Figura 7-13, percebe-se que à medida que o raio da esfera sonda aumenta, as duas distribuições (catalíticos e não catalíticos) se aproximam, tornando-se quase indistinguíveis, uma vez que para vizinhanças muito grandes ($r \geq 20\text{Å}$) o número de vizinhos em comum entre resíduos de ambas as classes torna-se maior. No entanto, fixado o raio da esfera, resíduos de aminoácidos catalíticos mostram estar mais próximos espacialmente de resíduos de aminoácidos distantes na sequência, em comparação com resíduos não catalíticos.

Apesar de as distribuições dos diversos descritores apresentarem diferenças entre as duas classes de resíduos, esta diferença não garante uma separação entre elas. Como a classe de resíduos não catalíticos é muito superior em número de amostras, mesmo uma pequena fração destas amostras ainda resulta em um grande número de amostras negativas e bem superior ao total de amostras positivas, ou seja, em muitos casos analisados, tem-se cerca de 10 a 30% de resíduos não catalíticos obedecendo um mesmo patamar que resíduos não catalíticos. Mas 10% do total de resíduos catalíticos ainda é um total expressivo de amostras que podem ou não ser separáveis das amostras positivas. No entanto, o mais importante desta seção é prover um entendimento do comportamento dos resíduos catalíticos para os diversos descritores utilizados neste estudo. Nas seções seguintes, estudou-se o emprego de técnicas de aprendizado de

máquina para a extração de regras visando a caracterização e predição de resíduos catalíticos.

As figuras com as FDA's para todos os descritores encontram-se no Apêndice C.

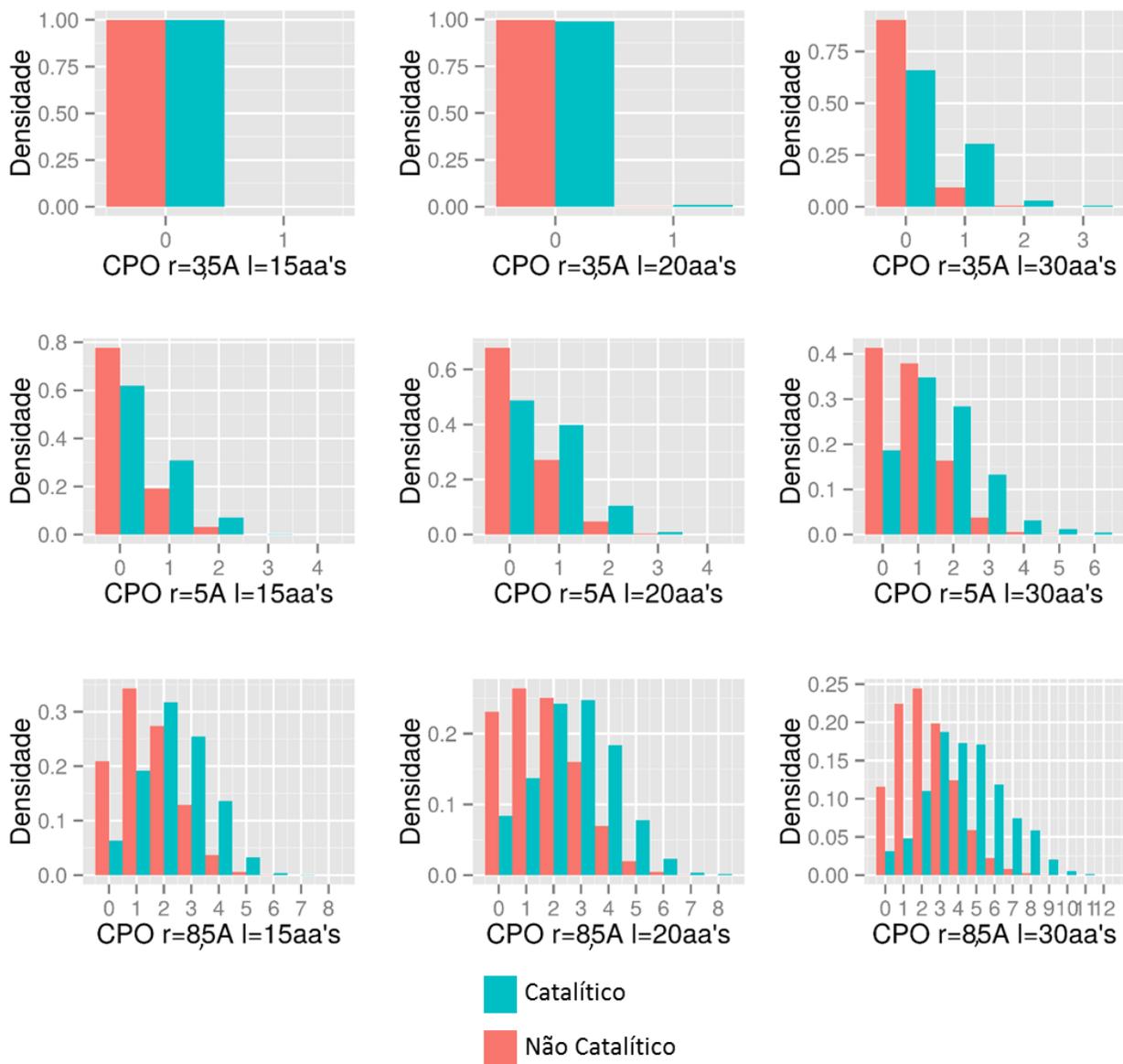


Figura 7-12: Histogramas para descritor de Cross Presence Order, considerando valores de raios e segmentos pré-calculados pelo Blue Star STING; raios (r) = {3.5, 5.0 8.5}Å e segmentos (l) = {15, 20, 30}aa's.

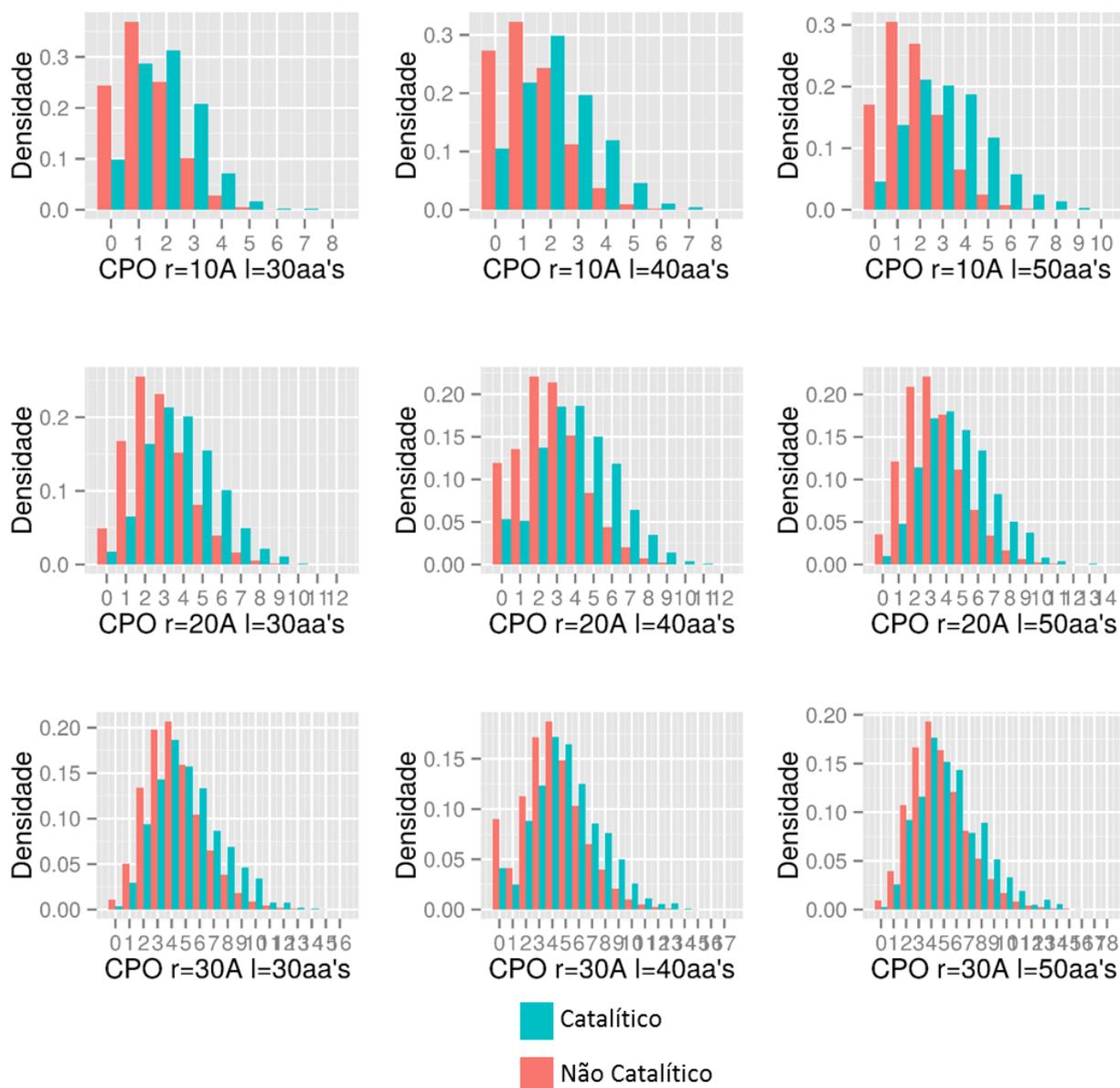


Figura 7-13: Descritores de Cross Presence Order calculados utilizando-se esferas de raios 10, 20 e 30Å e segmentos de tamanhos 30, 40 e 50 aminoácidos. Resíduos de aminoácidos catalíticos apresentam preferências por ordens mais elevadas, mas, a medida que o raio da esfera aumenta as distribuições de ambas as classes se aproximam, tornando-se quase indistinguível.

7.4 Extração de regras para a caracterização de resíduos catalíticos

Motivados pelos resultados obtidos a partir das regras geradas pelo ^JPD através dos métodos manual e automático discutidos nas Seções 7.1 e 7.2, procurou-se avaliar a possibilidade de obtenção de regras que pudessem caracterizar e selecionar resíduos catalíticos de um conjunto de enzimas compartilhando os mesmos três primeiros dígitos EC (sub-subclasse EC). Através do uso do algoritmo de indução de regras RIPPER (sem fase de poda), procurou-se obter regras específicas que caracterizam os resíduos de

aminoácidos catalíticos para cada sub-subclasse de enzimas. O algoritmo RIPPER permite configurar o número mínimo de amostras que uma regra deve cobrir (CF), de forma a ser incluída na solução final (conjunto de regras). Utilizou-se dois valores para o parâmetro CF do algoritmo: CF=2 (valor padrão do RIPPER) e CF=6 (número médio de CSR's).

As regras encontradas pelo RIPPER são capazes de selecionar com alta precisão os resíduos de aminoácidos catalíticos para as diversas sub-subclasses EC, fornecendo uma caracterização do nanoambiente desses resíduos de aminoácidos. Sem a fase de poda, as regras quando CF=2 e CF=6 tornam-se altamente específicas e são capazes de selecionar os resíduos de aminoácidos catalítico com alta precisão, ou seja, nenhum ou poucos falsos positivos são selecionados pelas regras. As sensibilidades dos conjuntos de regras, no entanto, mostram diferenças quando CF=2 e CF=6. No caso de CF=2 o algoritmo produz uma grande quantidade de regras capazes de selecionar todos os CSR's sem introduzir falsos positivos, o que pode ser um indicativo de sobreajuste das regras. Comparando-se a sensibilidade e o número de regras entre os casos onde CF=2 e CF=6, pode-se concluir que as regras geradas quando CF=2 são altamente específicas selecionando um número muito pequeno de CSR's (menor ou igual a 2), pois estas mesmas regras não aparecem no caso de CF=6, caso contrário, as sensibilidades seriam semelhantes e assim como o número de regras. Por exemplo, no caso das Pentosiltransferases (EC.2.4.2) para elevar a sensibilidade de 0.68 quando CF=6 para 1.00 quando CF=2, são adicionalmente criadas 7 novas regras. Uma vez que essa sub-subclasse possui 56 CSR's, tem-se que para cobrir 18 CSR's adicionais (não cobertos onde CF=6), o algoritmo está gerando regras com cobertura média de 2.57 CSR/regra. O mesmo é observado para outras sub-subclasses, indicando assim, a indução de regras altamente específicas que dificilmente irão possuir capacidades de generalização, ao ponto de, realizar a predição de CSR's em novas enzimas. O número de regras encontradas para cada sub-subclasse EC no onde CF=6, ainda que não seja elevado (média de 4 regras), demonstra que mesmo dentro de cada sub-subclasse EC observa-se uma heterogeneidade entre os resíduos catalíticos.

Os resultados contendo sensibilidade e precisão dos conjuntos de regras encontrados para as 12 sub-subclasses, utilizando-se o RIPPER sem fase de poda com CF=2 e CF=6, encontram-se na Tabela 7-5.

Uma vez que não utilizou-se a fase de poda, evitando a fragmentação do conjunto de treinamento em dois subconjuntos: conjunto para crescimento das regras (*grow set*) e conjunto de poda (*prune set*), o algoritmo buscou regras de maior cobertura e precisão em todo o conjunto de amostras. A medida que as regras são podadas busca-se elevar a cobertura, reduzindo-se assim a precisão devido a eventual introdução de falsos positivos devido à baixa representatividade das amostras positivas. Valores altos de precisão, como os apresentados, obtidos pelo algoritmo sem fase de poda podem caracterizar um sobreajuste das regras aos dados fornecidos, limitando suas aplicações em novas amostras.

Ainda que essas regras sejam limitadas no sentido de realizarem a predição de resíduos de aminoácidos catalíticos em novas enzimas, estipulam um limite para as predições e fornecem uma caracterização dos resíduos de aminoácidos catalíticos de enzimas já anotadas.

Tabela 7-5: Sensibilidade (Sens.), precisão (Prec.) e número de regras para todas as sub-subclasses EC estudadas encontradas pelo RIPPER-UNPRUNED (sem fase de poda) com CF=2 e CF=6.

DATASET	RIPPER-UNPRUNED (CF=2)			RIPPER-UNPRUNED (CF=6)		
	Sens.	Prec.	#Regras	Sens.	Prec.	#Regras
EC.1.1.1	1.00	0.99	10	0.67	1.00	5
EC.2.1.1	1.00	1.00	8	0.60	1.00	3
EC.2.3.1	1.00	1.00	12	0.49	0.87	4
EC.2.4.2	1.00	1.00	12	0.68	1.00	5
EC.2.5.1	1.00	0.99	10	0.60	1.00	5
EC.2.7.1	1.00	1.00	9	0.67	1.00	4
EC.3.1.1	1.00	1.00	14	0.47	0.87	3
EC.3.1.3	1.00	1.00	10	0.81	1.00	6
EC.3.2.1	1.00	1.00	15	0.59	1.00	5
EC.3.4.21	1.00	0.99	8	0.62	1.00	3
EC.4.1.1	1.00	0.99	8	0.54	1.00	4
EC.4.2.1	1.00	1.00	11	0.56	1.00	4

A Tabela 7-6 traz regras encontradas para duas sub-subclasses EC, sendo a sub-subclasse das Glicosidades (EC. 3.2.1) a de maior número de CSR's (148), e a sub-subclasse das Metiltransferases (EC. 2.1.1) a com menor número de CSR's (42) para o caso onde CF=6. As regras para as demais sub-subclasses incluídas neste trabalho encontram-se no Apêndice E.

Para avaliar a variabilidade das soluções (conjuntos de regras) e seus respectivos desempenhos fornecidas pelo RIPPER, o algoritmo multiobjetivo MOEA-RIPPER foi empregado. Os gráficos das Figuras 7-13 e 7-14 ilustram essa variabilidade das soluções não dominadas encontradas para cada sub-subclasse EC, onde é possível notar que o aumento do número de regras e antecedentes leva a um aumento do F-measure. No entanto, pode-se perceber casos em que com um menor número de regras e antecedentes foi possível obter maior F-measure. Devido ao caráter local do RIPPER, a própria busca pode estar causando uma separação do espaço das amostras, pois trata-se de uma abordagem *separate-and-conquer*. No caso do MOEA-RIPPER, foi ainda possível encontrar soluções com 100% de precisão e sensibilidade com menor número de regras do que as soluções encontradas pelo RIPPER-*Unpruned* quando CF=2, outra indicativa de que o algoritmo mostra dificuldades em selecionar os antecedentes para formar as regras, uma vez que o RIPPER-*Unpruned* (CF=2) não foi capaz de encontrar a regra

encontrada pelo MOEA-RIPPER, com menor número de antecedentes e mesmo desempenho.

Tabela 7-6: Regras encontradas pelo RIPPER sem fase de poda, utilizando todo o conjunto de amostras de duas sub-classes: EC's 2.1.1 e 3.2.1. As regras são altamente específicas e com baixo suporte, o que pode levar a um sobreajuste. Mais detalhes sobre os descritores presentes nas regras (antecedentes) estão apresentados nas seções 0 e 4.3.

Regras EC 2.1.1	Pos./Neg.
$ASA_{apolar} \geq 177.52 \wedge cpo_{r=8.5;l=30}^{VD} \geq 5.36 \wedge uc_{hbmwm}^{WNADist} \leq 47.03 \wedge clo_{r=3.5;l=15;LHA}^{WNASurf} \geq 0.3 \wedge density_{r=7;CA} \geq 0.55$	12/0
$ASA_{apolar} \geq 177.52 \wedge cpo_{r=5;l=20;LHA}^{GN} \geq 8.41 \wedge clo_{r=3.5;l=30;CA}^{GN} \leq 0.32 \wedge density_{r=3;LHA}^{WNADist} \geq 3.26 \wedge clo_{r=8.5;l=30;LHA}^{VD} \geq 1.86 \wedge charge - repu_{energy} \leq 140$	8/0
$alphaCentrality \leq 0.68 \wedge meanNeighborDegree \geq 0.63 \wedge uc_{hbmws}^{WNADist} \leq 60.59 \wedge uc_{charge-repu}^{VD} \geq 318.44 \wedge density_{r=5;CA}^{WNADist} \geq 3.87$	5/0
Regras EC 3.2.1	Pos./Neg.
$localCloseness \geq 0.84 \wedge ASA_{total} \geq 237.85 \wedge uc_{hbmwws}^{VD} \leq 22.87 \wedge distance_{CG} \leq 10.61 \wedge uc_{charge-repu}^{GN} \geq 4650.87 \wedge cpo_{r=5;l=20;LHA}^{WNADist} \leq 5.78$	29/0
$localCloseness \geq 0.91 \wedge ASA_{polar} \geq 145.09 \wedge uc_{aromatic} \leq 1.5 \wedge uv_{hbmwm}^{VD} \geq 17.36 \wedge density_{r=7;CA}^{WNADist} \geq 3.73$	19/0
$(localCloseness \geq 0.79) \wedge (uc_{aromatic} \leq 0) \wedge (polarASA \geq 94.82) \wedge (sponge_{r=6;CA}^{WNADist} \leq 2.13) \wedge (hb_{mwm}^{GN} \leq 23.97) \wedge (density_{r=3;LHA}^{VD} \geq 1.11)$	17/0
$localCloseness \geq 0.91 \wedge hydroKDI \leq -0.11 \wedge ced_{r=5;CA}^{WNADist} \geq 0.83 \wedge cpo_{r=8.5;l=15;LHA}^{GN} \leq 98.59 \wedge hb_{sws}^{WNASurf} \leq 0.43$	11/0
$localCloseness \geq 0.65 \wedge cpo_{r=8.5;l=30;CA}^{VD} \geq 10.93 \wedge hydroKDI \leq -0.06 \wedge volume\ do\ pocket \geq 865.04 \wedge clo_{r=3.5;l=20;CA}^{GN} \geq 4.22 \wedge clo_{r=5;l=20;CB}^{WNADist} \leq 6.76 \wedge hbmm_{energy} \leq 10.4$	11/0

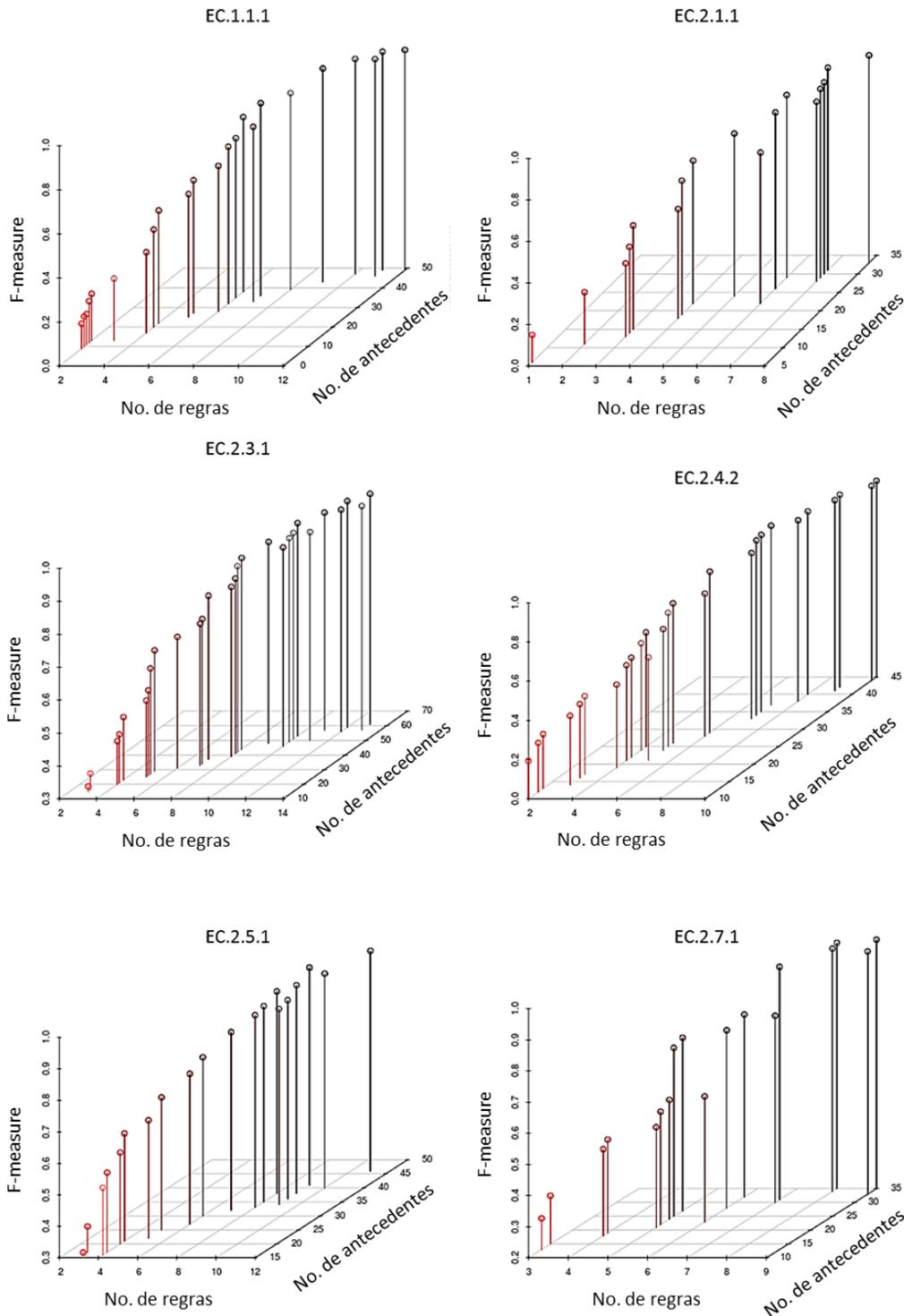


Figura 7-14: Soluções não-dominadas encontradas pelo MOEA-RIPPER. Gráficos com as soluções encontradas pelo emprego do algoritmo evolutivo multiobjetivo para a seleção de atributos considerando os três objetivos conflitantes: número de regras, número total de antecedentes nas regras e F-measure. A variabilidade das soluções ilustra os diferentes compromissos entre os objetivos e possibilita a percepção do impacto do número e regras e seus respectivos tamanhos no desempenho dos classificadores.

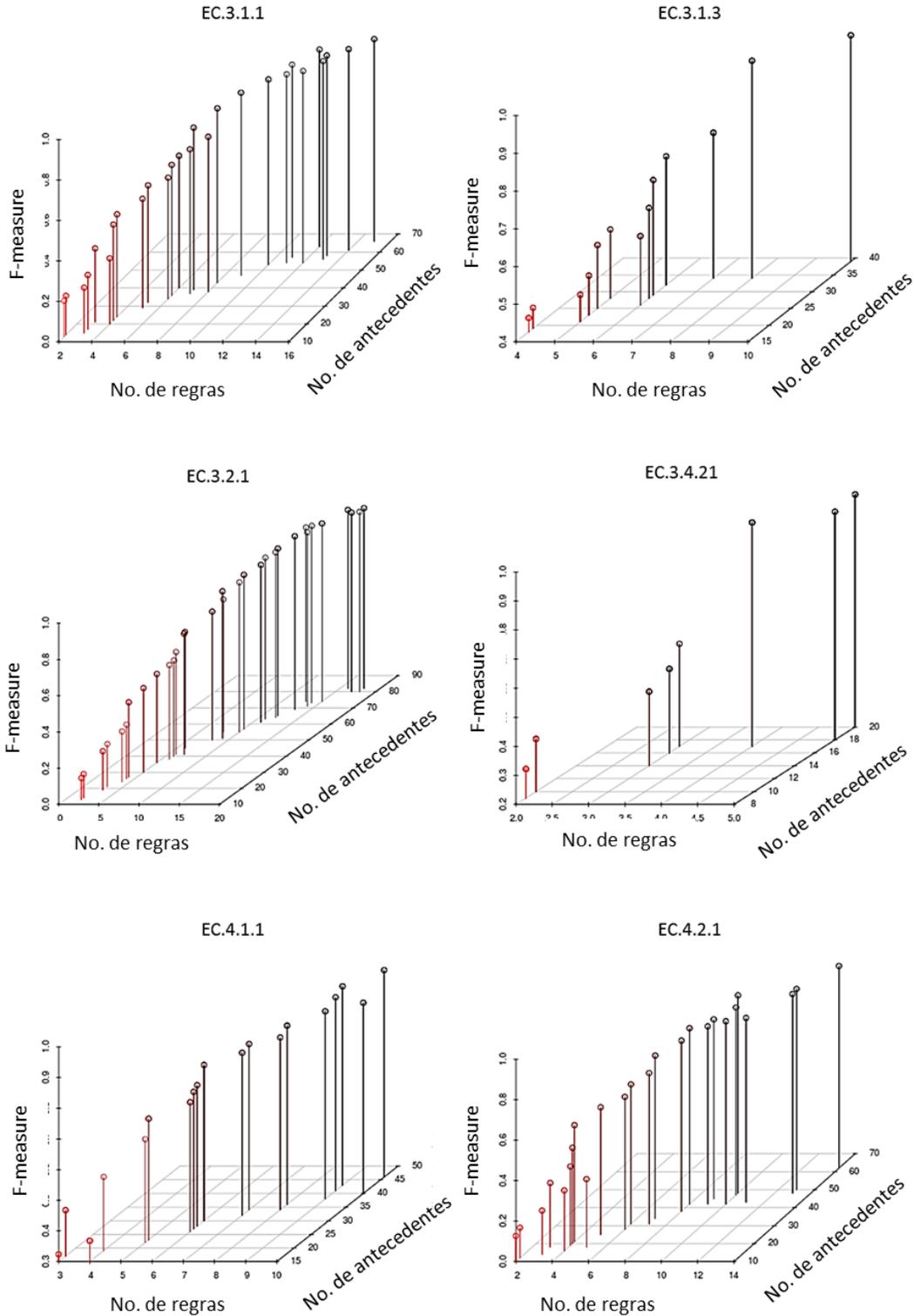


Figura 7-15: Soluções não-dominadas encontradas pelo MOEA-RIPPER. Gráficos com as soluções encontradas pelo emprego do algoritmo evolutivo multiobjetivo para a seleção de atributos considerando os três objetivos conflitantes: número de regras, número total de antecedentes nas regras e F-measure. A variabilidade das soluções ilustra os diferentes compromissos entre os objetivos e possibilita a percepção do impacto do número e regras e seus respectivos tamanhos no desempenho dos classificadores.

Ainda que não seja possível estimar como as regras encontradas irão comportar-se em novos conjuntos de dados, estes resultados fornecem indícios sobre a possibilidade de caracterização dos resíduos de aminoácidos catalíticos e algumas propriedades distintas entre estes e os demais resíduos de aminoácidos das enzimas. Avaliou-se, no entanto, o emprego das regras encontradas para uma sub-subclasse em outras sub-subclasses, de forma a averiguar se as regras são específicas para resíduos de aminoácido catalíticos pertencentes às enzimas de uma mesma sub-subclasse. Na Tabela 7-7, encontram-se a proporção dos resíduos de aminoácidos catalíticos cobertos pelos conjuntos de regras das sub-subclasses dispostas nas linhas quando aplicados em outras sub-subclasses (colunas). Ainda que as regras mostrem pouca especificidade em alguns casos, selecionando um alto número de resíduos de aminoácidos catalíticos de outras sub-subclasses, estes são sempre inferiores às suas sensibilidades. Dessa forma, as regras encontradas pelo RIPPER para as diferentes sub-subclasses misturam antecedentes que exploram características gerais (inter sub-subclasse) e locais (intra sub-subclasse) dos resíduos de aminoácidos catalíticos.

Tabela 7-7: Proporção dos resíduos de aminoácidos catalíticos cobertos pelas regras encontradas para as sub-subclasses (linhas) quando aplicadas a outras sub-subclasses (colunas). Valores na diagonal (em destaque preto) corresponde a sensibilidade das regras. Em destaque (vermelho) estão as máximas porcentagens obtidas ao se aplicarem as regras à sub-subclasses diferentes daquela para qual foram induzidas. As cores nos códigos EC corresponde às classes EC: EC.1 Oxirredutases (laranja), EC.2 Transferases (amarelo), EC.3 Hidrolases (verde) e EC.4 Liases (vermelho).

EC.	1.1.1	2.1.1	2.3.1	2.4.2	2.5.1	2.7.1	3.1.1	3.1.3	3.2.1	3.4.21	4.1.1	4.2.1
1.1.1	66.7%	0.0%	10.0%	10.7%	3.6%	15.4%	6.8%	10.3%	22.3%	8.9%	7.7%	11.8%
2.1.1	8.7%	59.5%	5.0%	3.6%	1.8%	1.9%	7.8%	2.9%	8.8%	0.0%	4.6%	1.5%
2.3.1	23.2%	9.5%	48.8%	12.5%	18.2%	19.2%	12.6%	11.8%	28.4%	7.1%	6.2%	5.9%
2.4.2	7.2%	4.8%	2.5%	68.8%	1.8%	9.6%	7.8%	2.9%	6.8%	1.8%	4.6%	5.9%
2.5.1	7.2%	16.7%	10.0%	10.7%	60.0%	5.8%	7.8%	8.8%	11.5%	8.9%	10.8%	1.5%
2.7.1	13.0%	11.9%	11.3%	1.8%	1.8%	67.3%	3.9%	8.8%	9.5%	1.8%	12.3%	13.2%
3.1.1	14.5%	4.8%	12.5%	7.1%	7.3%	3.8%	45.6%	13.2%	23.0%	16.1%	4.6%	5.9%
3.1.3	8.7%	21.4%	5.0%	8.9%	14.5%	7.7%	13.6%	80.9%	28.4%	14.3%	10.8%	10.3%
3.2.1	4.3%	4.8%	0.0%	1.8%	3.6%	13.5%	1.0%	8.8%	43.9%	3.6%	4.6%	2.9%
3.4.21	8.7%	2.4%	7.5%	3.6%	9.1%	5.8%	28.2%	4.4%	9.5%	62.5%	1.5%	4.4%
4.1.1	10.1%	2.4%	3.8%	1.8%	10.9%	11.5%	1.0%	13.2%	17.6%	0.0%	53.8%	10.3%
4.2.1	8.7%	0.0%	6.3%	7.1%	1.8%	9.6%	2.9%	4.4%	7.4%	0.0%	3.1%	56.0%

7.5 Predição de resíduos de aminoácidos catalíticos

Para estimar o desempenho dos classificadores em novos conjuntos de dados, utilizou-se validação cruzada 2x5. Inicialmente, foram considerados dois algoritmos para a obtenção de regras: RIPPER e C4.5.

Como esperado, sem um pré-processamento dos dados, seja via amostragem ou redução de dimensionalidade, o desempenho dos classificadores treinados pode ser considerado baixo para

finalidades práticas. Os classificadores obtidos possuem pouca capacidade de generalização e, assim, o emprego desses em novos conjuntos de enzimas torna-se inviável do ponto de vista prático, como mostra a Tabela 7-8. Um alto desvio padrão também indica grande variabilidade entre as predições realizadas junto aos conjuntos de teste durante a validação cruzada. Essa variabilidade está relacionada com a dependência dos classificadores para com o conjunto de treinamento, de forma que hipóteses (conjuntos de regras) bem diferentes umas das outras são construídas a cada treinamento.

Os resultados da validação cruzada para os dois algoritmos foram equivalentes, não apresentando diferenças significativas entre as predições realizadas para as sub-subclasses (Apêndice F-i). Segundo o teste de Welch (significância de 5%), somente no caso da sub-subclasse Carboxi-Liases (EC.4.1.1) houve diferenças significantes entre RIPPER e C4.5 (valor-p = 0.014). Dessa forma, não há evidências estatísticas de que as predições do RIPPER são superiores às fornecidas pelo C4.5, sendo que os dois algoritmos possuem, na média, desempenhos equiparáveis.

Algoritmos indutores de regras e de construção de árvores de decisão são bastante sensíveis ao tamanho do conjunto de treinamento. Estes algoritmos apresentam dificuldades de aprendizado caso o número de amostras seja relativamente baixo, sendo que geralmente um maior número de amostras leva a um melhor aprendizado. Tendo-se em vista que a classe positiva (minoritária) em todos os conjuntos de dados possui um número muito baixo de amostras, as hipóteses construídas por estes algoritmos serão sobreajustadas e altamente específicas, sendo que dificilmente irão se aplicar a novas amostras. O número de amostras positivas em cada sub-subclasse EC tem influência direta nos desempenhos dos classificadores treinados. Para conjuntos com um número de amostras positivas inferior a 80, nota-se uma variação dos desempenhos dos classificadores, bem como desvios padrões elevados para cada sub-subclasse EC, indicando grandes diferenças nos classificadores treinados durante a validação cruzada e suas predições. Porém, para conjuntos com mais de 80 amostras, nota-se uma tendência no aumento dos desempenhos assim como uma redução dos desvios padrões (Figura 7-16). Ainda que apenas duas sub-subclasses possuam número de amostras positivas superiores a 80, a diferença entre os desempenhos dos classificadores treinados para as Hidrolases de Ester-carboxílico (EC.3.1.1) e Glicosidasas (EC.3.2.1), em relação às outras sub-subclasses, levanta suspeitas sobre a influência do número de resíduos de aminoácidos catalíticos nestes conjuntos. Sendo todos os conjuntos compostos por enzimas não redundantes, segundo um mesmo nível de similaridade sequencial, essas diferenças podem ser um resultado do número de amostras positivas em cada sub-subclasse. Sendo necessário um alto número de amostras para o aprendizado eficiente por RIPPER e C4.5, sub-subclasses com maior número de CSR's tendem a apresentar melhor desempenho.

Neste caso, o desbalanceamento entre classes não é o maior responsável pelo baixo desempenho dos

classificadores, mas sim a sub-representação da classe positiva. Caso um maior número de amostras positivas seja introduzido nesses conjuntos, mesmo que com um aumento do número de negativas, mantendo a mesma proporção entre as classes, observa-se um aumento no desempenho dos classificadores.

Tabela 7-8: Valores médios para sensibilidade (Sens.), precisão (Prec.) e F-measure com respectivos desvios padrões (em parênteses) das validações cruzadas utilizando RIPPER e C4.5 para todas as sub-subclasses EC.

DATASET	RIPPER			C4.5		
	Sens.	Prec.	F-measure	Sens.	Prec.	F-measure
EC.1.1.1	0.17(0.09)	0.20(0.14)	0.18(0.10)	0.17(0.14)	0.20(0.12)	0.16(0.11)
EC.2.1.1	0.06(0.06)	0.11(0.13)	0.07(0.08)	0.03(0.05)	0.08(0.13)	0.04(0.07)
EC.2.3.1	0.15(0.09)	0.24(0.20)	0.17(0.13)	0.11(0.09)	0.19(0.17)	0.13(0.10)
EC.2.4.2	0.19(0.13)	0.19(0.12)	0.18(0.11)	0.14(0.14)	0.22(0.31)	0.16(0.17)
EC.2.5.1	0.08(0.10)	0.09(0.11)	0.09(0.10)	0.11(0.06)	0.22(0.29)	0.13(0.09)
EC.2.7.1	0.26(0.13)	0.27(0.13)	0.23(0.07)	0.15(0.07)	0.28(0.22)	0.18(0.10)
EC.3.1.1	0.27(0.12)	0.36(0.08)	0.29(0.09)	0.26(0.14)	0.31(0.18)	0.26(0.12)
EC3.1.3	0.31(0.22)	0.24(0.17)	0.24(0.13)	0.33(0.15)	0.37(0.17)	0.32(0.11)
EC.3.2.1	0.37(0.07)	0.40(0.08)	0.38(0.05)	0.31(0.09)	0.37(0.09)	0.32(0.07)
EC.3.4.21	0.23(0.14)	0.35(0.25)	0.24(0.10)	0.20(0.18)	0.18(0.13)	0.16(0.11)
EC.4.1.1	0.13(0.08)	0.29(0.18)	0.16(0.08)	0.05(0.06)	0.10(0.15)	0.06(0.09)
EC.4.2.1	0.14(0.09)	0.24(0.18)	0.17(0.10)	0.19(0.14)	0.23(0.16)	0.18(0.11)
MÉDIA	0.19(0.11)	0.25(0.14)	0.20(0.09)	0.17(0.11)	0.23(0.17)	0.17(0.10)

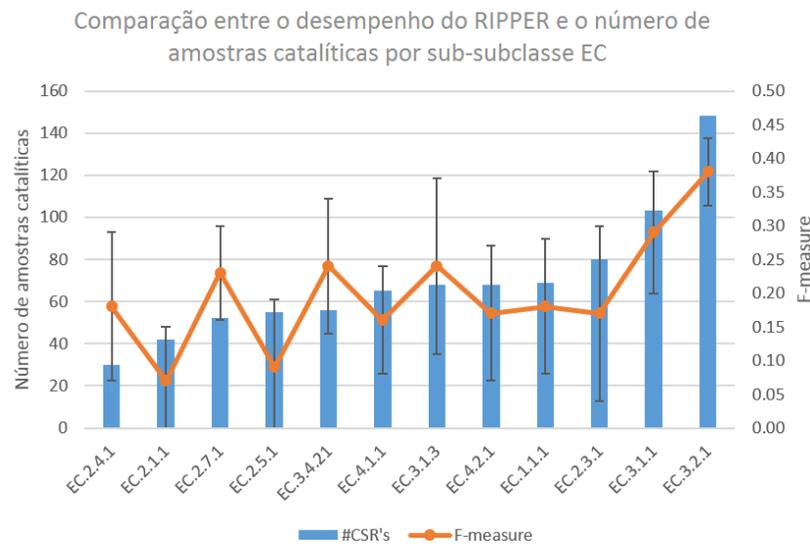


Figura 7-16: Relação entre o número de resíduos de aminoácidos catalíticos e desempenhos dos classificadores treinados para as diferentes sub-subclasses EC. O aumento do número de CSR's leva a um aumento do F-measure dos classificadores (fortalecimento da hipótese gerada pelo classificador).

7.6 Impacto da subamostragem no treinamento dos classificadores

Nesta etapa, procurou-se estudar o impacto da subamostragem aleatória da classe negativa no desempenho dos classificadores C4.5 e RIPPER. Para isto, utilizaram-se diferentes taxas de subamostragem, de forma que a proporção entre o número de amostras positivas e negativas fossem de: 1:1, 1:2, 1:6 (Gutteridge, et al., 2003), 1:10, 1:25 e 1:50.

A subamostragem aleatória pode ocasionar a remoção de um número muito alto de amostras negativas, incluindo aquelas contendo informações relevantes para o aprendizado dos classificadores. Assim, avaliou-se também a subamostragem da classe majoritária utilizando o algoritmo evolutivo para a seleção de conjunto de treinamento, denominado EA-TSS e descrito na Seção 5.2. Foi utilizado um número de 30 estratos para as amostras da classe negativa e o processo evolutivo foi mantido por 1000 gerações.

A subamostragem foi utilizada para reduzir o número de amostras negativas nos conjuntos de treinamento sendo os classificadores avaliados em conjuntos de teste contendo a proporção original entre as classes (sem amostragem).

À medida que os conjuntos tornam-se mais desbalanceados, percebe-se uma redução substancial na sensibilidade dos classificadores, acompanhada de um aumento mais discreto na precisão. A redução no desbalanceamento entre as classes traz ganhos importantes na sensibilidade dos classificadores, com uma queda na precisão quando comparado aos classificadores treinados sem subamostragem da classe majoritária. No entanto, na maioria dos casos, os desvios padrões para sensibilidade, precisão e *F-measure* indicam grande variabilidade dos classificadores treinados segundo diferentes conjuntos de treinamento. Ainda que estas regras obtenham maiores sensibilidades e precisões, elas não são capazes de fornecer predições eficazes. Na Tabela 7-9, é apresentado o resultado dos classificadores treinados utilizando-se o algoritmo RIPPER com base em seus os valores médios e desvio padrões para cada sub-subclasse e diferentes taxas de subamostragem (proporção entre classes). Na Tabela 7-10, é apresentado o mesmo resultado para os classificadores treinados utilizando-se o algoritmo para construção de árvores de decisão C4.5.

Para todas as proporções entre classes consideradas, observou-se que os maiores valores de *F-measure* são obtidos quanto a subamostragem é menor, ou seja, para proporções de 1:25 e 1:50. Isto porque, para outras proporções, há uma grande remoção de amostras negativas e uma queda da precisão dos classificadores, mesmo que seguido de aumento da sensibilidade. Dessa forma, a drástica redução do número de amostras da classe majoritária ocasiona a remoção de amostras importantes para o aprendizado, sendo que reduções em menor escala resultam em melhor desempenho.

Tabela 7-9: Desempenho dos classificadores treinados com RIPPER após subamostragem da classe negativa. Valores de sensibilidade (S), precisão (P) e F-measure (F) médios, com respectivos desvios padrões, durante validação cruzada. Em destaque, proporções que levaram aos maiores F-measures por sub-subclasse.

RIPPER							
#CSR:#ÑCSR		1:1	1:2	1:6	1:10	1:25	1:50
EC.1.1.1	S	0.82(0.17)	0.74(0.17)	0.60(0.15)	0.57(0.31)	0.48(0.20)	0.32(0.16)
	P	0.03(0.02)	0.06(0.02)	0.09(0.05)	0.09(0.05)	0.15(0.06)	0.16(0.10)
	F	0.06(0.03)	0.11(0.04)	0.15(0.07)	0.16(0.08)	0.22(0.08)	0.20(0.08)
EC.2.1.1	S	0.67(0.16)	0.65(0.23)	0.46(0.22)	0.43(0.18)	0.29(0.19)	0.24(0.19)
	P	0.02(0.01)	0.04(0.02)	0.08(0.05)	0.07(0.05)	0.09(0.07)	0.16(0.15)
	F	0.05(0.02)	0.08(0.04)	0.13(0.07)	0.12(0.07)	0.13(0.10)	0.18(0.16)
EC.2.3.1	S	0.83(0.11)	0.77(0.13)	0.58(0.21)	0.45(0.16)	0.26(0.13)	0.18(0.17)
	P	0.03(0.01)	0.04(0.01)	0.07(0.04)	0.08(0.04)	0.12(0.08)	0.11(0.06)
	F	0.07(0.01)	0.08(0.02)	0.12(0.04)	0.13(0.04)	0.16(0.10)	0.13(0.08)
EC.2.4.2	S	0.72(0.19)	0.71(0.20)	0.40(0.20)	0.51(0.27)	0.34(0.15)	0.15(0.11)
	P	0.03(0.01)	0.05(0.02)	0.06(0.03)	0.11(0.07)	0.15(0.10)	0.13(0.12)
	F	0.06(0.03)	0.09(0.03)	0.10(0.05)	0.17(0.10)	0.19(0.11)	0.13(0.10)
EC.2.5.1	S	0.66(0.19)	0.58(0.22)	0.41(0.23)	0.37(0.21)	0.20(0.21)	0.17(0.17)
	P	0.03(0.01)	0.04(0.02)	0.07(0.06)	0.08(0.04)	0.07(0.07)	0.14(0.12)
	F	0.05(0.01)	0.08(0.03)	0.12(0.09)	0.13(0.06)	0.09(0.09)	0.14(0.13)
EC.2.7.1	S	0.70(0.17)	0.84(0.14)	0.62(0.21)	0.47(0.22)	0.35(0.15)	0.37(0.18)
	P	0.04(0.02)	0.06(0.02)	0.07(0.03)	0.09(0.06)	0.14(0.08)	0.19(0.10)
	F	0.07(0.04)	0.11(0.03)	0.12(0.04)	0.15(0.08)	0.19(0.08)	0.24(0.10)
EC.3.1.1	S	0.76(0.11)	0.68(0.11)	0.51(0.13)	0.57(0.11)	0.44(0.16)	0.33(0.12)
	P	0.05(0.02)	0.06(0.02)	0.10(0.03)	0.20(0.10)	0.28(0.11)	0.24(0.08)
	F	0.10(0.04)	0.11(0.04)	0.17(0.05)	0.29(0.11)	0.32(0.09)	0.26(0.09)
EC.3.1.3	S	0.78(0.19)	0.79(0.17)	0.62(0.17)	0.67(0.21)	0.37(0.16)	0.38(0.21)
	P	0.06(0.02)	0.09(0.06)	0.11(0.04)	0.15(0.07)	0.17(0.06)	0.24(0.15)
	F	0.11(0.04)	0.16(0.09)	0.18(0.06)	0.23(0.09)	0.23(0.07)	0.27(0.14)
EC.3.2.1	S	0.90(0.06)	0.83(0.13)	0.70(0.13)	0.73(0.10)	0.61(0.07)	0.56(0.14)
	P	0.05(0.01)	0.07(0.02)	0.10(0.02)	0.16(0.04)	0.22(0.04)	0.28(0.07)
	F	0.10(0.02)	0.14(0.03)	0.18(0.04)	0.27(0.05)	0.32(0.05)	0.37(0.08)
EC.3.4.21	S	0.70(0.22)	0.59(0.22)	0.47(0.25)	0.50(0.23)	0.44(0.22)	0.28(0.14)
	P	0.04(0.02)	0.05(0.02)	0.07(0.04)	0.12(0.07)	0.22(0.20)	0.29(0.16)
	F	0.08(0.03)	0.08(0.03)	0.12(0.07)	0.19(0.10)	0.25(0.17)	0.27(0.13)
EC.4.1.1	S	0.64(0.09)	0.55(0.23)	0.48(0.11)	0.40(0.19)	0.19(0.21)	0.17(0.11)
	P	0.02(0.01)	0.03(0.01)	0.06(0.02)	0.07(0.03)	0.06(0.06)	0.15(0.07)
	F	0.05(0.03)	0.06(0.03)	0.11(0.03)	0.12(0.03)	0.09(0.09)	0.14(0.06)
EC.4.2.1	S	0.76(0.13)	0.67(0.20)	0.52(0.17)	0.44(0.18)	0.30(0.13)	0.22(0.08)
	P	0.04(0.02)	0.05(0.02)	0.07(0.03)	0.10(0.04)	0.12(0.06)	0.15(0.06)
	F	0.07(0.03)	0.09(0.04)	0.13(0.04)	0.15(0.05)	0.16(0.07)	0.17(0.05)
Médias	S	0.75(0.15)	0.70(0.18)	0.53(0.18)	0.51(0.20)	0.36(0.17)	0.28(0.15)
	P	0.04(0.02)	0.05(0.02)	0.08(0.04)	0.11(0.06)	0.15(0.08)	0.19(0.10)
	F	0.07(0.03)	0.10(0.04)	0.14(0.05)	0.18(0.07)	0.20(0.09)	0.21(0.10)

Tabela 7-10: Desempenho dos classificadores treinados com RIPPER após subamostragem da classe negativa. Valores de sensibilidade (S), precisão (P) e F-measure (F) médios, com respectivos desvios padrões, durante validação cruzada. Em destaque, proporções que levaram aos maiores F-measures por sub-subclasse.

C4.5			1:1	1:2	1:6	1:10	1:25	1:50
#CSR/#ÑCSR								
EC.1.1.1	S		0.74(0.19)	0.65(0.17)	0.49(0.17)	0.50(0.18)	0.31(0.17)	0.28(0.13)
	P		0.03(0.01)	0.04(0.02)	0.07(0.05)	0.09(0.03)	0.10(0.06)	0.15(0.10)
	F		0.06(0.01)	0.08(0.04)	0.12(0.07)	0.15(0.04)	0.15(0.08)	0.19(0.10)
EC.2.1.1	S		0.76(0.12)	0.53(0.22)	0.40(0.21)	0.29(0.12)	0.20(0.14)	0.09(0.09)
	P		0.02(0.01)	0.02(0.01)	0.04(0.03)	0.04(0.02)	0.06(0.05)	0.05(0.06)
	F		0.04(0.02)	0.05(0.02)	0.07(0.05)	0.07(0.03)	0.09(0.06)	0.06(0.07)
EC.2.3.1	S		0.74(0.15)	0.66(0.18)	0.48(0.18)	0.40(0.19)	0.28(0.13)	0.21(0.14)
	P		0.03(0.01)	0.04(0.01)	0.07(0.02)	0.07(0.03)	0.10(0.08)	0.13(0.06)
	F		0.05(0.02)	0.08(0.02)	0.11(0.04)	0.12(0.05)	0.13(0.09)	0.14(0.07)
EC.2.4.2	S		0.77(0.14)	0.57(0.21)	0.47(0.16)	0.35(0.11)	0.22(0.10)	0.17(0.10)
	P		0.03(0.01)	0.03(0.01)	0.06(0.02)	0.07(0.03)	0.08(0.03)	0.11(0.07)
	F		0.06(0.02)	0.06(0.03)	0.10(0.03)	0.11(0.04)	0.11(0.04)	0.13(0.08)
EC.2.5.1	S		0.77(0.15)	0.61(0.17)	0.40(0.20)	0.30(0.19)	0.16(0.15)	0.16(0.013)
	P		0.02(0.00)	0.03(0.01)	0.06(0.03)	0.07(0.05)	0.07(0.05)	0.21(0.24)
	F		0.04(0.01)	0.05(0.02)	0.10(0.06)	0.09(0.09)	0.09(0.07)	0.16(0.13)
EC.2.7.1	S		0.80(0.11)	0.69(0.16)	0.59(0.14)	0.47(0.13)	0.30(0.16)	0.23(0.26)
	P		0.04(0.01)	0.04(0.02)	0.07(0.04)	0.09(0.04)	0.08(0.06)	0.15(0.14)
	F		0.07(0.02)	0.08(0.04)	0.13(0.06)	0.15(0.06)	0.12(0.07)	0.15(0.11)
EC.3.1.1	S		0.77(0.14)	0.68(0.15)	0.51(0.16)	0.52(0.13)	0.41(0.12)	0.30(0.12)
	P		0.04(0.01)	0.06(0.02)	0.10(0.03)	0.14(0.06)	0.20(0.06)	0.24(0.08)
	F		0.08(0.03)	0.11(0.04)	0.16(0.05)	0.22(0.08)	0.26(0.07)	0.26(0.09)
EC.3.1.3	S		0.77(0.11)	0.71(0.20)	0.64(0.20)	0.52(0.15)	0.35(0.11)	0.29(0.15)
	P		0.05(0.02)	0.08(0.05)	0.12(0.06)	0.16(0.09)	0.19(0.10)	0.23(0.17)
	F		0.10(0.03)	0.14(0.07)	0.20(0.08)	0.23(0.09)	0.23(0.09)	0.25(0.14)
EC.3.2.1	S		0.90(0.06)	0.85(0.08)	0.69(0.06)	0.61(0.17)	0.59(0.08)	0.47(0.11)
	P		0.05(0.01)	0.07(0.02)	0.12(0.02)	0.13(0.03)	0.22(0.06)	0.26(0.07)
	F		0.09(0.02)	0.13(0.04)	0.21(0.03)	0.21(0.04)	0.32(0.06)	0.33(0.05)
EC.3.4.21	S		0.73(0.11)	0.62(0.08)	0.42(0.25)	0.40(0.21)	0.31(0.06)	0.27(0.16)
	P		0.04(0.02)	0.05(0.02)	0.07(0.04)	0.10(0.04)	0.13(0.04)	0.33(0.20)
	F		0.07(0.03)	0.10(0.03)	0.12(0.07)	0.15(0.07)	0.17(0.05)	0.28(0.16)
EC.4.1.1	S		0.60(0.14)	0.56(0.18)	0.43(0.19)	0.37(0.27)	0.27(0.19)	0.14(0.06)
	P		0.02(0.01)	0.02(0.01)	0.04(0.02)	0.05(0.03)	0.08(0.05)	0.12(0.06)
	F		0.03(0.01)	0.04(0.02)	0.07(0.03)	0.08(0.06)	0.13(0.08)	0.12(0.06)
EC.4.2.1	S		0.70(0.19)	0.59(0.14)	0.47(0.21)	0.36(0.16)	0.27(0.18)	0.11(0.07)
	P		0.03(0.01)	0.04(0.01)	0.07(0.03)	0.09(0.05)	0.14(0.08)	0.13(0.15)
	F		0.05(0.02)	0.07(0.03)	0.13(0.04)	0.14(0.07)	0.18(0.10)	0.11(0.09)
Médias	S		0.75(0.11)	0.64(0.13)	0.50(0.18)	0.42(0.17)	0.31(0.13)	0.23(0.12)
	P		0.03(0.01)	0.04(0.02)	0.07(0.03)	0.09(0.04)	0.12(0.06)	0.18(0.12)
	F		0.06(0.02)	0.08(0.03)	0.13(0.05)	0.14(0.06)	0.17(0.07)	0.18(0.10)

Observando-se os valores de sensibilidade e precisão para as Glicosidases (EC 3.2.1), nota-se que, neste caso, os classificadores adquirem um desempenho bem acima dos demais classificadores para outras sub-subclasses EC. Com máximo *F-measure* para a proporção de 1 resíduo de aminoácido catalítico para 50 não catalíticos, os classificadores treinados utilizando-se RIPPER para a sub-subclasse das Glicosidases atingem sensibilidade e precisão médias de 56% e 28%, respectivamente. No caso dos classificadores C4.5, sensibilidade e precisão médias atingem valores ligeiramente inferiores, de 47% e 26%, respectivamente. A sub-subclasse das Glicosidases é representada pelo maior número de cadeias proteicas e, conseqüentemente, maior número de CSR's, o que permite uma melhor aproximação da real distribuição desta classe de resíduos, assim com maior representatividade da classe positiva, sendo a única sub-subclasse a apresentar um valor de *F-measure* acima de 0.3, com sensibilidade e precisão consideráveis, além de baixo desvio padrão. Em adição, os desvios padrões em todos os casos, exceto para Glicosidases, apresentam valores elevados, indicando um desempenho bastante diversificado dos classificadores treinados durante a validação cruzada.

De forma análoga, explorou-se a amostragem da classe minoritária via sobreamostragem aleatória (ROS) e atribuição de pesos às amostras minoritárias, de forma que as proporções entre as classes nos conjuntos de treinamentos fossem de 1:1, 1:2, 1:6, 1:10, 1:25 e 1:50. Enquanto na sobreamostragem aleatória as amostras minoritárias são aleatoriamente amostradas com reposição, ocasionando a introdução de cópias das amostras positivas. O cálculo das métricas utilizadas pelos algoritmos é realizada utilizando-se o peso atribuído a cada amostra (padrão de 1.0 para todas as amostras), assim, a atribuição de pesos maiores do que 1.0 às amostras positivas simula uma amostragem com reposição. Por exemplo, caso seja atribuído peso igual de 2.0 a todas as amostras positivas, durante o cálculo do número de verdadeiros positivos, caso o número de amostras corretamente preditas como catalíticas seja igual a n , esta métrica apresentará um valor igual a $2n$, pois cada amostra contribui com peso 2.0.

De acordo com os resultados obtidos com a sobreamostragem, a introdução de cópias das amostras positivas aos conjuntos de treinamento não traz benefícios claros em relação aos classificadores treinados sem amostragem e subamostragem. Devido aos algoritmos empregarem uma fase de poda para aumentar a generalização das hipóteses construídas, a introdução de cópias prejudica a execução desta fase. No caso do RIPPER, tem-se 1/3 das amostras para realizar a poda das regras (*pruneSet*), enquanto 2/3 são utilizados para indução das regras (*growSet*). Se houver grande sobreposição entre as amostras nestes conjuntos, poucas regras serão podadas e o classificador terá sua capacidade de generalização reduzida. O mesmo não acontece no caso da atribuição de pesos, uma vez que os conjuntos *pruneSet* e *growSet* são independentes, por não existirem amostras duplicadas nos conjuntos de treinamento.

Em todos os casos, nota-se uma tendência similar entre os desempenhos dos classificadores, de

forma que, independentemente da técnica empregada (subamostragem ou sobreamostragem), o desempenho de cada sub-subclasse permanece muito similar, fornecendo mais indícios do impacto do número de amostras positivas nos conjuntos de dados considerados. Nos gráficos da Figura 7-17, fica evidente esse padrão no desempenho dos classificadores, onde o comportamento das linhas em cada um dos gráficos segue padrões muito similares, com quedas e picos concomitantes.

Ainda, o emprego do teste estatístico de Welch (Welch, 1947) para uma comparação entre os desempenhos dos classificadores de máximo desempenho com pré-processamento (sub e sobreamostragem) e os classificadores sem pré-processamento, não resultou em diferenças significativas entre as médias de seus *F-measures* para a maioria das sub-subclasses (Apêndice F-ii). No caso das sub-subclasses Hidrolases de ester carboxílico (EC.3.1.1), Serino Proteases (EC.3.4.21) e Monoester fosfórico hidrolases (EC.3.1.3), houve diferenças significativas entre os classificadores treinados com atribuição de pesos às amostras (RWOS) quando a proporção foi de 1:50 com valores-p de 1.06×10^{-3} , 3.28×10^{-4} e 4.8×10^{-2} , no teste de Welch, respectivamente.

No caso do algoritmo evolutivo a para seleção de amostras (EA-TSS), apesar de fornecer resultados ligeiramente superiores aos apresentados pelos classificadores sem amostragem, os resultados dos testes estatísticos (teste-t de Welch) não apresentaram diferenças entre as predições realizadas (Apêndice F-iii).

De todos os classificadores avaliados, aqueles treinados com proporção entre as classes de 1:50 apresentaram no geral melhores resultados, comparando-se com outras proporções, além de desempenhos equivalentes aos classificadores sem pré-processamento. Uma vez que subamostragem e sobreamostragem aleatórias não introduzem novas informações aos dados (no caso de subamostragem a informação é reduzida), foi realizada a introdução de novas amostras sintéticas, criadas a partir do SMOTE, como uma tentativa de melhorar o desempenho dos classificadores. Os resultados da introdução das amostras sintéticas estão apresentados na Seção 7.8.

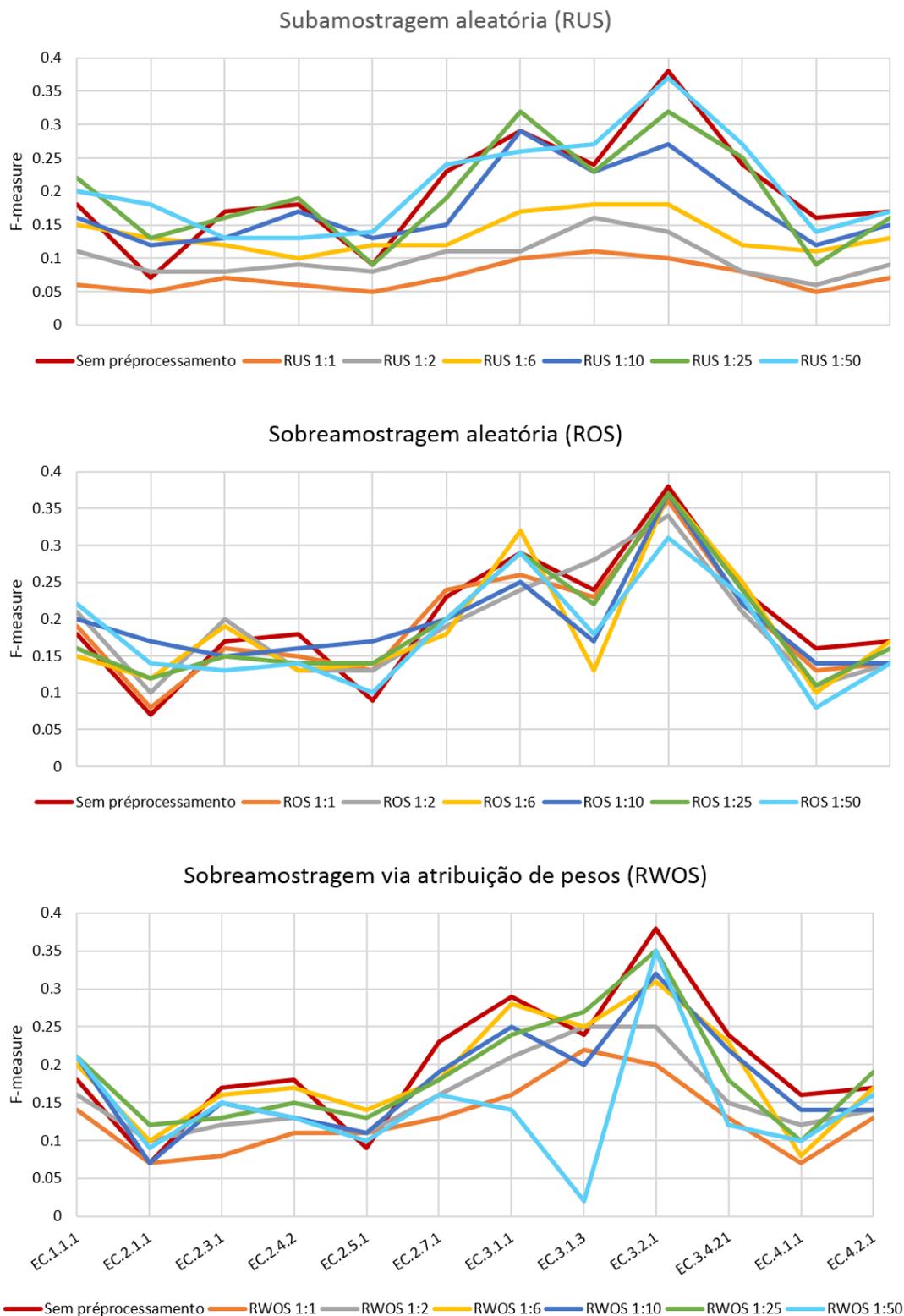


Figura 7-17: Comparação entre o desempenho dos classificadores treinados com subamostragem aleatória (RUS), sobreamostragem aleatória (ROS) e atribuição de pesos (RWOS) para diferentes proporções entre as classes. Nota-se um comportamento semelhante dos desempenhos dos classificadores nos três gráficos, indicando uma dependência ao conjunto de treinamento.

Comparando-se o desempenho dos classificadores treinados sem amostragem com aqueles que obtiveram maior desempenho com subamostragem e sobreamostragem, não percebe-se ganhos gerais no *F-measure* com uso de tais técnicas, como pode ser visto pelo gráfico da Figura 7-18. Ainda que em alguns casos haja um ligeiro aumento no desempenho, a remoção de amostras ou introdução de cópias e pesos diferentes não aliviam o desbalanceamento, nem elevam a representatividade da classe positiva.

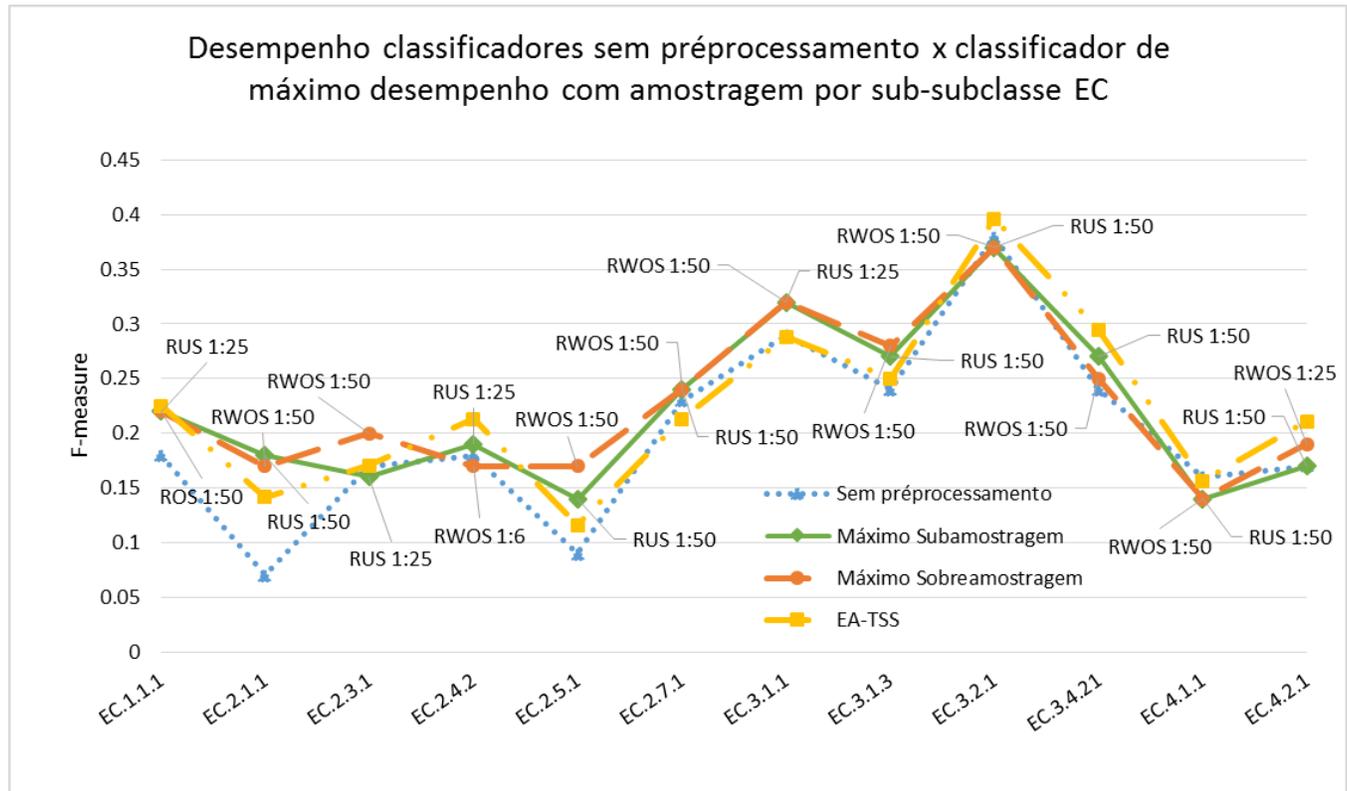


Figura 7-18: Comparação de desempenho dos classificadores treinados sem pré-processamento com aqueles que obtiveram máximo desempenho utilizando subamostragem e sobreamostragem, para cada sub-subclasse. Não é perceptível diferença no desempenho com o uso de tais técnicas.

7.7 Ensembles de classificadores via *Boosting* e *Bagging*

Muitos algoritmos têm sido propostos para superar os desafios de classificação em problemas desbalanceados, dentre estes encontram-se técnicas de *boosting* e *bagging*.

O método RUSBoost, descrito na seção 5.2, realiza a subamostragem da classe majoritária para treinar diversos classificadores, utilizando a técnica de *boosting*, e construir um *ensemble* contendo classificadores treinados em conjuntos de dados balanceados (ou com menor desbalanceamento). Foram utilizados para a construção de modelos classificadores os doze conjuntos de dados compostos por enzimas de mesma sub-subclasse EC, com proporções entre as classes minoritária e majoritária idênticas às utilizadas anteriormente, ou seja, 1:1, 1:2, 1:6, 1:10, 1:25 e 1:50.

Similar aos classificadores da seção anterior, taxas de subamostragem que geram uma distribuição

mais homogênea entre as classes (1:1 e 1:2) levam à remoção de grande número de amostras da classe majoritária, ocasionando a redução de desempenho dos classificadores. No entanto, ainda é possível notar a mesma tendência de desempenho das sub-subclasses EC, de forma que a construção de *ensembles* via *boosting* e subamostragem não traz ganhos de desempenho em relação ao classificador treinado sem pré-processamento, como pode ser visto no gráfico da Figura 7-19. Percebe-se ainda pelo gráfico que o desempenho dos classificadores para as diversas sub-subclasses EC variam diferentemente de acordo com a taxa de subamostragem. No entanto, para proporções mais balanceadas (1:1 e 1:2), o desempenho é sempre inferior ao classificador treinado sem pré-processamento. Sendo assim, não existe uma proporção única para todos os conjuntos de dados, e cada sub-subclasse mostra um melhor desempenho para diferentes taxas de subamostragem, mesmo que as diferenças não sejam estatisticamente significantes em relação ao classificador sem amostragem, segundo teste de Welch com significância de 5% (Apêndice F-iv). O único caso em que o teste estatístico indicou melhora no desempenho foi para as enzimas Metiltransferases (EC.2.1.1) e empregado o classificador RUSBoost treinado com proporção de 1:10 entre as classes (valor-p = 0.0214).

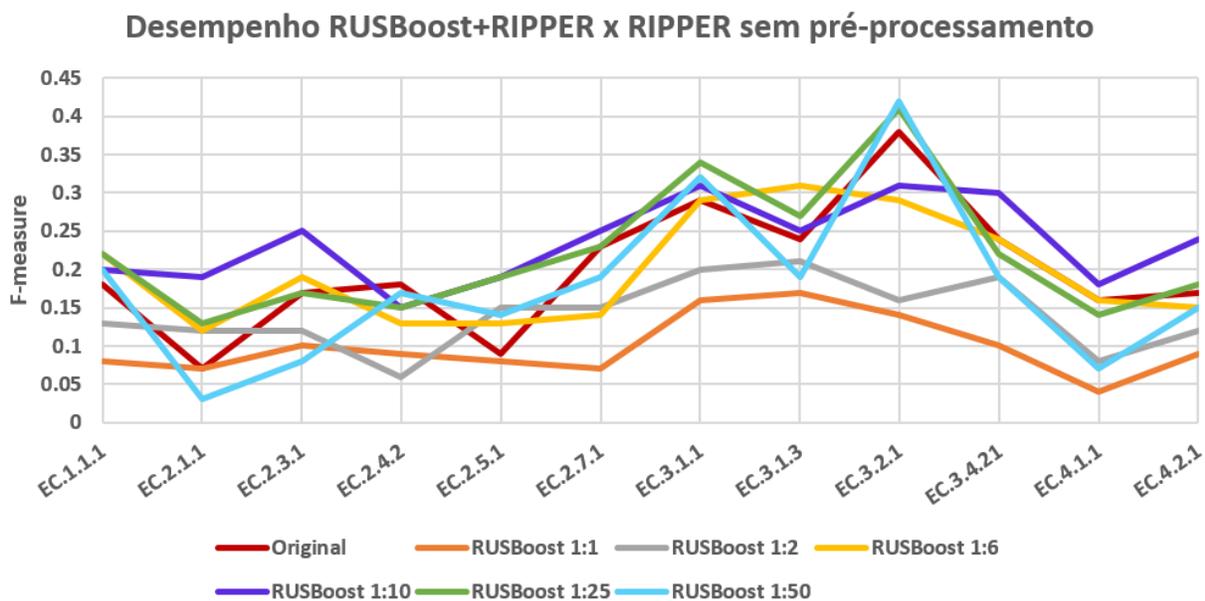


Figura 7-19: Comparação de desempenho dos classificadores treinados utilizando RUSBoost para diferentes taxas de subamostragem e o classificador sem pré-processamento (Original). Exceto em alguns casos (EC. s 1.1.1, 2.1.1, 2.3.1 e 2.5.1) os classificadores treinados pelo RUSBoost não apresentaram maiores F-measures do que o classificador sem pré-processamento. No entanto, mesmo nos casos em que houve ganhos, estes não foram estatisticamente significantes.

Novamente, o número de amostras da classe minoritária para grande parte das sub-subclasses é muito baixo para que algoritmo de indução de regras e árvores de decisão possam apresentar desempenho satisfatório. Como exceção, tem-se a sub-subclasse das Glicosidases (EC.3.2.1) onde o número de

amostras catalíticas, ainda que baixo, é o maior dentre todos os subconjuntos (148 CSR's, ver Tabela 3-1 na página 28). Para esta sub-subclasse de enzimas, percebe-se um melhor desempenho dos classificadores treinados em todas as configurações utilizadas, mostrando uma boa relação entre sensibilidade e precisão, resultando em valores de *F-measure* superiores a 0.3 em grande parte dos casos e com menor desvio padrão no desempenho.

O aumento do número de amostras da classe minoritária, de outras formas que não a sobreamostragem aleatória, pode, portanto, trazer ganhos aos classificadores.

7.8 Redução de dimensionalidade e Sobreamostragem via SMOTE

Apesar do algoritmo RIPPER, assim como o C4.5, realizar a seleção de atributos internamente (antecedentes das regras), suas escolhas estão sujeitas à eficiência da métrica empregada pelo algoritmo. No caso do RIPPER, é empregada uma modificação do ganho de informação, enquanto que no C4.5 emprega-se a taxa de ganho de informação. Estas duas métricas baseiam-se na entropia da informação do conjunto resultante do emprego de uma condição sobre um certo atributo. No entanto, existem muitas outras métricas que podem ser empregadas, conduzindo assim a diferentes escolhas e soluções para os algoritmos.

Antes de realizar a sobreamostragem da classe minoritária através de algum método baseado em distâncias entre as amostras, é preciso realizar uma redução da dimensionalidade do problema, a fim de amenizar problemas na formulação de vizinhança em altas dimensionalidades. Dessa forma, antes do emprego da técnica de sobreamostragem para a introdução de amostras sintéticas, conhecida como SMOTE, deve-se selecionar os atributos de forma a reduzir a dimensionalidade dos dados e deixar o problema mais factível para a definição de uma vizinhança.

Para a seleção de atributos, utilizou-se a técnica de *wrapper* com busca sequencial incremental (*forward feature selection*), como descrita na Seção 6.3. A redução de dimensionalidade foi realizada antes da aplicação do SMOTE e para diferentes taxas de sobreamostragem: de 100% a 1000%, com incrementos de 100%, de forma que uma taxa de $k\%$ resulte na introdução de $n_p * k / 100$ amostras sintéticas, sendo n_p o número de amostras positivas.

Além do uso do RIPPER, foram também avaliados os métodos SMOTEBoost e SMOTEBagging. O algoritmo C4.5 não foi avaliado devido ao seu desempenho ter se mostrado equivalente ao do RIPPER nas etapas anteriores.

A sobreamostragem utilizando o SMOTE, ao contrário da sobreamostragem aleatória e a atribuição de pesos, trouxe ganhos aos classificadores treinados, como pode ser visto pelo gráfico da Figura 7-20. No gráfico, encontram-se os desempenhos (*F-measure*) dos métodos SMOTEBagging, SMOTEBoost,

RIPPER + SMOTE e RIPPER, todos treinados após a redução de dimensionalidade, e o RIPPER sem pré-processamento. No caso dos classificadores SMOTEBoost e RIPPER+SMOTE, onde variou-se a taxa de sobreamostragem do SMOTE de 100% a 1000%, o gráfico mostra o *F-measure* dos classificadores com maior desempenho para cada sub-subclasse EC. No entanto, uma relação completa dos resultados encontra-se no Apêndice H.

Apesar da redução de dimensionalidade viabilizar o uso do método SMOTE e, conseqüentemente, a introdução de amostras sintéticas promover melhores proporções entre as classes de resíduos de aminoácidos, não é notável uma melhora no desempenho para todas as sub-subclasses EC. No caso das enzimas Glicosidades (EC 3.2.1), Carboxilesterases (EC. 3.1.1), Serino Proteases (EC. 3.4.21), Carboxiliases (EC 4.1.1), as Transferases da sub-subclasse EC. 2.5.1 e as Oxirredutases da sub-subclasse EC 1.1.1, percebe-se aumento significativo do desempenho dos classificadores, que pôde ser confirmado pelo teste estatístico de Welch (Apêndice F-v). Para as demais sub-subclasses, ainda que possa haver aumento no desempenho, estes não são estatisticamente significantes e podem ser considerados similares ao classificador treinado utilizando-se somente o algoritmo RIPPER, sem amostragem e sem redução de dimensionalidade.

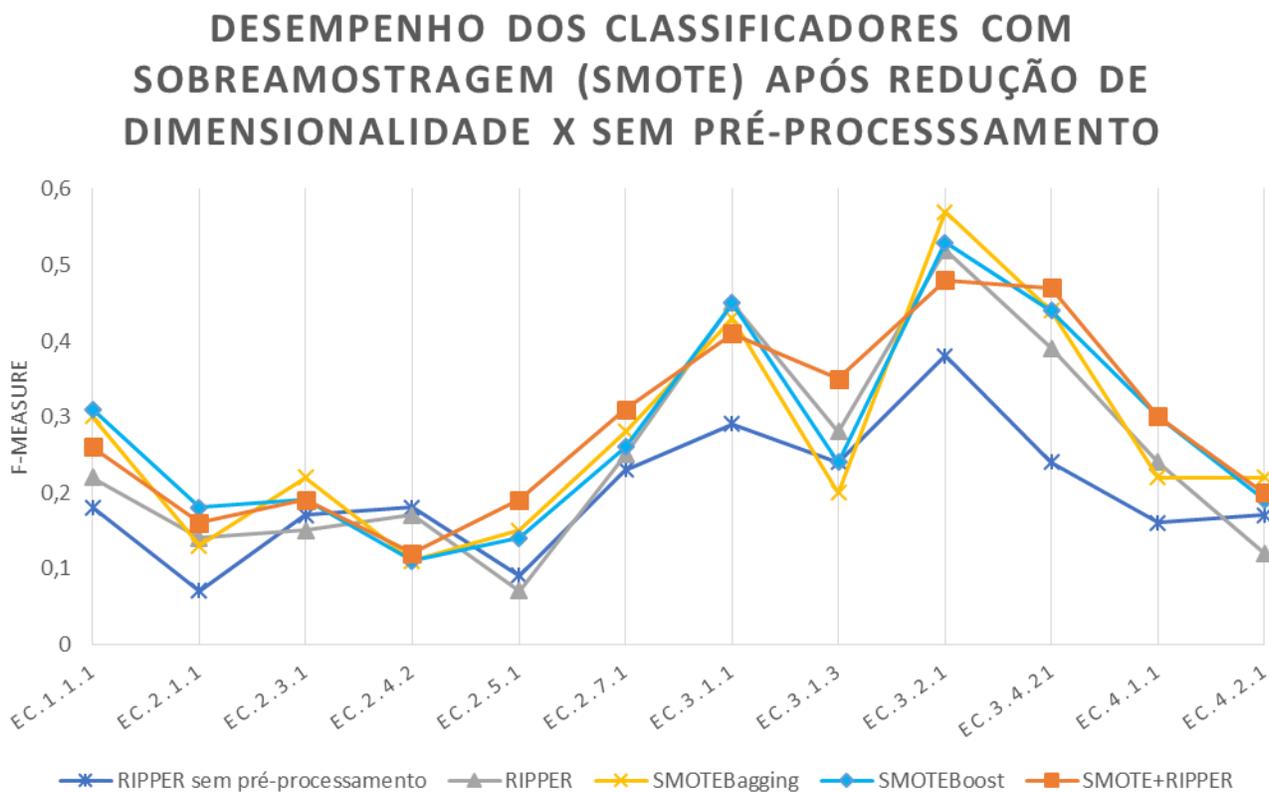


Figura 7-20: Comparação de desempenho dos classificadores treinados aplicando SMOTE após redução de dimensionalidade e o classificador sem pré-processamento. O uso de SMOTE e a redução de dimensionalidade trouxeram ganhos expressivos em algumas sub-subclasses.

A lista dos descritores selecionados pelo algoritmo de seleção de atributos, para cada sub-subclasse, encontra-se no Apêndice I. Em alguns casos, por conta do baixo número de amostras catalíticas, os conjuntos de atributos selecionados envolvem pouquíssimos descritores. Tendo em vista que nesses casos observou-se também um baixo desempenho dos classificadores, entende-se que a busca por um subconjunto de atributos com maior desempenho sofreu também com baixa representação da classe de resíduos de aminoácidos catalíticos. Devido ao baixo número de amostras catalíticas, os classificadores tendem a apresentar desempenhos similares, pois mesmo com o emprego de subamostragem ou sobreamostragem aleatórias, a redução de dimensionalidade ainda sofrerá com a baixa amostragem, já que utiliza o classificador para avaliar o desempenho de um subconjunto de atributos. Não haverá, portanto, grandes diferenças entre os classificadores treinados com apenas um atributo e aqueles treinados com todos os atributos ou qualquer outro subconjunto de atributos. Ainda que o algoritmo selecione o subconjunto de melhor desempenho, esta escolha ficará sujeita a subotimalidade e, ao final, o subconjunto escolhido não trará ganhos aos classificadores.

7.9 Descritores de conservação na predição de resíduos catalíticos

Por serem funcionalmente importantes, resíduos de aminoácidos catalíticos são bastante conservados nas estruturas primárias das enzimas. A maioria dos estudos realizados para predição de resíduos catalíticos utiliza um ou mais descritores de conservação de aminoácidos nas sequências proteicas. Apesar de inicialmente estas informações não terem sido utilizadas, uma vez que a conservação de um aminoácido está ligada a um conjunto de proteínas homólogas e não representa uma característica individual do resíduo de aminoácido, avaliou-se nesta etapa o impacto destes descritores na predição dos resíduos de aminoácidos catalíticos. O algoritmo RIPPER foi novamente utilizado para indução de regras de classificação, considerando descritores de conservação de sequência além dos descritores estruturais anteriormente utilizados. Antes de realizar o treinamento dos classificadores, uma redução da dimensionalidade foi realizada seguindo as mesmas configurações anteriores, ou seja, utilizou-se busca sequencial incremental juntamente com o RIPPER, visando selecionar os subconjuntos de atributos mais relevantes para a predição. Após as escolhas dos subconjuntos de atributos, todos os classificadores anteriores foram treinados novamente, incluindo subamostragem, sobreamostragem, *boosting*, *bagging* e amostragem híbrida (subamostragem seguida de sobreamostragem).

A introdução de descritores de conservação proporcionou melhores resultados, comparados aos classificadores treinados pelo RIPPER sem pré-processamento, o que pôde ser confirmando através do teste estatístico de Welch para comparação entre médias. Para comparar os desempenhos dos classificadores que incorporam descritores de conservação com os demais, foi selecionado o classificador

de maior desempenho de cada configuração. Dessa forma, quatro classificadores com os maiores F-measure's foram selecionados, para quatro configurações avaliadas: subamostragem, seleção de atributo + SMOTE, uso de descritores de conservação e sem pré-processamento. Os resultados para cada sub-subclasse EC encontram-se nos gráficos da Figura 7-21, contendo medidas de sensibilidade, precisão e máximo F-measure de cada configuração.

De forma geral, a introdução de conservação trouxe melhorias nos desempenhos dos classificadores treinados para todas as sub-subclasses, quando comparados aos classificadores sem pré-processamento. Considerando-se os classificadores com uso de RUS e FS+SMOTE, os classificadores com uso de conservação trouxeram ganhos para a maioria das sub-subclasses em relação ao RUS (7 de 12), e em relação ao FS+SMOTE os ganhos foram menores (3 de 12), sendo que em 8 sub-subclasses o desempenho de ambas as configurações foi similar.

Por outro lado, as sensibilidades e precisões dos classificadores treinados com e sem descritores de conservação mostraram-se diferentes. Apesar de possuírem *F-measures* muito próximos, estes valores são derivados de diferentes contribuições da sensibilidade e precisão. Enquanto classificadores treinados com descritores de conservação mostraram maiores taxas de sensibilidade do que outros classificadores, suas taxas de precisão ficaram abaixo das dos classificadores sem uso de conservação (FS+SMOTE). Dado que o *F-measure* é uma média harmônica entre sensibilidade e precisão, classificadores que mostrarem compromissos semelhantes entres essas métricas tendem a possuir maiores valores de *F-measure*. Dessa forma, mesmo que a conservação eleve a sensibilidade, como o mesmo não ocorre com a precisão, onde verifica-se uma redução em relação aos outros classificadores, estes classificadores mostram desempenho similar aos treinados sem conservação.

Ainda que a introdução de conservação resulte em melhor desempenho, é a redução de dimensionalidade seguida de sobreamostragem (SMOTE) que fornece melhores resultados, mesmo quando não se considera o uso de descritores de conservação (FS+SMOTE).

Assim, com a introdução de descritores de conservação, é possível elevar as sensibilidades dos classificadores, mas com uma redução de suas precisões. Em alguns casos, essa redução é mais elevada do que em outros. Porém, F-measure's similares podem ser encontrados em outros classificadores treinados sem conservação.

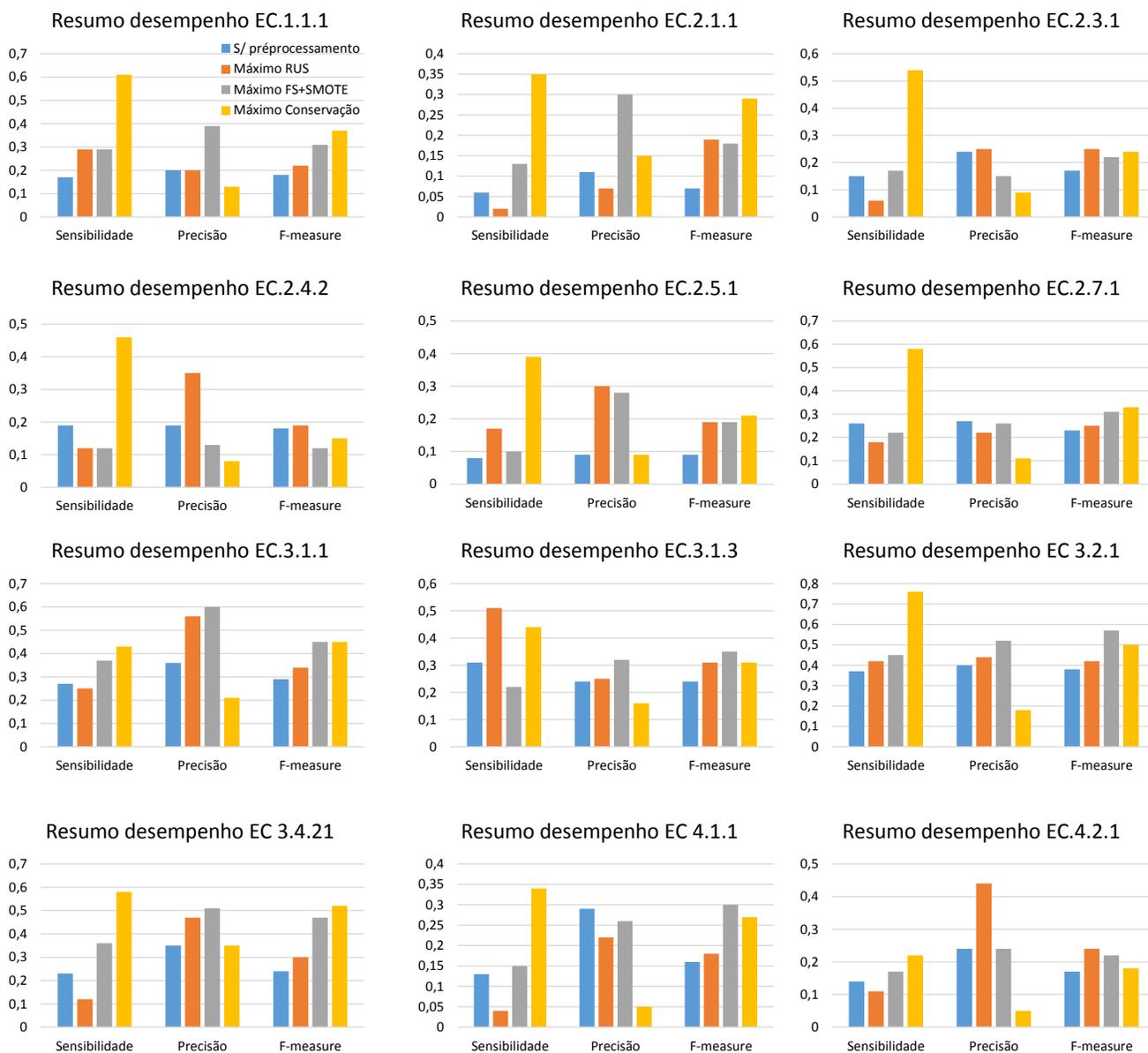


Figura 7-21: Desempenho dos melhores classificadores treinados para quatro configurações diferentes: Sem pré-processamento, RUS, FS+SMOTE e FS+SMOTE com conservação para cada sub-subclasse. Descritores de conservação elevam a sensibilidade dos classificadores, porém com redução na precisão, resultando em F-measures similares ao caso de uso do FS+SMOTE sem descritores de conservação.

7.10 Desafios à predição de resíduos de aminoácidos catalíticos

A partir dos resultados apresentados nas seções anteriores, fica evidente a dificuldade de se encontrar regras que possam ser utilizadas para a predição de resíduos de aminoácidos catalíticos. Uma análise mais detalhada dos desempenhos de cada classificador, para cada conjunto de enzimas e reações, fornece indícios de que os classificadores enfrentam grandes dificuldades em lidar com dados desbalanceados. No entanto, à medida que o número de resíduos catalíticos aumenta, percebe-se uma melhora nos desempenhos dos classificadores, como pode ser observado no caso das Glicosidades (EC. 3.2.1) e

Carboxilesterases (EC. 3.1.1).

Durante o treinamento dos classificadores utilizando-se o RIPPER, o conjunto de amostras é dividido em dois subconjuntos, sendo que 1/3 das amostras é utilizado para poda das regras e os 2/3 restantes para indução das regras. Quando o número de amostras positivas é baixo, a distribuição da classe de resíduos catalíticos nos 2/3 dos dados apresenta diferenças significativas aos outros 1/3 dos dados. Assim, as regras inicialmente induzidas pelo algoritmo, com alta cobertura e precisão, sofrem uma redução destas capacidades ao serem aplicadas ao conjunto de poda. Sendo realizadas diversas remoções de antecedentes na fase de poda, a hipótese inicial gerada pela regra é drasticamente modificada. No entanto, não realizar a poda das regras leva a soluções altamente específicas e com capacidade de generalização ainda menores (sobreajuste dos classificadores).

Além disto, a baixa quantidade de amostras positivas não fornece uma boa aproximação da real distribuição dos resíduos de aminoácidos catalíticos, influenciando diretamente o processo de aprendizado do algoritmo de indução de regras e do algoritmo de construção de árvores de decisão. Estes algoritmos requerem uma grande quantidade de amostras para serem eficazes e fornecerem regras com boa capacidade de generalização. Com poucas amostras, os ganhos de informação obtidos durante a escolha dos antecedentes são muito próximos uns dos outros, sendo quase impossível encontrar um antecedente com um valor muito discrepante dos demais, de forma que sua escolha possa ser justificada, reforçando sua escolha pelo algoritmo. No entanto, o algoritmo ainda seleciona o atributo com maior ganho, mesmo que este ganho não seja muito diferente dos demais. Dessa forma, pequenas variações nos conjuntos de dados (remoção de algumas poucas amostras positivas) podem ocasionar mudanças significativas nas regras induzidas, uma vez que pequenas modificações nos ganhos de informação são introduzidas. Este fenômeno pode ser melhor observado através dos gráficos das Figuras 7-21 e 7-22 onde encontram-se os ganhos de informação dos 10 melhores atributos para as sub-subclasses das Glicosidades (EC 3.2.1), com maior número de CSR's, e Metiltransferases (EC 2.1.1), com menor número de CSR's. Para a construção dos gráficos, foram consideradas 11 execuções do RIPPER, sendo a primeira delas considerando-se todo o conjunto de enzimas de cada sub-subclasse, e as outras execuções considerando cada *fold* de uma validação cruzada 2x5. Assim, é possível comparar as mudanças ocorridas nas escolhas dos atributos quando o classificador é treinado utilizando-se todos os dados, e quando é treinado com parte dos dados, da mesma forma como fora feito nas seções anteriores. Considerando-se somente a escolha do primeiro atributo, uma vez que esta irá impactar nas escolhas posteriores, os atributos foram ordenados segundo seus ganhos de informação e os 10 melhores de cada execução foram escolhidos para avaliar a variação e a ordem das escolhas pelos melhores atributos. Na Figura 7-22, tem-se o número de vezes que um dado atributo apareceu nas primeiras 10 posições, segundo

as 11 execuções realizadas para as duas sub-subclasses. Por exemplo, comparando-se as duas sub-subclasses, tem-se que no caso das Metiltransferases os atributos *apolarASA*, *score* (do *pocket*) e *cpo_85_30_LHA_VD* compartilham a primeira escolha do algoritmo (barras azuis no gráfico), sendo que o atributo *apolarASA* apareceu como primeiro antecedente das regras em 5 execuções de 11; o descritor *score* apareceu em 3 execuções e o descritor *cpo_85_30_LHA_VD* aparece as outras 2 execuções, sendo que em apenas uma execução nenhum dos três descritores é selecionado para compor uma regra.

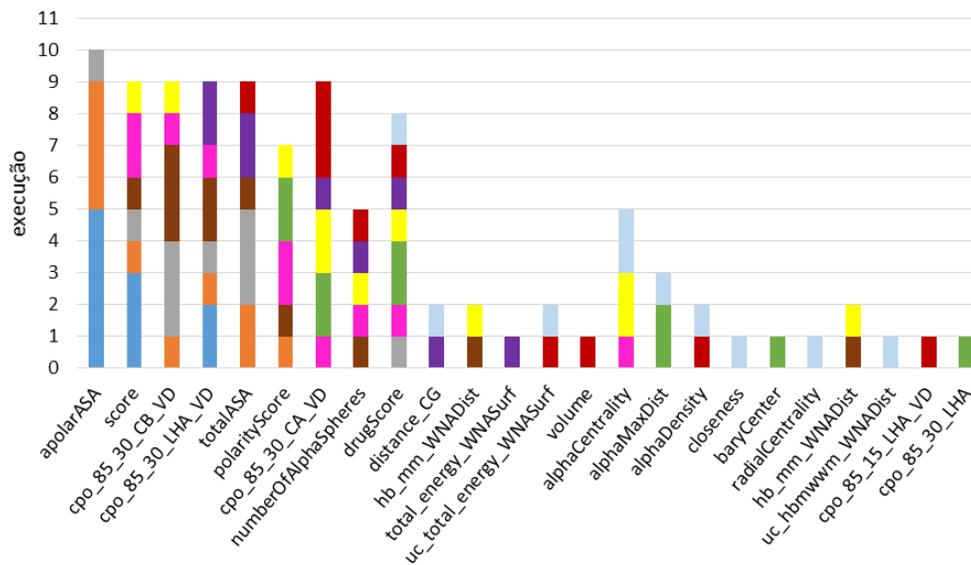
Por outro lado, no caso das Glicosidades, o atributo *localCloseness* aparece em todas as 11 execuções na primeira posição (azul), como pode ser visto pela barra totalmente azul para este atributo. Assim, durante o treinamento, no caso das Metiltransferases, cinco regras teriam como primeiro antecedente o descritor *apolarASA*, três regras o descritor *score* e, duas o descritor *cpo_85_30_LHA_VD*, de forma que as regras induzidas pelo algoritmo seriam bem diferentes considerando as iterações futuras. Enquanto isso, no caso das Glicosidades o atributo de *localCloseness* aparecerá em todas as regras induzidas como o primeiro antecedente das regras geradas.

Uma comparação entre os gráficos revela que, no caso das Glicosidades, as escolhas sofreram menores variações durante as 11 execuções do que no caso das Metiltransferases, uma vez que um menor número de atributos foi escolhido durante as 11 execuções, e ainda pelo gráfico apresentar menor variação de cores (posições) para cada atributo. Como as escolhas posteriores são diretamente influenciadas pelas escolhas anteriores, no caso das Metiltransferases, perceber-se-ia conjuntos de regras diferentes durante as execuções, ainda que não se possa confirmar se o mesmo ocorreria para as Glicosidades. Somente o fato do descritor de *localCloseness* ser escolhido para incorporar as regras fornece pistas sobre como estas escolhas impactam nas regras induzidas e na variabilidade das mesmas.

Já nos gráficos da Figura 7-23, encontram-se os ganhos de informação normalizados dos 10 melhores atributos selecionados em cada execução. Novamente, nota-se que, no caso das Glicosidades, as escolhas pelo melhor atributo sofreram menor variação do que em relação às Metiltransferases. Pelas diferenças entre os tamanhos dos *boxplots* (diferença entre primeiro e terceiro quartil), verifica-se que o ganho de informação sofreu maior variação no caso das Metiltransferases, além de um maior número de atributos ter aparecido no resultado final, devido às alterações nas escolhas dos 10 melhores atributos em cada execução.

Acredita-se que, com isto, seja possível notar as alterações nas escolhas dos melhores atributos segundo o ganho de informação e observar o impacto do número de amostras no cálculo desta métrica. Dessa forma, um aumento do número de resíduos de aminoácidos catalíticos traria benefícios diretos a todas as sub-subclasses, como consequência da redução da variação das escolhas realizadas pelo RIPPER, desde que as novas amostras não se dispersassem ainda mais pelo espaço de variáveis.

Posições dos atributos relativas ao ganho de informação
 Metiltransferases – EC.2.1.1 (#CSR's: 42 de 4979)



Glicosidases – EC.3.2.1 (#CSR's: 148 de 21764)

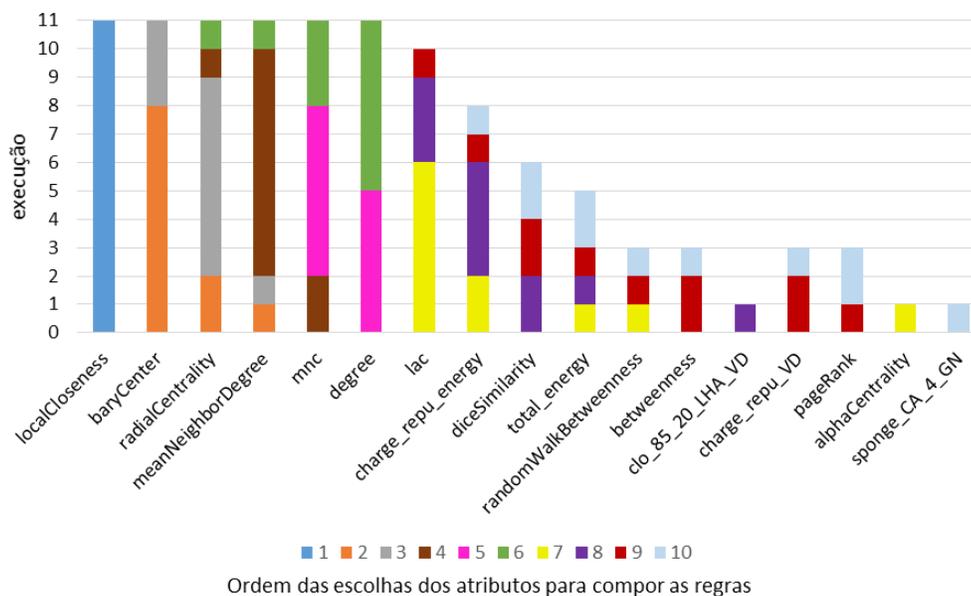


Figura 7-22: Número de vezes em que os atributos apareceram nas posições de 1 a 10 considerando as 11 execuções para as duas sub-subclasses: Glicosidases (EC.3.2.1) e Metiltransferases (EC.2.1.1). Escolhas dos melhores atributos para Metiltransferases mostram maior variação em relação às Glicosidases.

Ganhos de informação para os 10 melhores atributos de cada execução

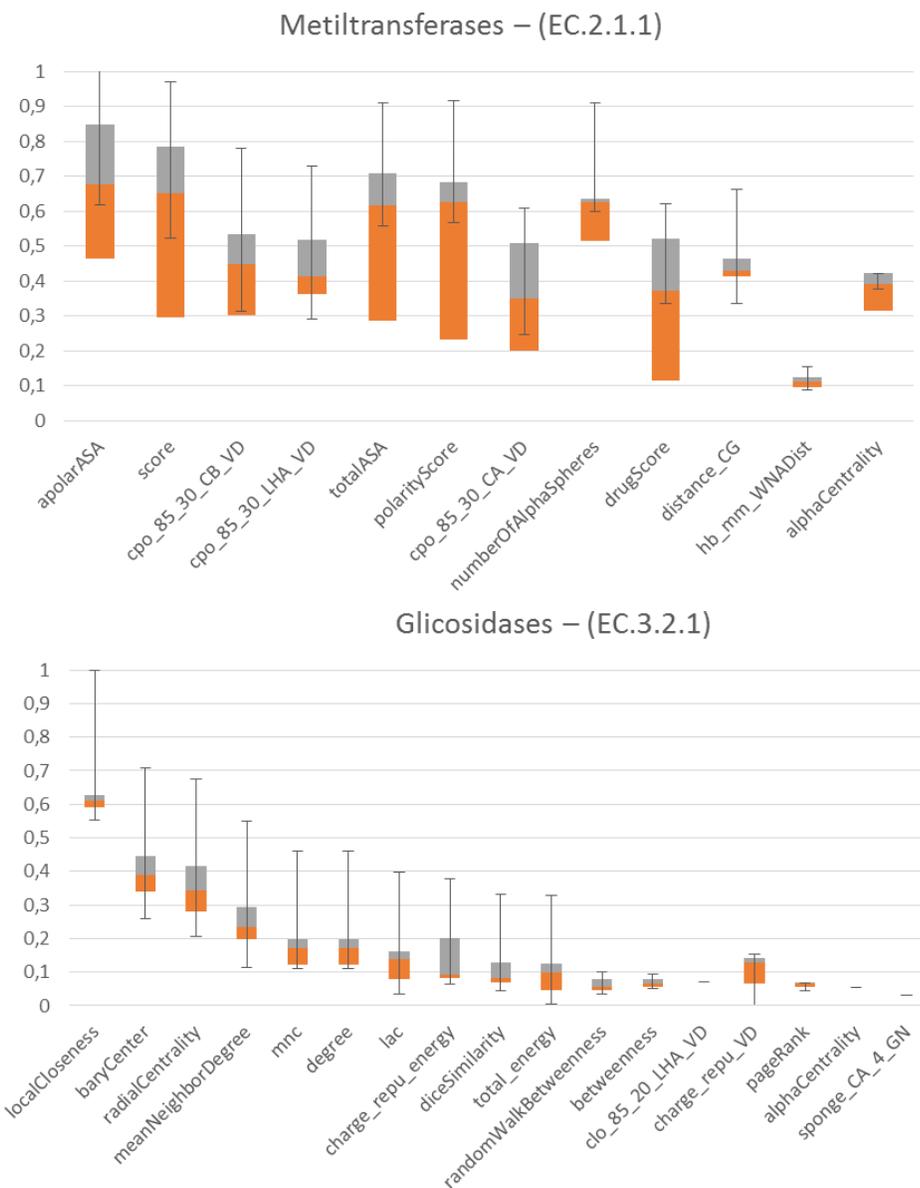


Figura 7-23: Gráfico tipo boxplot para ganho de informações dos 10 melhores atributos, considerando as 11 execuções do algoritmo RIPPER para as sub-subclasses das Glicosidases (EC.3.2.1) e Metiltransferases (EC.2.1.1). Os ganhos de informação dos atributos para o caso das Metiltransferases sofrem maior variação em relação às Glicosidases (diferença entre primeiro e terceiro quartil de cada boxplot e maior número de atributos).

Os descritores que figuram nas primeiras escolhas do algoritmo envolvem aqueles ligados à geometria das cadeias proteicas. Os descritores de *local closeness* e *barycenter* fornecem indicações da localização dos resíduos de aminoácido catalíticos na superfície da proteína. Enquanto valores elevados de *local closeness* indicam a presença do resíduo de aminoácido em uma cavidade na superfície da molécula, o descritor de *barycenter* apresenta valores menores para aqueles resíduos localizados em

posições mais centrais em relação ao centro de geométrico da proteína. Dessa forma, as primeiras escolhas do algoritmo sugerem que as regras estarão preferencialmente selecionando resíduos de aminoácido em regiões próximas ao centro geométrico da proteína que apresentem-se como uma cavidade. Esta é uma propriedade que foi observada para diversos CSR's de diferentes sub-subclasses, sendo portanto, uma característica global dos CSR's, ao invés de uma característica local e específica de cada sub-subclasse (Mitternacht & Berezovsky, 2011).

Dessa forma, uma melhor amostragem da classe positiva pode trazer ganhos, justamente por proporcionar ao algoritmo escolhas locais com maiores relevâncias e contribuição para que sucessivas iterações levem às soluções ótimas globais ou o mais próximo destas. Uma vez que alguns descritores irão apresentar distribuições mais homogêneas para os CSR's, estes terão um ganho de informação mais discrepante que outros descritores, fortalecendo sua escolha como antecedente de uma regra e aumentando sua capacidade de generalização.

7.11 Expansão do número de enzimas via alinhamentos estruturais

Há uma tendência natural dos algoritmos de indução de regras em criar regras bastante específicas, com baixo suporte e alta confiança, devido à fragmentação da classe positiva. Apesar do pré-processamento aumentar o número de amostras da classe positiva, existem algumas desvantagens com estas abordagens. Primeiramente, a sobreamostragem aleatória não introduz novas amostras, sendo somente as amostras existentes replicadas, o que equivale a fornecer um peso maior às amostras positivas. Como consequência, a classe positiva continua mal amostrada e confinada a pequenas regiões do espaço, pois a replicação não aumenta a dispersão dessas amostras. Outras técnicas, como SMOTE e LN-SMOTE, utilizam amostras vizinhas para criar novas amostras sintéticas próximas a estas. Ainda que seja possível expandir o número das amostras positivas e, de certa forma, melhorar a dispersão dos dados, criando regiões positivas de maior densidade, considerando-se a alta dimensionalidade dos dados, a noção de vizinhança torna-se limitada. Assim, o uso de tais técnicas em espaços de alta dimensão torna-se ineficiente e acaba por produzir novas amostras em regiões diversas do espaço (Blagus & Lusa, 2013). No entanto, a redução de dimensionalidade através de algum método para seleção de atributos pode trazer ganhos expressivos, quando combinados a tais técnicas. Como apresentado na seção 7.8, a sobreamostragem utilizando o SMOTE após a redução de dimensionalidade traz ganhos expressivos aos classificadores. No entanto, devido ao baixo número de amostras positivas, a seleção do melhor subconjunto de atributos fica também sujeita aos mesmos problemas.

Outra forma, mais natural, de aumentar o número de amostras positivas, seria então introduzir amostras provenientes de outras enzimas não contidas na base de dados do CSA. Através de alinhamentos

estruturais (superposição) entre as cadeias enzimáticas contidas no CSA e aquelas sem anotação de resíduos catalíticos contidas no PDB, é possível expandir o número de cadeias no conjunto de dados e, conseqüentemente, aumentar o número de amostras positivas.

Ainda que a introdução de novas cadeias também cause um aumento do número de amostras negativas, o aumento do número de amostras positivas proporciona um aumento da densidade de regiões positivas no espaço amostral, de forma que as regras geradas pelos algoritmos possam, então, englobar regiões com maiores volumes, cobrindo um maior número de amostras positivas e realizando a separação dessas das amostras negativas.

Um alinhamento estrutural consiste em realizar a superposição de estruturas terciárias, de forma a minimizar as distâncias interatômicas das cadeias através da minimização da raiz do erro quadrático médio (RMSD) entre as estruturas. Através da superposição de estruturas e através do valor do RMSD, é possível avaliar como duas ou mais estruturas se relacionam do ponto de vista estrutural (conformação tridimensional). Assim, mesmo estruturas proteicas com seqüências muito dissimilares podem apresentar uma superposição estrutural com baixo RMSD, indicando uma relação evolutiva entre elas. Para realizar os alinhamentos estruturais, utilizou-se o software *MultiProt*, que oferece soluções acuradas com extrema rapidez (Shatsky, et al., 2002).

Considerando-se somente alinhamentos estruturais entre cadeias contendo os mesmos três primeiros dígitos do número EC (mesma sub-subclasse) e com similaridades sequenciais abaixo de 40%, as anotações contidas no CSA puderam ser transferidas nos casos em que todos os resíduos de aminoácidos catalíticos anotados foram alinhados com resíduos de aminoácido de mesmo tipo. Ou seja, para se transferir a anotação de uma cadeia contendo três resíduos catalíticos (e.g. HIS, SER, ASP), estes devem estar alinhados com outros três resíduos de aminoácidos de mesmo tipo, formando os pares: HIS-HIS, SER-SER, e ASP-ASP. Alinhamentos onde a raiz do erro quadrático médio (RMSD) entre as posições dos pares de átomos alinhados foram maiores do que 2.0Å foram descartados. Considerou-se, portanto, somente os alinhamentos com alta similaridade e manualmente curados para averiguar a correta transferência de anotações, fornecendo um incremento de 654 cadeias distribuídas entre as 12 sub-subclasses EC consideradas. Uma comparação entre o número de cadeias antes e após a expansão dos dados, para as diferentes sub-subclasses EC, encontra-se no gráfico da Figura 7-24.

Número de cadeias por sub-subclasse EC antes e após as transferências de anotações por alinhamentos estruturais

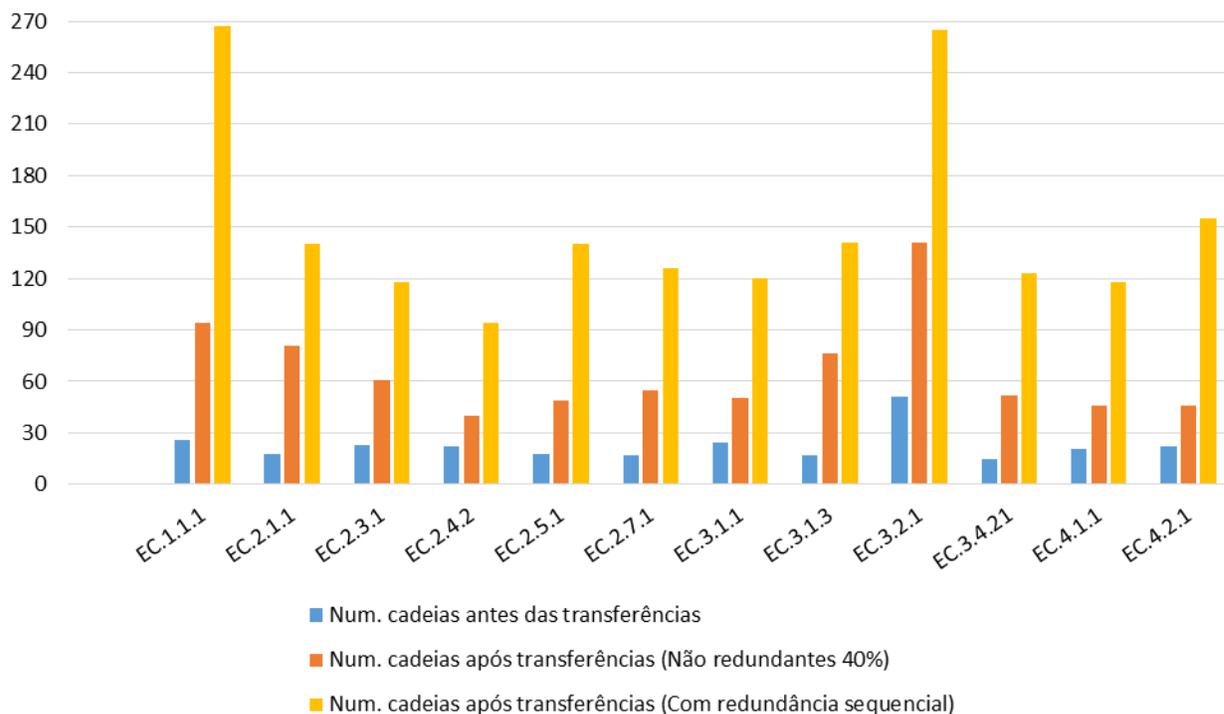


Figura 7-24: Comparação entre o número de cadeias enzimáticas presentes no banco de dados utilizado antes e após a transferência de anotação com e sem redundâncias sequenciais. O incremento de cadeias não é homogêneo, porém foi possível dobrar o número de cadeia em todos os casos.

Devido à distribuição das cadeias não ser homogênea para as diferentes classes EC, nota-se uma diferença entre os números de novas cadeias adicionadas em cada sub-subclasse. Algumas conformações tridimensionais são mais abundantes do que outras. Assim, após a transferência de anotações, algumas conformações continuam sub-representadas. Por serem responsáveis pela catálise de reações idênticas, enzimas que compartilham os mesmo três primeiros dígitos EC, ou seja, mesma sub-subclasse, tendem a apresentar grandes similaridades estruturais, uma vez que suas funções estão intimamente ligadas a suas formas. Ainda que enzimas com estruturas muito diferentes possam catalisar a mesma reação, tem-se este fato como uma exceção (Chothia & Lesk, 1986). Portanto, a redundância estrutural dos conjuntos de enzimas de uma mesma sub-subclasse é natural e pelos propósitos deste trabalho é justamente nesta redundância que regras podem ser construídas para a caracterização de resíduos de aminoácidos catalíticos. Caso contrário, uma diversidade estrutural levaria a uma diversidade de nanoambiente, dificultando a caracterização através de regras que justamente exploram as semelhanças entres esses nanoambientes estruturais. Como exemplo dessa similaridade estrutural e consequentemente físico-química dos nanoambientes dos resíduos de aminoácidos catalíticos, pode-se verificar o caso de duas

Serino Proteases que evoluíram convergentemente. As enzimas Subtilisinas (EC. 3.4.21.62) compreendem um caso de evolução convergente com outras Serino Proteases, como as Elastases (EC. 3.4.21.37). Assim, mesmo possuindo pouca similaridade sequencial (< 5%), a enzimas Subtilisina de *Bacillus licheniformis* (PDB: 1SBC cadeia A) e Elastase de *Homo sapiens* (PDB: 1PPF cadeia E) catalisam a mesma reação química e, possuem a mesma tríade catalítica de outras Serino Proteases (HIS, ASP e SER). O alinhamento estrutural entre essas duas estruturas revela pouca similaridade entre as estruturas como pode ser visto na Figura 7-25.

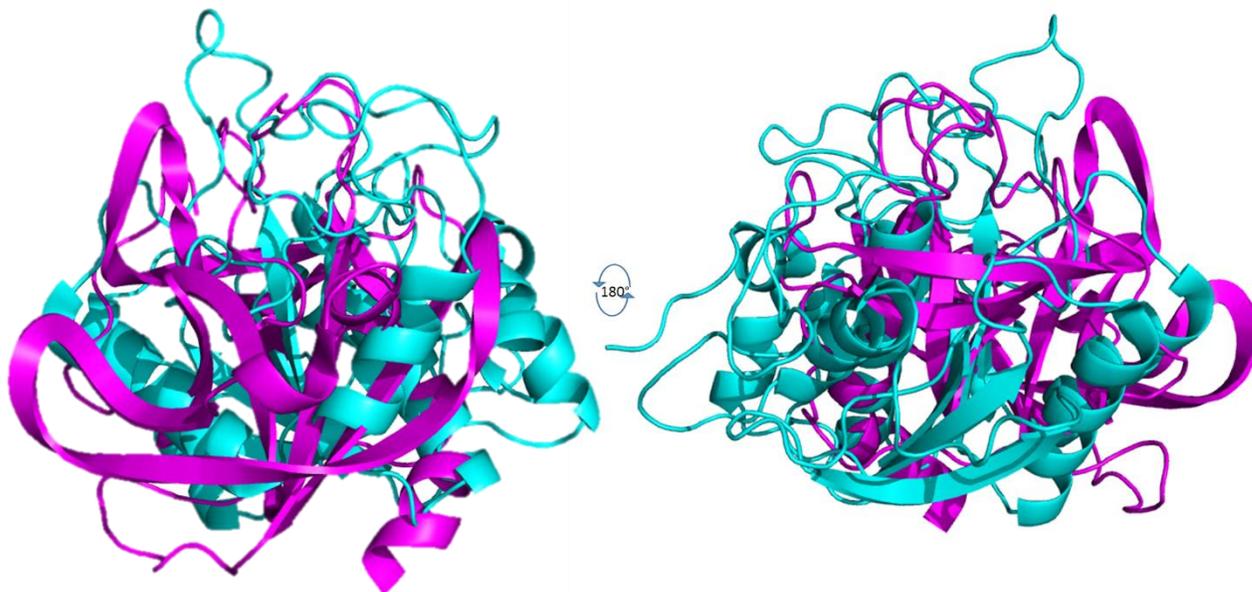


Figura 7-25: Alinhamento estrutural entre Subtilisina de *Bacillus licheniformis* (PDB: 1SBC, magenta) e Elastase de *Homo sapiens* (PDB: 1PPF, ciano), demonstra que as estruturas são bastante diferentes, porém catalisam a mesma reação química.

No entanto, ainda assim, é possível encontrar uma mesma regra que selecione os resíduos de aminoácido catalíticos de ambas as estruturas (Figura 7-26). Esse exemplo, fortalece a hipótese de um nanoambiente comum entre as duas enzimas, de forma a desempenharem o papel de catalisar a mesma reação química, ainda que bastante dissimilares em sequência e estrutura.

Alguns resíduos de aminoácidos catalíticos foram transferidos em maior número do que outros, uma vez que os alinhamentos estruturais forneceram um pareamento maior para estruturas contendo esses resíduos, em detrimento de outras. Ainda que alguns CSR's identificados no CSA possam ser considerados específicos, sendo encontrados somente em algumas enzimas de certos organismos ou famílias enzimáticas, pouco se pode concluir a respeito desses resíduos de aminoácidos, uma vez que não possuem o mesmo papel em todas as enzimas de uma mesma família ou sub-subclasse EC. Embora a maior ocorrência de certos tipos de aminoácidos catalíticos para uma mesma sub-subclasse EC possa

sugerir que esses são mais gerais em detrimento de outros encontrados somente em uma pequena parcela das enzimas de mesma sub-subclasse, devido à maior amostragem de certas conformações tridimensionais, alguns tipos de aminoácidos catalíticos puderam ser transferidos em maior número do que outros, considerando baixa similaridade sequencial (menor que 40%).



Figura 7-26: Regra selecionando CSR's de enzimas evolutivamente convergentes: Subtilisina de *Bacillus licheniformis* (PDB: 1SBC) e Elastase de *Homo sapiens* (PDB: 1PPF cadeia E). Apesar de grandes diferenças sequenciais e estruturais, as propriedades do nanoambiente dos CSR's são preservadas e ambas estruturas, sendo possível que a mesma regra (quadro em cinza) selecione os CSR's de ambas as enzimas (destaque em vermelho).

Como exemplo, tem-se o caso das Clp Proteases (EC 3.4.21.92), onde no CSA encontra-se somente anotação para uma estrutura de código PDB: 1TYF. Sendo seus resíduos de aminoácidos catalíticos compostos pelos resíduos: SER97, HIS122 e ASP171, componentes da tríade catalítica presente na grande maioria das Serino Proteases (EC. 3.4.21); o resíduo de aminoácido GLY68 (comumente encontrado em outras Serino Proteases); e MET98 (anotado como catalítico somente nas Clp Proteases). Durante a transferência de anotações, foram alinhadas estruturalmente com a enzima 1TYF seis outras

Clp Proteases, sendo cinco dessas com similaridade sequencial acima de 40% com a enzima de referência (PDB: 1TYF) e apenas uma com similaridade abaixo de 40% (PDB: 2C8T). Assim, foi introduzido ao conjunto das Serino Proteases (EC3.4.21) apenas uma Clp Protease, totalizando duas Clp Proteases em todo o conjunto de dados. O número de metioninas catalíticas, portanto, foi inferior ao número de outros tipos de aminoácidos (HIS, ASP e SER) para a sub-subclasse das Serino Proteases, sugerindo a especificidade deste tipo de aminoácido para as Clp Proteases, uma vez que o mesmo não é encontrado em outras Serino Proteases. Para identificar os CSR's da Clp Protease de código PDB 1TYF o CSA utilizou o trabalho de WANG et. al (1997). No artigo os autores determinaram a estrutura da Clp Protease de *Escherichia coli* complexada com di-isopropil-fluorofosfato (DFP) a uma resolução de 2.5Å. Eles relataram que o resíduo de aminoácido MET98 participa na estabilização do oxiânion intermediário (fosfato), através de uma ligação de hidrogênio com o oxigênio do oxiânion. Esta estabilização permite uma reorientação do oxigênio do SER97 formando uma ligação covalente com o grupo fosfato do substrato. Com base nesta evidência da ação da MET98, esse resíduo de aminoácido foi identificado como catalítico pela curadoria do CSA e por isso é encontrado em sua base de dados. Por outro lado, o resíduo de aminoácido GLY68 também foi identificado pelos mesmos autores como participante desta reação de estabilização, da mesma forma que em outras Serino Proteases, onde uma GLY desempenha o papel de estabilização do oxiânion (Topf, et al., 2002; Frigerio, et al., 1992; James, et al., 1980). De acordo com a definição de resíduos catalíticos utilizada pela CSA, resíduos que participam da estabilização de um estado transitório ou intermediário, são considerados como catalíticos. Por esta razão os resíduos de MET, GLY e ASN (no caso das Subtilisinas (Bryan, et al., 1986)) foram anotados como catalíticos pelo CSA.

Apesar de não se ter investigado se o mesmo ocorre para outras sub-subclasses, dada a diversidade de CSR's encontrada dentro de cada sub-subclasse EC, acredita-se que o mesmo efeito possa ocorrer também para outras sub-subclasses. Ou seja, espera-se que existam anotações de resíduos de aminoácidos catalíticos que introduzam uma variabilidade nos grupos de resíduos catalíticos de cada enzima de uma mesma sub-subclasse, dificultando assim a tarefa de se obter regras que classifiquem os CSR's de cada sub-subclasse EC.

Apesar do CSA não conter anotações de todos os CSR's para as enzimas presentes em sua base de dados, alguns aminoácidos podem ser considerados específicos para um grupo de enzimas compartilhando os mesmos quatro dígitos EC, em comparação a outros CSR's encontrados em grande parte das enzimas, compartilhando os mesmo três primeiros dígitos EC. Sabendo-se que o quarto nível da hierarquia EC diferencia as enzimas segundo o substrato de ligação e, dada a diversidade dos tipos de aminoácidos identificados como catalíticos em casa sub-subclasse, podemos afirmar que certos resíduos

de aminoácido desempenham o mesmo papel em diferentes enzimas, porém, são de tipos de aminoácidos diferentes. Assim, esses resíduos de aminoácido considerados como catalíticos para enzimas compartilhando os mesmo quatro dígitos EC, podem ser considerados variáveis dentro de cada sub-subclasse, ao passo que outros são mais gerais e podem ser utilizados para caracterizar uma sub-subclasse EC, como a tríade catalítica presente na maioria das Serino Proteases.

As propensidades catalíticas dos resíduos de aminoácidos foram calculadas antes e após a transferência de anotações e, devido aos fatos apresentados acima, houve modificações nesses valores. Devido à baixa amostragem das estruturas ou à especificidade de alguns resíduos de aminoácidos catalíticos, aminoácidos que anteriormente possuíam propensidades altas (maiores que 1) tiveram suas propensidades reduzidas. Em muitos casos, houve inversão da propensidade, ou seja, resíduos considerados propensos a serem catalíticos passaram a ter valores negativos, ressaltando a propensão de outros tipos de aminoácidos que podem ser considerados gerais dentro de uma sub-subclasse EC, como ilustram os gráficos da Figura 7-27. A expansão dos dados levando a alterações observadas nas propensidades dos CSR's reforça a hipótese de que alguns resíduos de aminoácidos catalíticos podem ser considerados mais gerais, em detrimento de outros mais específicos de cada tipo de enzima. Uma vez que a expansão dos dados introduz novos CSR's aos conjuntos de dados, estes novos resíduos elevam a representatividade de alguns tipos de aminoácidos elevando assim suas propensidades. O contrário ocorre para aqueles tipos de aminoácidos que não tiveram suas quantidades alteradas e por consequência tiveram suas propensidades reduzidas.

Com a expansão dos dados, novos classificadores foram treinados e seus desempenhos comparados com os classificadores anteriormente obtidos. Dado que os melhores desempenhos foram obtidos com o uso de redução de dimensionalidade e sobreamostragem da classe positiva, somente classificadores treinados sob estas configurações foram considerados para avaliar o impacto da transferência de anotação. Obedecendo as mesmas configurações de sobreamostragem anteriores e utilizando a seleção sequencial incremental de atributos, avaliaram-se os desempenhos dos novos classificadores através de validação cruzada 2x5. No gráfico da Figura 7-28, encontram-se os desempenhos dos classificadores que retornaram maiores F-measure's, bem como o número de cadeias antes e após a transferência de anotações (com e sem redundância sequencial). Pelo gráfico, nota-se que o algoritmo SMOTEBagging foi o que apresentou maiores valores de F-measure em quase todas as sub-subclasses (8 de 12). Ainda que uma comparação desses desempenhos com os desempenhos dos classificadores treinados anteriormente não possa ser realizada diretamente, devido aos conjuntos de teste serem diferentes, nota-se um aumento no desempenho dos classificadores (com exceção dos classificadores para as Pentosiltransferases (EC.2.4.2)). No caso das Pentosiltransferases, observa-se que, mesmo após a

transferência de anotações, poucas cadeias não redundantes (18) foram adicionadas a esta sub-subclasse. Logo, a transferência teve pouco impacto nos classificadores treinados para esta sub-subclasse.

Para avaliar se houve ganhos nos classificadores após a transferência de anotação, em relação aos outros classificadores, um conjunto de teste (*benchmark*) foi criado para estimar o desempenho dos melhores classificadores antes e após a transferência. Construído utilizando-se cadeias retiradas do conjunto de enzimas anotadas pela transferência, de forma que cada enzima do conjunto original contribuísse com uma enzima alinhada estruturalmente, o conjunto de teste possui cadeias cobrindo todas as diferentes conformações tridimensionais encontradas no conjunto original. Retiradas dos conjuntos de treinamentos, as cadeias do conjunto de teste foram utilizadas para estimar o desempenho dos classificadores, que foram novamente treinados.

Um resumo dos resultados encontra-se nos gráficos da Figura 7-29, onde percebe-se que em muitos casos os classificadores treinados com os dados após a transferência de anotação (gráfico inferior da Figura 7-29) trouxeram aumento considerável no desempenho dos classificadores, quando comparados aos classificadores treinados com os dados originais (sem transferência) (gráfico superior da Figura 7-29). Somente para a sub-subclasse das Pentosiltransferases (EC.2.4.2) a introdução de mais cadeias não trouxe ganho de desempenho. Novamente, isto deve-se em parte ao baixo número de cadeias adicionadas com a transferência de anotações para esta sub-subclasse. De um conjunto com 22 cadeias, houve uma expansão para 40 cadeias, ou seja, um aumento de 18 cadeias, algo inexpressivo comparado às outras sub-subclasses, onde o menor aumento foi de 24 cadeias no caso das Hidro-Liases (EC.4.2.1), que também mostrou ganhos mais tímidos com a transferência de anotação.

Apesar de apresentarem melhoras no desempenho, classificadores treinados utilizando-se técnicas de *boosting* e *bagging* dificultam a interpretação dos conjuntos de regras gerados, uma vez que as predições são realizadas segundo composições de predições de diferentes regras e, possivelmente, divergentes. Em um *ensemble* cada classificador (conjunto de regras) irá fornecer uma classificação para o resíduo de aminoácido (catalítico ou não catalítico). O resíduo será então classificado como catalítico se a soma dos pesos associados aos classificadores membros do *ensemble* que forneceram uma classificação deste como catalítico for superior a soma dos pesos dos que forneceram uma classificação de não catalítico. Dessa forma, um mesmo resíduo será classificado como catalítico por alguns membros do *ensemble* segundo suas regras e não catalíticos por outros, o que dificulta uma conclusão a respeito de qual regra adotar para caracterizar o resíduo de aminoácido. Por isso, é preferível classificadores gerados sem uso de técnicas de *ensemble*, mesmo com o emprego de algum pré-processamento.

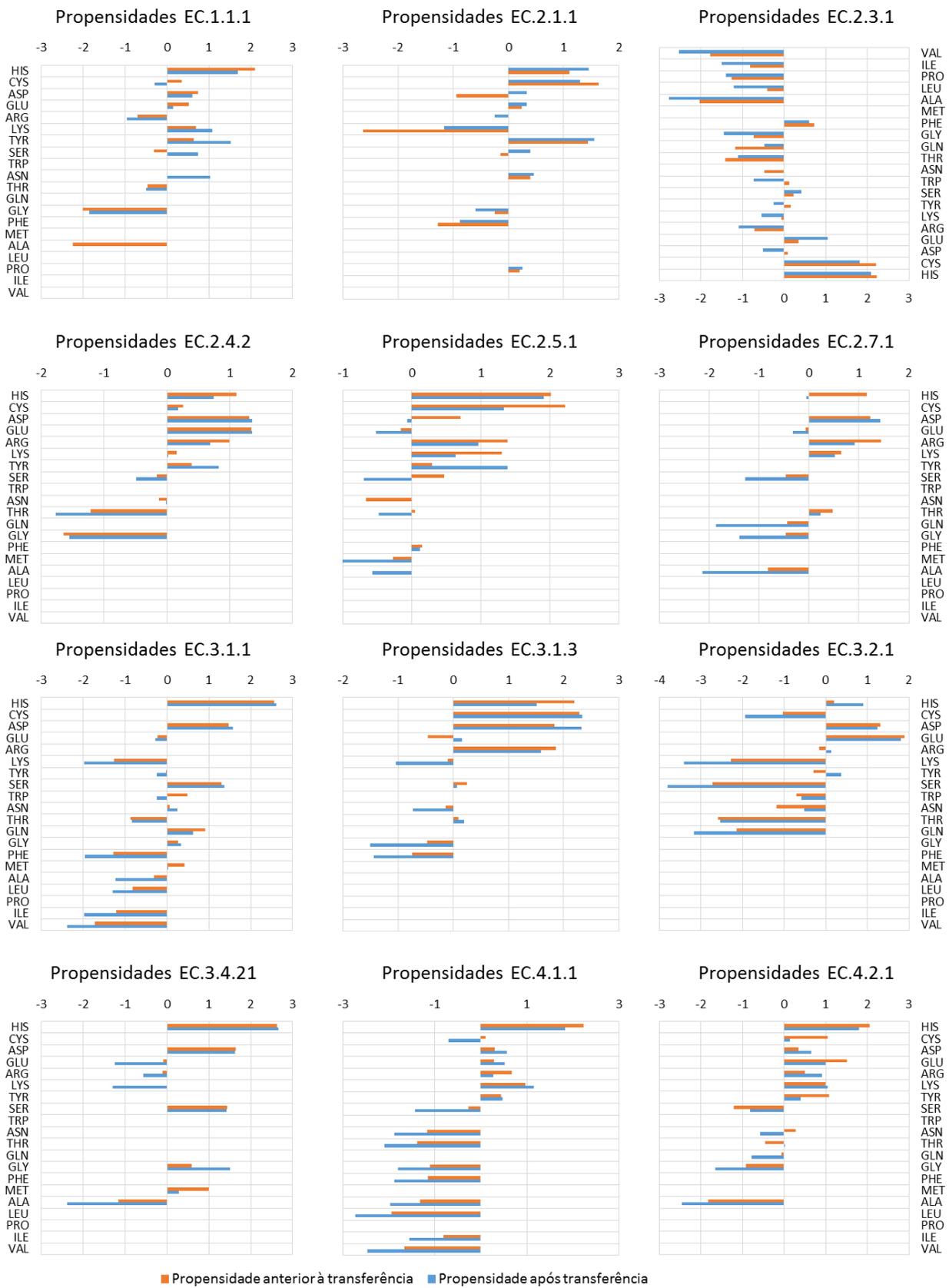


Figura 7-27: Propensidades dos 20 tipos de aminoácidos em serem catalíticos para as 12 sub-subclasses consideradas antes e após a transferência de anotação.

Número de cadeias por sub-subclasse EC antes e após as transferências de anotações por alinhamentos estruturais

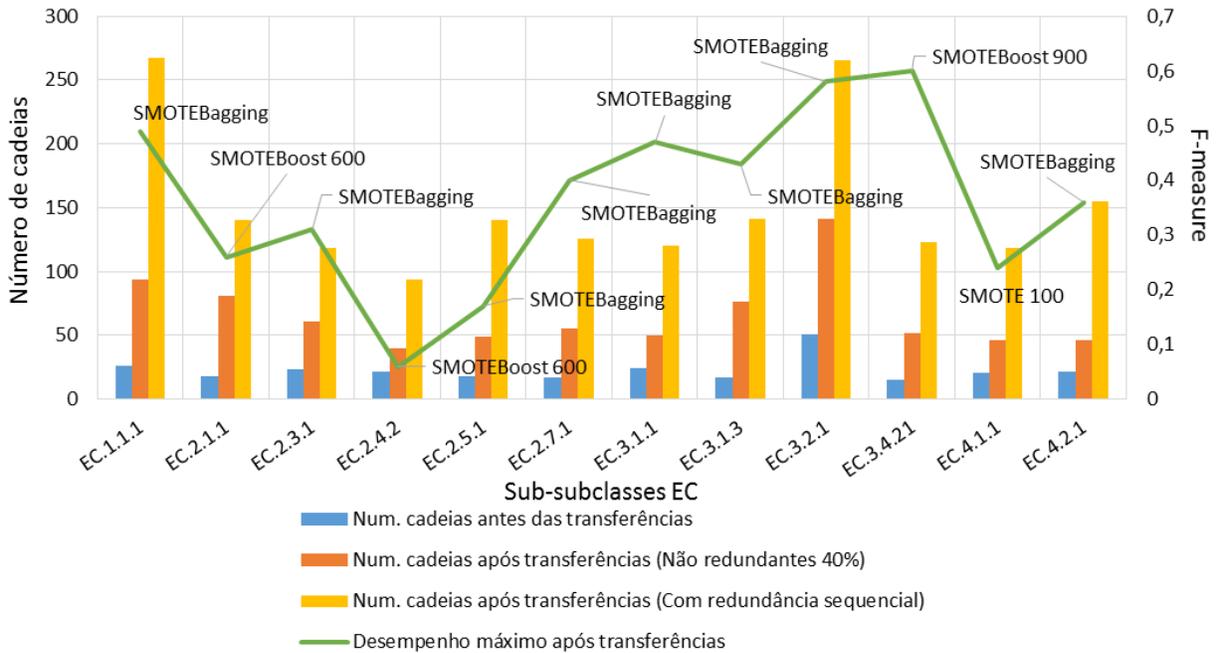


Figura 7-28: Desempenho dos melhores classificadores treinados após a transferência de anotação e o número de cadeias em cada sub-subclasse antes e após as transferências.

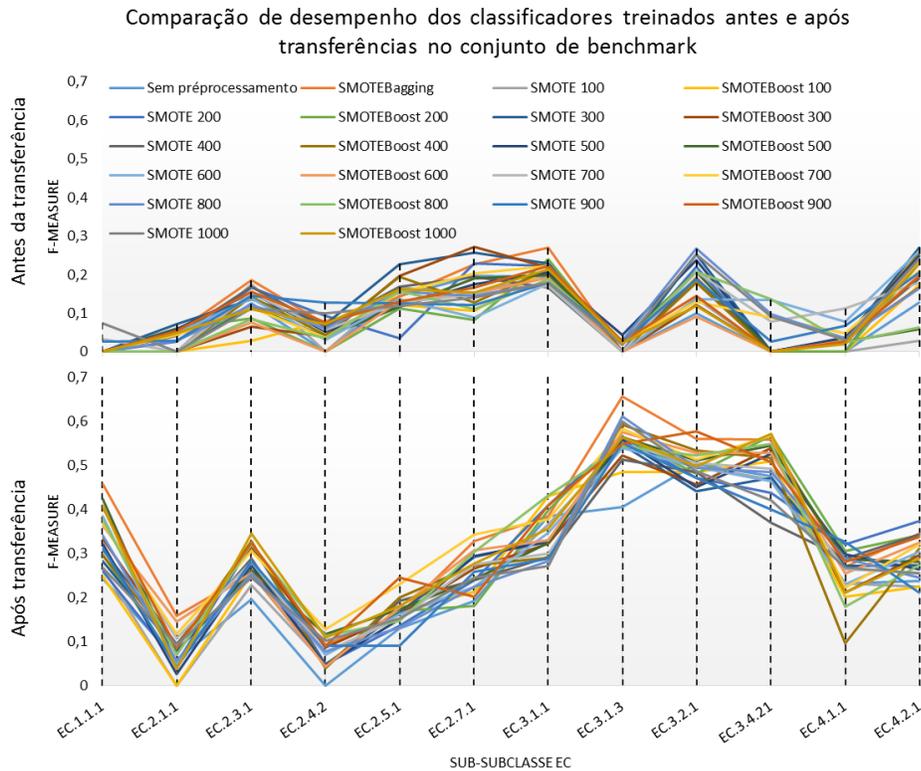


Figura 7-29: Comparação de desempenho dos classificadores treinados antes e após a transferência de anotações.

7.12 Classificador genérico e comparação com outros métodos na literatura

Como discutido no capítulo Capítulo 2, os métodos propostos para a predição de resíduos de aminoácidos catalíticos não realizam separação das enzimas segundo as sub-subclasses a que essas pertencem, sendo utilizado um único classificador para predição, ou seja, são considerados classificadores genéricos. Devido à impossibilidade de construir classificadores para todas as sub-subclasses, seja por falta de dados, inconsistências ou múltiplas atribuições e falta de classificação EC para muitas enzimas, tem-se explorado na literatura o caso de predição sem considerar separação por números EC. Apesar de objetivos pouco semelhantes nos dois casos, a predição dos CSR's pode ser realizada. No entanto, ressalta-se que, no caso de separação por sub-subclasse EC, procurou-se buscar características estruturais comuns aos resíduos de aminoácidos catalíticos de enzimas responsáveis pela catálise de uma mesma reação química. Visto como um problema de caracterização tendo a predição como consequência, a separação por sub-subclasse EC difere do objetivo dos outros métodos propostos, onde a predição é tida como único objetivo, não sendo almejada a caracterização das enzimas segundo as reações químicas que estas catalisam. Por estas razões, não é possível realizar uma comparação entre os classificadores apresentados nas seções anteriores e os classificadores encontrados na literatura. Tendo em vista avaliar as contribuições dos descritores estruturais do Blue Star STING aplicados à predição de resíduos catalíticos e estabelecer uma comparação com outros métodos, foram treinadas diversas Máquinas de Vetores de Suporte (SVM) considerando os mesmos conjuntos de dados utilizados pelos demais métodos. Essa iniciativa se justifica devido à boa capacidade de SVM's em operar em espaços de alta dimensão (Boser, et al., 1992; Aizerman, et al., 1964) e por serem amplamente empregadas para a predição de resíduos de aminoácidos catalíticos (Petrova & Wu, 2006; Pugalenthia, et al., 2008; Xin, et al., 2010).

Foram comparados com os classificadores treinados neste trabalho três métodos recentes e de maior sucesso que utilizam informações estruturais para predição, além de um método também de sucesso que utiliza somente informações provenientes da estrutura primária das enzimas. Os métodos *Partial Order Optimum Likelihood* (POOL), *Enzyme Catalytic residue Side-chain Arrangement* (EXIA) e o método proposto por CILIA & PASSERINI (2010) (C&P) utilizam informações extraídas das estruturas terciárias das enzimas e possuem versões com e sem uso de descritores de conservação de estrutura primária. O método POOL foi avaliado pelos seus autores utilizando-se informações estruturais extraídas de curvas de titulação teóricas microscópicas (THEMATICS) (Wei, et al., 2007), descritores de cavidades na superfície na molécula e descritores de conservação, respectivamente rotulados de POOL(T), POOL(G) e POOL(C). Foram consideradas para comparação as versões com e sem conservação do método *Enzyme*

Catalytic residue Side-chain Arrangement, EXIA+PSSM e EXIA, respectivamente, bem como versões do método POOL com e sem conservação, POOL(T+G) e POOL(T+G+C) respectivamente, além da versão do método C&P que faz uso de descritores de conservação. Da mesma forma, foram consideradas duas versões do classificador proposto, sendo uma sem uso de conservação (STING-CSR) e outra com uso de descritores de conservação (STING-CSR-Conserv.).

A comparação com o método POOL foi realizada utilizando-se um conjunto de dados contendo 160 enzimas retiradas do CSA e fornecido pelos autores do método (POOL160), empregando-se validação cruzada com 10 pastas e 10 repetições (10x10 treinamentos). As curvas ROC e PR para os classificadores STING-CSR e STING-CSR-Conserv em todos os conjuntos utilizados para comparação encontram-se na Figura 7-30.

Ambos os classificadores, STING-CSR e STING-CSR-Conserv, mostraram desempenho superior ao método POOL para o conjunto de dados POOL160. O classificador STING-CSR foi superior em comparação ao POOL(T+G) tanto em sensibilidade (64.61% para igual precisão) quanto em precisão (21.33% para igual sensibilidade), apresentando AUCROC também superior, como mostra o gráfico da Figura 7-31. Em comparação ao POOL(T+G+C), o STING-CSR atingiu desempenhos equivalentes mesmo sem uso de descritores de conservação, sendo poucas as diferenças na sensibilidade, precisão e AUCROC entre os classificadores.

A introdução de conservação ao método POOL traz ganhos, mas estes são menores que os ganhos obtidos considerando-se o uso de conservação juntamente com descritores do Blue Star STING. Para uma precisão igual à obtida pelo POOL(T+G+C) (19.07%) o método STING-CSR-Conserv atingiu sensibilidade de 72.38% em comparação aos 64.68%, e para igual sensibilidade (64.68%) o STING-CSR-Conserv mostrou precisão de 28.33% contra os 19.07% do POOL(T+G+C), além de um valor de AUCROC superior (0.963 contra 0.925), como pode ser visto também no gráfico da Figura 7-31. O desempenho médio do STING-CSR, STING-CSR-Conserv. e as duas versões do método POOL encontram-se na Tabela 7-11.

A comparação com os métodos EXIA, EXIA+PSSM e C&P foi realizada utilizando-se cinco diferentes conjuntos de dados, extraídos de outros trabalhos da literatura. Os conjuntos de dados incluem o PW79, com 79 enzimas retirados do trabalho de PETROVA & WU (2006), EFold, EFamily e EFSuperfamily, contendo diferentes níveis de homologia, e o POOL160. Os conjuntos de dados EF foram criados de acordo com os níveis de *fold*, superfamília, e família do SCOP ASTRAL 40 v1.65 (Youn, et al., 2007). O método EXIA foi ainda comparado utilizando o conjunto UB78 (Chien & Huang, 2012), contendo 78 enzimas cristalizadas sem o substrato (*unbound*) selecionadas dos demais conjuntos. Devido a possíveis alterações nas conformações das enzimas após estabelecerem ligações com o substrato, ligado

ao mecanismo de ajuste induzido (*induced-fit*), este conjunto foi utilizado para avaliar se os padrões explorados pelos descritores estruturais sofrem alterações após a ligação com o substrato, ao ponto de favorecer as predições.

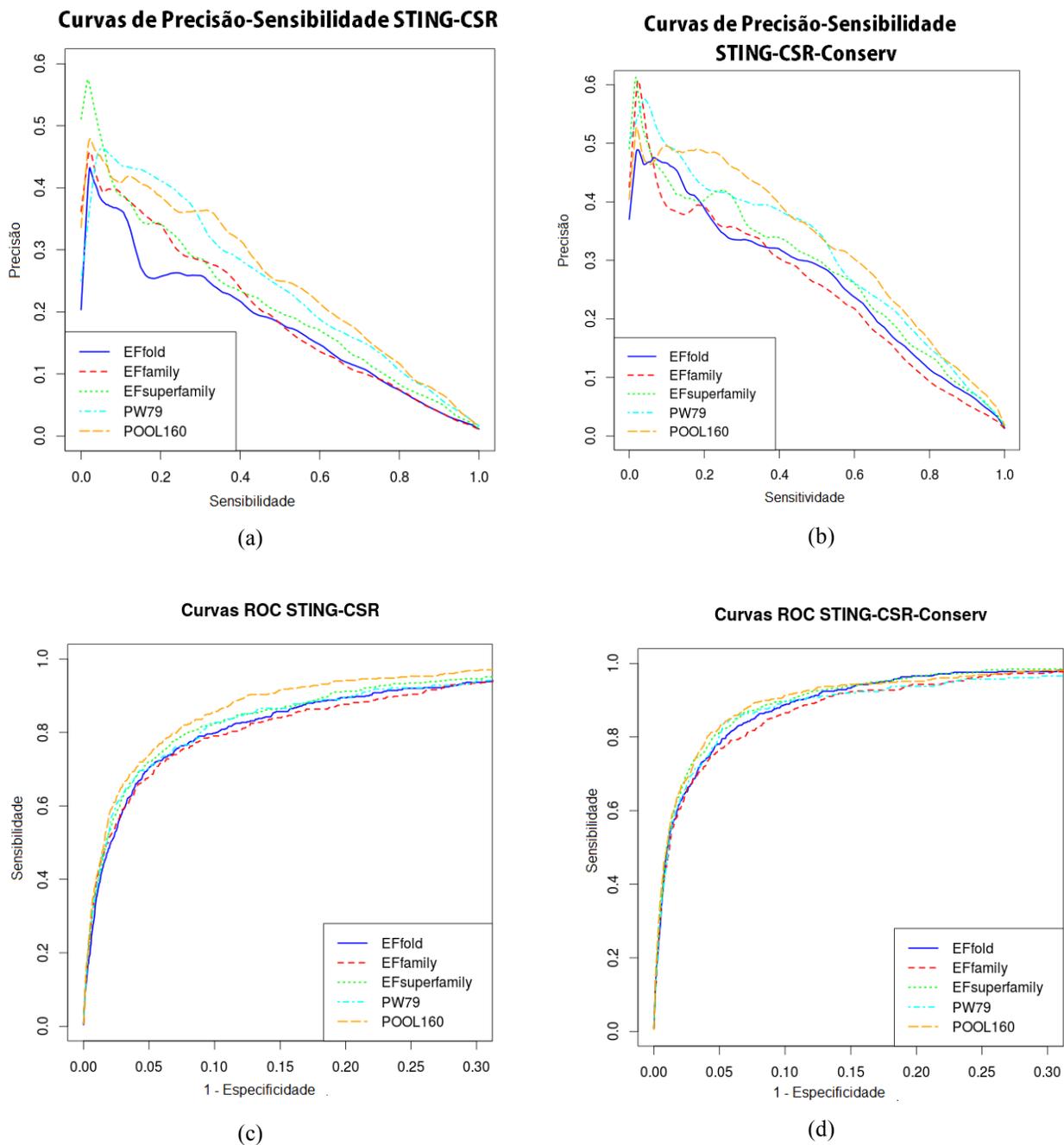


Figura 7-30: Curvas PR (a) e (b) e ROC (c) e (d) para os classificadores STING-CSR e STING-CSR-Conserv nos conjuntos de dados utilizados para comparação.

Para os cinco conjuntos de dados, STING-CSR e STING-CSR-Conserv mostraram desempenhos muito próximos aos obtidos pelos métodos EXIA e EXIA+PSSM, sendo que estes últimos mostraram desempenhos ligeiramente superiores. Sem uso de conservação, o STING-CSR mostrou taxas de

sensibilidade e precisão superiores ao EXIA para os conjuntos EFfold, EFsuperfamily e EFamily, apesar de ter um desempenho geral (AUCROC) levemente inferior ou similar. As previsões sofrem maiores divergências para o conjunto PW79, onde STING-CSR obteve AUCROC de 0.926 enquanto o EXIA mostrou um valor de 0.962, assim como sensibilidade e precisão também superiores. Para o conjunto POOL160, os desempenhos de ambos os métodos foram muito parecidos, sendo AUCROC de 0.96 para EXIA e 0.95 para STING-CSR com taxas de sensibilidade e precisão semelhantes (Tabela 7-12). Um fator importante para que o método EXIA obtenha tal desempenho deve-se ao descritor de orientação das cadeias laterais dos resíduos de aminoácido, descrito na seção 2.2. No entanto, o método utiliza outros descritores incluindo um descritor de propensidade que avalia a probabilidade de um resíduo de aminoácido e seus vizinhos de formarem um sítio catalítico baseado em análises estatísticas. Por outro lado, o método STING-CSR utiliza somente descritores físico-químicos e estruturais extraídos diretamente da proteína para qual seus resíduos de aminoácido serão classificados, evitando assim a necessidade de descritores estatísticos como propensidade. Da mesma forma como descritores de conservação da estrutura primária de proteínas, descritores de propensidade não fornecem informações sobre as características físico-químicas e estruturais para um resíduo ser considerado como catalítico, além de sofrerem alterações à medida que as bases de dados crescem, uma vez que, as probabilidades e entropias tendem a se modificar com a inserção de novas sequências e novos experimentos de mutagêneses.

Tabela 7-11: Comparação entre STING-CSR e STING-CSR-Conserv. com os métodos POOL(T+G) e POOL(T+G+C) no conjunto de dados POOL160. ¹Sensibilidade para igual precisão; ²Precisão para igual sensibilidade.

	POOL(T+G)	POOL(T+G+C)
<i>SENSIBILIDADE</i>	61.74	64.68
<i>PRECISÃO</i>	18.06	19.07
<i>AUCROC</i>	0.907	0.925
	STING-CSR	
<i>SENSIBILIDADE</i> ¹	64.61	64.32
<i>PRECISÃO</i> ²	21.33	19.61
<i>AUCROC</i>	0.95	
	STING-CSR-Conserv	
<i>SENSIBILIDADE</i> ¹	73.84	72.38
<i>PRECISÃO</i> ²	30.16	28.33
<i>AUCROC</i>	0.963	

Para os casos em que se utilizaram descritores de conservação, o EXIA+PSSM mostrou

desempenhos superiores ao STING-CSR-Conserv segundo AUCROC, sendo que as taxas de sensibilidade e precisão foram superiores em todos os conjuntos, exceto para o EFsuperfamily (Tabela 7-12). A introdução de descritores de conservação nitidamente eleva os desempenhos de ambos os métodos em todos os casos avaliados, sendo a sensibilidade a que sofre maior aumento, como foi visto na seção 7.9. O aumento da sensibilidade sugere que descritores de conservação são importantes para propiciar maior cobertura dos resíduos de aminoácidos catalíticos, sem acarretar uma grande introdução de falsos positivos, causando um aumento ou ligeira redução da precisão. Como pode ser percebido através das curvas ROC e PR dos gráficos da Figura 7-30, descritores de conservação elevam consideravelmente o desempenho dos classificadores.

O método C&P possui um desempenho geral (AUCROC) melhor em relação ao STING-CSR para o conjunto PW79 e similar para o POOL160 (Tabela 7-13). Para os conjuntos EF, o STING-CSR apresentou desempenhos inferiores segundo sensibilidade e precisão. Porém, em comparação com o STING-CSR-Conserv, seu desempenho foi inferior para os conjuntos PW79, EFsuperfamily e EFamily, se comparadas as taxas de sensibilidade e precisão. Nos outros dois conjuntos, Efold e POOL160, o STING-CSR-Conserv mostrou melhor precisão no primeiro caso (22.59% contra 17.1% para igual sensibilidade de 64.2%) porém menor sensibilidade (60.47 % contra 64.20% para igual precisão de 17.1%). Para o conjunto POOL160, apesar de o STING-CSR-Conserv ter obtido sensibilidade e precisão inferiores, o desempenho geral do STING-CSR-Conserv foi maior (0.963 AUCROC contra 0.948 AUCROC).

Para o conjunto UB78 com enzimas cristalizadas sem o substrato, os classificadores STING-CSR e STING-CSR-Conserv mostraram valores de AUCROC de 0.911 e 0.95, respectivamente, valores estes equivalentes aos apresentados para os outros conjuntos. Os classificadores EXIA e EXIA+PSSM apresentaram AUCROC mais elevados do que o STING-CSR e STING-CSR-Conserv no UB78 (0.941 e 0.961, respectivamente), seguindo os resultados das comparações em outros conjuntos de dados. Na Figura 7-32, são apresentadas as curvas ROC e PR dos classificadores STING-CSR e STING-CSR-Conserv para o conjunto UB78.

Os resultados das predições obtidas, tanto pelo STING-CSR quanto STING-CSR-Conserv, estão de acordo com os encontrados na literatura e ilustram as capacidades e limitações dos métodos atuais, bem como dos descritores físico-químicos e estruturais, em promoverem a classificação de resíduos de aminoácidos catalíticos. O uso de descritores de conservação eleva o desempenho das classificações em todos os métodos apresentados acima, indicando que estes descritores possuem informações relevantes para a predição que não podem ser obtidas a partir dos descritores físico-químicos e estruturais, diferentemente do que foi observado no caso da classificação de resíduos formadores de interface (de

Moraes, et al., 2014).

Tabela 7-12: Comparação entre STING-CSR e STING-CSR-Conserv com os métodos EXIA e EXIA+PSSM em vários conjuntos de dados. ¹Sensibilidade para igual precisão; ²Precisão para igual sensibilidade.

	PW79	EF FOLD	EF SUPERFAMILY	EF FAMILY	POOL160
EXIA					
SENSIBILIDADE	48.90	44.80	50.00	46.30	68.60
PRECISÃO	28.00	17.10	16.90	18.50	19.00
AUCROC	0.962	0.940	0.940	0.944	0.960
STING-CSR					
SENSIBILIDADE ¹	37.34	47.69	55.60	48.48	63.10
PRECISÃO ²	26.59	19.95	20.43	20.52	18.17
AUCROC	0.926	0.930	0.940	0.925	0.950
EXIA+PSSM					
SENSIBILIDADE	63.0	72.3	72.4	69.0	78.0
PRECISÃO	28.0	17.1	16.9	18.5	18.9
AUCROC	0.978	0.968	0.965	0.966	0.969
STING-CSR-CONSERV					
SENSIBILIDADE ¹	54.0	60.5	73.6	62.8	72.5
PRECISÃO ²	26.0	16.6	18.6	17.0	18.1
AUCROC	0.953	0.959	0.962	0.952	0.963

Tabela 7-13: Comparação entre STING-CSR e STING-CSR-Conserv e o método C&P (Cilia & Passerini, 2010) em vários conjuntos de dados. ¹Sensibilidade para igual precisão; ²Precisão para igual sensibilidade.

	PW79	EF FOLD	EF SUPERFAMILY	EF FAMILY	POOL160
CILIA & PASSERINI					
SENSIBILIDADE	46	64.2	67.3	61.7	78.1
PRECISÃO	28	17.1	16.9	18.5	19.00
AUCROC	0.963	-	-	-	0.948
STING-CSR					
SENSIBILIDADE ¹	37.34	47.69	55.6	48.48	63
PRECISÃO ²	27.13	13.16	14.6	13.35	13.05
AUCROC	0.926	0.928	0.937	0.925	0.95
STING-CONSERV-CSR					
SENSIBILIDADE ¹	54	60.47	73.58	62.81	72.34
PRECISÃO ²	36.93	22.59	21.25	21.95	18.14
AUCROC	0.953	0.959	0.962	0.952	0.963

Comparação entre Curvas ROC

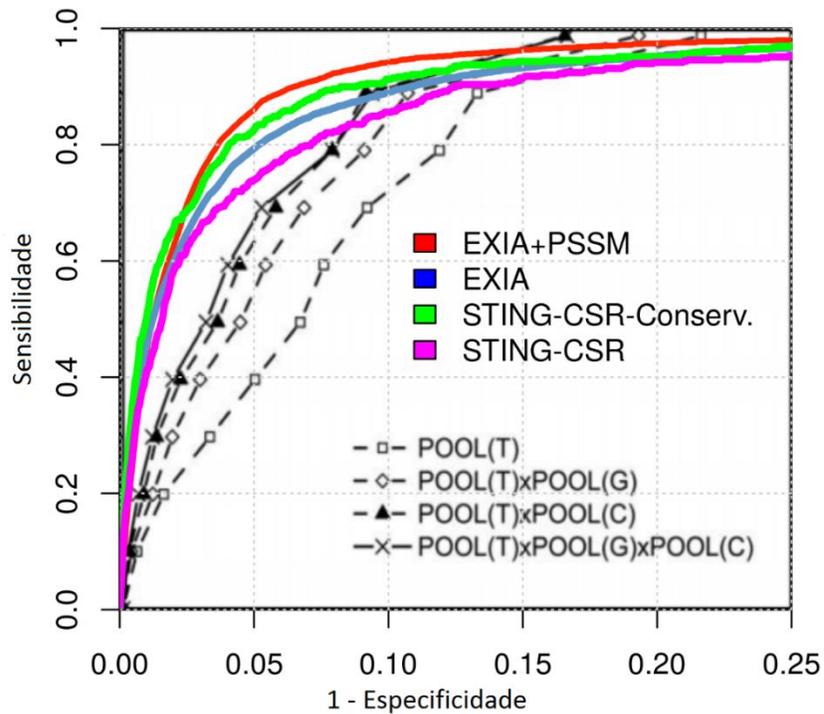


Figura 7-31: Curvas ROC dos métodos EXIA, EXIA+PSSM, POOL(T), POOL(T+G), POOL(T+C), POOL(T+G+C), STING-CSR e STING-CSR-Conserv. Figura retirada de (Chien & Huang, 2012) e editada para inserção das Curvas ROC para os classificadores STING-CSR e STING-CSR-Conserv.

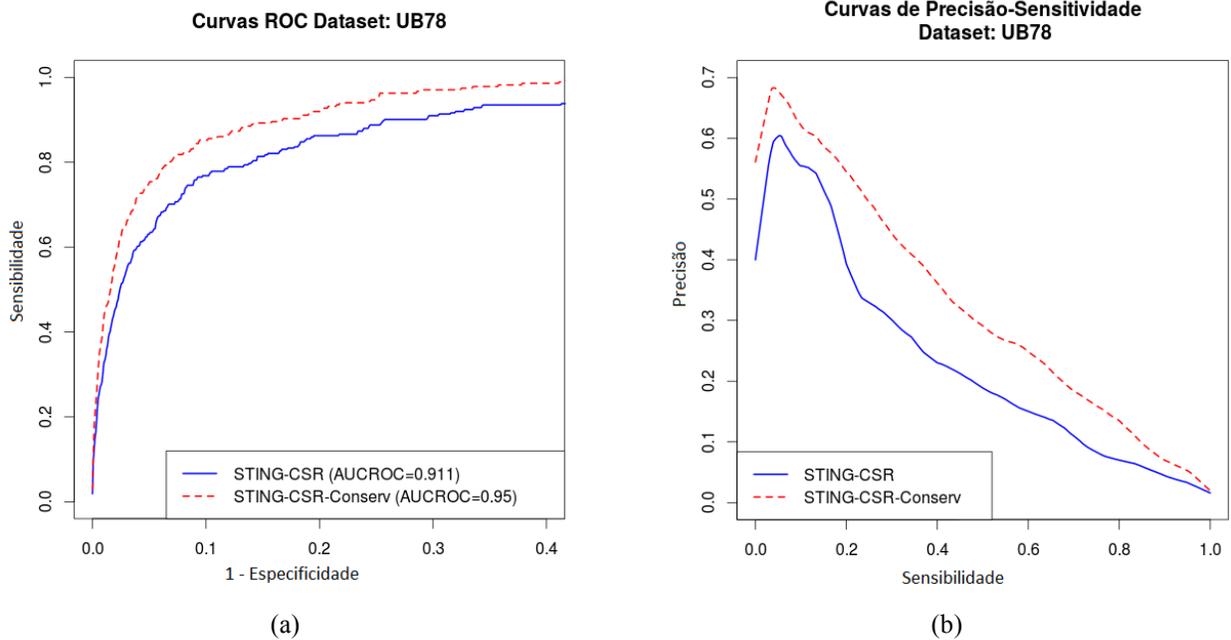


Figura 7-32: Curva ROC e PR para os classificadores STING-CSR e STING-CSR-Conserv aplicados ao conjunto de dados UB78, contendo enzimas sem o substrato (unbonded). Desempenhos são semelhantes aos apresentados em outros conjuntos de dados com enzimas cristalizadas com o substrato.

Ainda que, do ponto de vista prático, os resultados podem ser considerados baixos para a construção de uma ferramenta automatizada de predição de resíduos de aminoácidos catalíticos com altas taxas de acerto e precisão, é relevante considerar que, devido a diversos desafios inerentes ao problema, tais como o desbalanceamento entre classes, baixo número de amostras catalíticas e anotações incompletas no CSA, estes resultados dificilmente poderão ser elevados a patamares superiores sem um tratamento adequado destes desafios. Um classificador de resíduos de aminoácidos catalíticos com alta precisão pode não ser igualmente eficiente na prática, visto que suas predições irão excluir possíveis resíduos de aminoácidos catalíticos indevidamente anotados como não catalíticos, devido à falta de experimentos e estudos que comprovem a importância destes para a atividade enzimática. Por outro lado, classificadores com alta sensibilidade, mesmo que com baixa precisão, podem fornecer predições de resíduos de aminoácidos catalíticos ainda não anotados, cabendo portanto uma inspeção humana precisa das classificações.

Capítulo 8 - Conclusão

Como objeto de estudo deste trabalho, procurou-se caracterizar os resíduos de aminoácidos catalíticos divididos de acordo com a hierarquia *Enzyme Commission*. Devido às diferenças físicas, químicas e estruturais nas enzimas que compõem as diferentes sub-subclasses EC, procurou-se identificar essas diferenças e caracterizar aquelas que possuem uma mesma identificação e, conseqüentemente, catalisam a mesma reação química. Pelo fato de catalisarem a mesma reação para substratos diferentes, explorou-se a busca por padrões nos nanoambientes em torno dos resíduos de aminoácidos catalíticos, que pudessem ser úteis para caracterizarem e classificarem um conjunto de enzimas compartilhando os mesmos três primeiros dígitos da nomenclatura EC.

Considerado um problema de classificação com alto desbalanceamento, a predição de resíduos de aminoácidos catalíticos apresenta os mesmos desafios que outros problemas similares, como detecção de fraudes em ligações telefônicas (Fawcett & Provost, 1997), filtragem e recuperação de informação (Lewis & Catlett, 1994) e detecção de derramamentos de petróleo em imagens de satélite (Kubat, et al., 1998). Estes problemas de classificação oferecem dificuldades aos algoritmos de aprendizado de máquina que buscam minimização do erro ou maximização de acurácia. Por isso, faz-se necessário o emprego de algum pré-processamento para promover o balanceamento entre as classes dos problemas. Na área de predição de resíduos catalíticos, nenhuma atenção tem sido dada a este fato, sendo que os métodos que realizam algum pré-processamento dos dados utilizam somente da subamostragem aleatória. Neste trabalho, explorou-se o uso da subamostragem aleatória, assim como as sobreamostragens aleatórias com replicação e atribuição de pesos, e também a introdução de amostras sintéticas (SMOTE).

Foram consideradas também diversas propostas de classificadores, particularmente aquelas baseadas em regras e consideradas estados da arte em aprendizado de máquina, capazes de fornecerem regras facilmente interpretáveis por especialistas, como RIPPER e C4.5. Métodos para a construção de comitês de classificadores (*ensembles*), como *boosting* e *bagging*, adaptados a problemas desbalanceados, também foram avaliados, dentre eles RUSBoost integrando subamostragem com *boosting*, SMOTEBoost e SMOTEBagging, integrando SMOTE com *boosting* e *bagging*, respectivamente. Devido à dificuldade em lidar com um alto número de atributos e à perda da eficácia de métodos como SMOTE, uma redução de dimensionalidade foi realizada através de busca sequencial incremental (*forward feature selection*), num esquema de *wrapper* selecionando os melhores atributos.

Os classificadores inicialmente treinados apresentaram diversos problemas de generalização, devido à indução de regras específicas e com baixo suporte. Um estudo detalhado desses problemas revelou ser

a baixa representatividade das amostras de resíduos de aminoácidos catalíticos o maior desafio ao treinamento dos classificadores. A variabilidade das soluções encontradas pelos algoritmos durante repetidas execuções da validação cruzada, não somente indicam a existência de múltiplas regras capazes de selecionar os CSR's, com diferentes coberturas e precisões, mas também que para obter uma caracterização para todos os CSR's de uma mesma sub-subclasse, com alta cobertura e precisão, é preciso um número relativamente elevado de regras (maior que 10). Esse número de regras sugere que há segregações mesmo dentro da classe de resíduos de aminoácidos catalíticos, como foi de fato observado no caso das Clp Proteases, onde adicionalmente à tríade catalítica das Serino Proteases, foi anotado também o resíduo MET98 como membro do seu grupo de CSR's.

Para aumentar a representatividade da classe de resíduos de aminoácido catalíticos, uma transferência de anotação via alinhamentos estruturais foi realizada. Os classificadores, treinados e avaliados em um benchmark preparado para comparação, revelaram um aumento significativo no desempenho, com a construção de conjuntos de regras de maior suporte e precisão.

Neste caso, o desbalanceamento entre classes não é o maior responsável pelo baixo desempenho dos classificadores, mas sim a sub-representação da classe positiva. Caso um maior número de amostras positivas seja introduzido nesses conjuntos, mesmo que com um aumento do número de negativas, mantendo a mesma proporção entre as classes, observa-se um aumento no desempenho dos classificadores. Isto devido ao fato de que quando o número de amostras é muito baixo, as regras tendem a ser muito específicas, cobrindo poucas amostras catalíticas, e os algoritmos ficam sujeitos a soluções ótimas locais. Com um baixo número de amostras catalíticas, os algoritmos são obrigados a gerar regras que cubram somente uma pequena porção destas amostras, uma vez que uma regra mais geral irá cobrir também um maior número de amostras negativas e comprometer sua precisão. O aumento do número de amostras positivas produz regras com maior suporte e, conseqüentemente, mais gerais e que possam ser utilizadas para classificar resíduos de aminoácidos em outras enzimas com maior desempenho.

Dentre os melhores classificadores, encontram-se aqueles treinados utilizando-se técnicas de *boosting* e *bagging*, sendo o SMOTEBagging o que mostrou melhor desempenho para a maioria das sub-subclasses estudadas (9 de 12 de sub-subclasses), seguido do método SMOTEBoost. Apesar da subamostragem aleatória aliviar o desbalanceamento, é a sobreamostragem que traz ganhos substanciais aos classificadores, sendo portanto uma melhor escolha do que a subamostragem para o tratamento do desbalanceamento no caso de resíduos de aminoácidos catalíticos.

A introdução de descritores de conservação não ofereceu ganhos ou diferenças significativas em comparação com outros classificadores. Foi observado um aumento na sensibilidade dos classificadores treinados com conservação. No entanto, uma redução na precisão ocasionou em valores de F-measure

semelhantes àqueles obtidos sem o uso de conservação. Dessa forma, no caso de classificadores baseados em regras, descritores de conservação elevam o suporte das regras, mas levam a uma queda na confiabilidade (precisão) destas.

Os casos de maior sucesso na caracterização e predição de resíduos de aminoácidos catalíticos incluem as sub-subclasses das Glicosidasas (EC.3.2.1), com sensibilidade de 44% e precisão de 59%; as Monoester fosfórico hidrolases (EC.3.1.3), com sensibilidade de 62% e precisão de 70%; e as Serino Proteases (EC.3.4.21), com sensibilidade de 50% e precisão de 66.7% avaliadas no conjunto de teste elaborado neste trabalho (*benchmark*). No outro extremo (piores predições), encontram-se os classificadores para as sub-subclasses Pentosiltransferases (EC.2.4.2), com sensibilidade de 7% e precisão de 37.5%; e a sub-subclasse EC.2.5.1, com sensibilidade de 16.2% e precisão de 50%. Essas duas últimas estão conseqüentemente entre as sub-subclasses com menor número de resíduos de aminoácidos catalíticos após a transferência de anotações, 133 e 139, respectivamente, contra 544 no caso das Glicosidasas (EC.3.2.1).

Foi avaliado também o uso de SVM's a para predição de resíduos de aminoácidos catalíticos sem a separação por sub-subclasse EC, da mesma forma como empregado por outros métodos da literatura. Os classificadores genéricos, como foram chamados, atingiram desempenhos com base em sensibilidade, precisão e AUCROC próximos ou equivalentes aos métodos disponíveis na literatura, de forma que os descritores estruturais de proteínas presentes no Blue Star STING possuem capacidades similares aos descritores utilizados em outros trabalhos, quando aplicados na predição de resíduos de aminoácidos catalíticos. Nesses casos, o uso de descritores de conservação mostrou melhoras nos desempenhos dos classificadores, muito provavelmente devido ao tratamento adequado dos falsos positivos realizados pelas SVM's durante o treinamento, utilizando todos os atributos disponíveis, ao invés das regras que, uma vez construídas, ainda são podadas, resultando em antecedentes majoritariamente compostos por descritores de conservação que elevam a sensibilidade (maior cobertura).

Futuramente, pretende-se estudar abordagens de busca globais para a extração de regras, tais como abordagens evolutivas. Dentre as desvantagens desses métodos, encontram-se o alto custo computacional e a dificuldade em lidar com dados contínuos. No entanto, com a evolução da computação paralela e do processamento por placas gráficas (GPU's), algoritmos evolutivos podem se beneficiar de tais capacidades, viabilizando seu uso em problemas com alto número de atributos e amostras (Cano, et al., 2014a; Cano, et al., 2014b).

Devido ao alto desempenho dos descritores extraídos a partir da representação das cadeias proteicas como grafos não-direcionados, um estudo mais aprofundado sobre o uso exclusivo de grafos deve ser realizado para a obtenção de outros descritores com capacidades preditivas similares aos considerados

neste e em outros trabalhos. Inclusive, métodos de busca de padrões em grafos e subgrafos e classificação em grafos, via introdução de funções de *kernel* apropriadas, também podem trazer novas informações úteis sobre o comportamento dos resíduos de aminoácidos catalíticos, inclusive oferecendo melhores desempenhos (Xin, et al., 2010; Sanjaka & Changhui, 2013; Vacic, et al., 2010).

Classificadores baseados em regras utilizam uma abordagem de correspondência total (*full match*) para classificar uma amostra, ou seja, para que uma amostra seja coberta por uma regra, a amostra deve obedecer a todos os antecedentes da regra. Abordagens alternativas incluem correspondência parcial (*partial match*), onde uma amostra pode obedecer a apenas um subconjunto dos antecedentes de uma regra e, assim, recebe uma pontuação referente a esta correspondência, que pode ser utilizada para classificar a amostra (Zhang, et al., 2011).

Para viabilizar todo o trabalho descrito foram criados um banco de dados relacional para os descritores do Blue Star STING e uma biblioteca para cálculo, leitura e escrita dos descritores estruturais de proteínas. Foram também introduzidos e propostos novos descritores de proteínas baseados em grafos e descritores de vizinhança.

Espera-se que este trabalho tenha contribuído para um melhor entendimento dos mecanismos catalíticos em enzimas e forneça diretrizes para trabalhos futuros na exploração de novas técnicas e abordagens que se beneficiem da separação de enzimas segundo as reações catalisadas por estas. Com o rápido crescimento dos bancos de dados públicos de estruturas (e.g. PDB), acredita-se que a metodologia empregada neste trabalho possa ser expandida para outras sub-subclasses EC, fornecendo uma caracterização e uma classificação das enzimas segundo propriedades físico-químicas e estruturais de seus resíduos de aminoácidos catalíticos.

Capítulo 9 - Referências

- Adam, A. et al., 2010. *A Modified Artificial Neural Network Learning Algorithm for Imbalanced Data Set Problem*. s.l., s.n., pp. 44-48.
- Aizerman, M. A., Braverman, E. M. & Rozonoer, L. I., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, Volume 25, pp. 821-837.
- Aloy, P., Querol, E., Aviles, F. & Sternberg, M., 2001. Automated Structure-based Prediction of Functional Sites in Proteins: Applications to Assessing the Validity of Inheriting Protein Function from Homology in Genome Annotation and to Protein Docking. *Journal of Molecular Biology*, 311(2), pp. 395-408.
- Alterovitz, R. et al., 2009. ResBoost: characterizing and predicting catalytic residues in enzymes. *BMC Bioinformatics*, Volume 10, p. 197.
- Altschul, S. et al., 1990. Basic local alignment search tool. *J. Mol. Biol.*, Volume 215, pp. 403-410.
- Altschul, S. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.. *Nucleic Acids Research*, Volume 25, pp. 3389-402.
- Amitai, G. et al., 2004. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, Dec, 344(4), pp. 1135-1146.
- Andrews, R., Diederich, J. & Tickle, A., 1995. A survey and critique of techniques for extracting rules from. *Knowledge Based Systems*, Volume 8, pp. 373-389.
- Bairoch, A., 2000. The ENZYME database in 2000. *Nucleic Acids Research*, 28(1), pp. 304-305.
- Bairoch, A. & Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, Jan, 28(1), pp. 45-8.
- Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E., 2004. Swiss-Prot: juggling between evolution and stability. *Briefings in bioinformatics*, 5(1), pp. 39-55.
- Barandela, R., Valdovinos, R. & Sánchez, J., 2003. New Applications of Ensembles of Classifiers. *Pattern Analysis & Applications*, 6(3), pp. 245-256.
- Barlett, G., Porter, C., Borkakoti, N. & Thornton, J., 2002. Analysis of catalytic residues in enzyme active site. *Journal of Molecular Biology*, Volume 324, pp. 105-121.
- Basheer, M. A.-m. & Hamid, S., 2012. A Genetic Algorithm for Discovering Classification Rules in Data Mining. *International Journal of Computer Applications* 41(18):40-44, March 2012. *International Journal of Computer Applications*, 41(18), pp. 40-44.
- Bennett, K. & Mangasarian, O., 1992. Robust Linear Programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, Volume 1, pp. 23-34.
- Bennett, K. & Mangasarian, O., 1994a. Multicategory discrimination via linear programming. *Optimization Methods and Software*, Volume 3, pp. 29-39.
- Bennett, K. & Mangasarian, O., 1994b. Serial and parallel multicategory discrimination. *SIAM J. Optim.*, 4(4), p. 722-734.
- Ben-Shimon, A. & Eisenstein, M., 2005. Looking at enzymes from the inside out: The proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *Journal of Molecular Biology*, Volume 351, pp. 309-326.
- Benson, D. et al., 2014. GenBank. *Nucleic Acids Res.*, Jan, 42(1), pp. 32-7.
- Berman, H. M. et al., 2000. The Protein Data Bank. *Nucleic Acids Research*, Volume 28, pp. 235-242.
- Blagus, R. & Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics 2013*, Volume 14, p. 106.
- Błaszczczyński, J., Stefanowski, J. & Idkowiak, Ł., 2013. Extending Bagging for Imbalanced Data. Em: R. Burduk, et al. eds. *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Advances in Intelligent Systems and Computing ed. s.l.:Springer International

Publishing, pp. 269-278.

Bobadilla, L., Nino, F., Cepeda, E. & Patrorrovo, M. A., 2007. *Characterizing and Predicting Catalytic Residues in Enzyme Active Sites based on local properties: A machine learning approach*. s.l., s.n., pp. 938-945.

Boser, B., Guyon, I. & Vapnik, V., 1992. *A training algorithm for optimal margin classifiers*. s.l., s.n., p. 144.

Bray, T., Doig, A. J. & Warwicker, J., 2009. Sequence and structural features of enzymes and their active sites by EC class. *Journal of Molecular Biology*, Volume 386(5), pp. 1423-36.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24(2), pp. 123-140.

Breiman, L., Friedman, J., Olshen, R. & Stone, C., 1984. *Classification and regression trees*. Monterey(CA): Chapman and Hall/CRC.

Broadley, C. & Utgoff, P., 1995. Multivariate Decision Trees. *Machine Learning*, Volume 19, pp. 45-77.

Bryan, P. et al., 1986. Site-directed mutagenesis and the role of the oxyanion hole in subtilisin.. *Proc. Natl Acad. Sci USA*, 83(11), pp. 3743-5.

Cano, A., Zafra, A. & Ventura, S., 2014a. Parallel evaluation of Pittsburgh rule-based classifiers on GPUs. *Neurocomputing*, Volume 126, pp. 45-47.

Cano, A., Zafra, A. & Ventura, S., 2014b. Speeding up multiple instance learning classification rules on GPUs. *Knowledge and Information Systems*, pp. 1-19.

Cano, J. R., Herrera, F. & Lozano, M., 2003. Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study. *EEE Transactions on Evolutionary Computation*, 7(6), pp. 561-575.

Cano, J. R., Herrera, F. & Lozano, M., 2007. Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability. *Data Knowl. Eng.*, 60(1), pp. 90-108.

Capra, J. et al., 2009. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Computational Biology*, 5(12).

Ceroni, A., Costa, F. & Frasconi, P., 2007. Classification of small molecules by two- and three-dimensional decomposition kernels.. *Bioinformatics*, 23(16), pp. 2038-2045.

Chakravarti, I. M., Laha, R. G. & Roy, J., 1967. *Handbook of Methods of Applied Statistics*. New York: Wiley.

Chakravarty, S., Hutson, A., Estes, M. & Prasad, B., 2005. Evolutionary trace residues in noroviruses: importance in receptor binding, antigenicity, virion assembly, and strain diversity. *Journal of Virology*, 79(1), pp. 554-568.

Chapelle, O., Schölkopf, B. & Zien, A., 2006. *Semi-supervised learning*. Cambridge: MIT Press..

Chávez, E., Navarro, G., Baeza-Yates, R. & Marroquín, J. L., 2001. Searching in Metric Spaces. *ACM Computing Surveys*, 33(3), pp. 273-321.

Chawla, N., 2005. Data mining for imbalanced datasets: An overview.. Em: *The Data Mining and Knowledge Discovery Handbook*. s.l.:Springer, pp. 853-867.

Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. of Artificial Intelligence Research*, Volume 16, pp. 341-378.

Chawla, N., Japkowicz, N. & Kolcz, A., 2004. Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter*, 6(1), pp. 1-6.

Chawla, N., Lazarevic, A., Hall, L. & Bowyer, K., 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. Em: *Knowledge Discovery in Databases: PKDD 2003*. s.l.:Springer Berlin Heidelberg, pp. 107-119.

Chien, T. et al., 2008. E1DS: catalytic site prediction based on 1D signatures of concurrent conservation. *Nucleic Acids Research*, Volume 36, pp. 291-296.

Chien, Y.-T. & Huang, S.-W., 2013. On the Structural Context and Identification of Enzyme Catalytic Residues. *BioMed Research International*, Volume 2013, p. 9 pages.

- Chien, Y. T. & Huang, S., 2012. Accurate Prediction of Protein Catalytic Residues by Side Chain Orientation and Residue Contact Density. *Plos One*.
- Chitale, M. & Kihara, D., 2011. Computational Protein Function Prediction: Framework and Challenges. Em: *Protein Function Prediction for Omics Era*. Netherlands: Springer Netherlands, pp. 1-17.
- Chothia, C. & Lesk, A. M., 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, Volume 5, pp. 823-826.
- Cilia, E. & Passerini, A., 2010. Automatic prediction of catalytic residues by modeling residue structural neighborhood. *BMC Bioinformatics 2010, 11:115*, 11(1), p. 115.
- Cohen, W., 1995. *Fast Effective Rule Induction*. s.l., s.n., pp. 115-123.
- Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), pp. 273-297.
- da Silveira, C. et al., 2009. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins*, 74(3), pp. 727-43.
- de Beer, T. A. P., K., B., Thornton, J. M. & Laskowski, R. A., 2014. PDBsum additions. *Nucleic Acids Res*, Volume 42, pp. 292-296.
- de Moraes, F. et al., 2014. Improving predictions of protein-protein interfaces by combining amino acid-specific classifiers based on structural and physicochemical descriptors with their weighted neighbor averages. *PLoS One*, 9(1), p. e87107.
- Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on*, 6(2), pp. 182-197.
- Decherchi, S. & Rocchia, W., 2013. A general and Robust Ray-Casting-Based Algorithm for Triangulating Surfaces at the Nanoscale. *PLOS One*.
- Deng, H. & Runger, G., 2012. Feature Selection via Regularized Trees. *CoRR*, Volume abs/1201.1587.
- Dou, Y., Yang, J. & Zhang, C., 2012. L1pred: A Sequence-Based Prediction Tool for Catalytic Residues in Enzymes with the L1-logreg Classifier. *PLOS One*.
- Duda, R., Hart, P. & Stork, D., 2001. *Pattern Classification*. 2 ed. New York: John Wiley & Sons.
- Dukka, B. & Livesay, D., 2008. Improving position-specific predictions of protein functional sites using phylogenetic motifs. *Bioinformatics*, 24(20), pp. 2308-2316.
- Edelsbrunner, H. & Koehl, P., 2003. The weighted-volume derivative of a space-filling diagram. *Proc. Natl Acad. Sci.*, Volume 100, pp. 2203-2208.
- Elcock, A., 2001. Prediction of functionally important residues based solely on the computed energetics of protein structure. *Journal of Molecular Biology*, Volume 312, pp. 885-896.
- Eshelman, J., 1991. The CHC Adaptive Search Algorithm : How to Have Safe Search When Engaging in Nontraditional Genetic Recombination. *Foundations of Genetic Algorithms*, pp. 265-283.
- Fajardo, J. E. & Fiser, A., 2013. Protein structure based prediction of catalytic residues. *BMC Bioinformatics*, Volume 14, p. 63.
- Fawcett, T., 2006. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp. 861-874.
- Fawcett, T. E. & Provost, F., 1997. Adaptive fraud detection.. *Data Mining and Knowledge Discovery*, Volume 1, pp. 291-316.
- Fitzgerald, P. et al., 2005. The Macromolecular dictionary (mmCIF). Em: *Definition and exchange of crystallographic data*. s.l.:Springer, pp. 295-443.
- Freilich, S. et al., 2005. The complement of enzymatic sets in different species. *Journal of Molecular Biology*, Volume 349, pp. 745-63.
- Freud, Y. & Mason, L., 1999. *The Alternating Decision Tree Algorithm*. s.l., s.n., pp. 124-133.
- Freund, Y. & Schapire, R. E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, 55(1), pp. 119-139.
- Friedman, J., 1991. Multivariate Adaptive Regression Splines. *Annals of Statistics*, 19(1), pp. 1-67.
- Frigerio, F. et al., 1992. Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c

complex at 2.0 Å resolution. *J. Mol. Biol.*, 225(1), pp. 107-23.

Frishman, D. & Argos, P., 1995. Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics*, Volume 23, pp. 566-579.

Furnham, N. et al., 2014. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, Jan, 42(1), pp. 485-9.

Furnkranz, J. & Widmer, G., 1994. *Incremental Reduced Error Pruning*. s.l., s.n., pp. 70-77.

Furnkranz, J. & Widmer, G., 1994. Incremental Reduced Error Pruning.

Galar, M. et al., 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4), pp. 463-484.

Galar, M., Fernández, A., Barrenechea, E. & Herrera, F., 2013. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, 46(12), pp. 3460-3471.

García, S., Derrac, J., Cano, J. & Herrera, F., 2012. Prototype selection for nearest neighbor classification: taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3), pp. 417-435.

García, S. & Herrera, F., 2009. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. *Evolutionary Computation*, Volume 17, pp. 275-306.

Gasteiger, E. et al., 2005. Protein Identification and Analysis Tools on the ExPASy Server. Em: J. M. Walker, ed. *The Proteomics Protocols Handbook*. s.l.:Humana Press, pp. 571-607.

Gasteiger, E. et al., 2005. Protein Identification and Analysis Tools on the ExPASy Server. Em: J. M. Walker, ed. *The Proteomics Protocols Handbook*. s.l.:Humana Press, pp. 517-607.

George, R. et al., 2004. SCOPEC: a database of protein catalytic domains. *Bioinformatics*, Volume 20, pp. 130-136.

Goldgur, Y. et al., 1998. Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc Natl Acad Sci*, 95(16), pp. 9150-4.

Guerra-Salcedo, C., Chen, S., Whitley, D. & Smith, S., 1999. *Fast and accurate feature selection using hybrid genetic strategies*. s.l., s.n., pp. 177-184.

Guilloux, V., Schmidtke, P. & Tuffery, P., 2009. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, Volume 10, p. 168.

Gutteridge, A., Bartlett, G. J. & Thornton, J. M., 2003. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *Journal of Molecular Biology*, Volume 330, pp. 719-34.

Hall, M. et al., 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Han, L. et al., 2012. Identification of Catalytic Residues Using a Novel Feature that Integrates the Microenvironment and Geometrical Location Properties of Residues. *PLOS One*, July.

Han, S., Yuan, B. & Liu, W., 2009. *Rare Class Mining: Progress and Prospect*. s.l., s.n., pp. 1-5.

Hegyí, H. & Gerstein, M., 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology*, Volume 288, pp. 147-164.

He, H. & Garcia, E., 2009. Learning from imbalanced data. *IEEE Transactions on Data and Knowledge Engineering*, 21(9), pp. 1263-1284.

Higa, R. et al., 2006. Building multiple sequence alignments with a flavor of HSSP alignments. *Genet Mol Res.*, 5(1), pp. 127-37.

Huang, Y. et al., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5), pp. 680-682.

Hutchison, C. et al., 1978. Mutagenesis at a Specific Position in a DNA Sequence. *Journal of Biological Chemistry*, Volume 253, pp. 6551-6560.

Imam, T., Ting, K. & Kamruzzaman, J., 2006. z-SVM: An SVM for Improved Classification of Imbalanced Data. Em: *AI 2006: Advances in Artificial Intelligence*. s.l.:s.n., pp. 264-273.

- Innis, C., Shi, J. & Blundell, T., 2000. Evolutionary trace analysis of TGF- β and related growth factors: implications for site-directed mutagenesis. *Protein Engineering design & selection*, 13(12), pp. 839-847.
- James, M. et al., 1980. Crystal structure studies and inhibition kinetics of tripeptide chloromethyl ketone inhibitors with *Streptomyces griseus* protease B. *J. Mol. Biol.*, 139(3), pp. 423-38.
- Japkowicz, N., 2003. *Class Imbalances: Are we Focusing on the Right Issue?*. s.l., s.n.
- Johnstone, I. & Titterington, D., 2009. Statistical challenges of high-dimensional data. *Phil. Trans. R. Soc. A*, 367(1906), pp. 4237-4253.
- Joosten, R. et al., 2010. A series of PDB related databases for everyday needs.. *NAR*.
- Jordan, D. B. et al., 1999. Catalytic mechanism of scytalone dehydratase from *Magnaporthe grisea*. *Pesticide Science*, March, 55(3 - 9th International Congress of Pesticide Chemistry (IUPAC)), pp. 277-280.
- Kabsch, W. & Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.. *Biopolymers*, Volume 22, pp. 2577-2637.
- Kanehisa, M. et al., 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, Jan, 32(Database issue), pp. 277-80.
- Karp, G., 2008. *Cell and Molecular Biology: Concepts and Experiments*. New Jersey: John Wiley.
- Kass, G., 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), pp. 119-127.
- Kleywegt, G. & Jones, T., 1997. Model Building and Refinement Practice. *Methods in Enzymology*, Volume 277, pp. 208-230.
- Knowles, J. R., 1987. Tinkering with enzymes: what are we learning?. *Science*, Volume 236, pp. 1252-1258.
- Kohavi, R., 1996. *Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid*. s.l., AAAI Press., pp. 202-207.
- Ko, J. et al., 2005. Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. *Proteins*, Volume 59, pp. 183-195.
- Kotsiantis, S. B., 2013. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), pp. 261-283.
- Kristensen, D. M. et al., 2008. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, Volume 9, p. 17.
- Kubat, M., Holte, R. & Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, Volume 30, pp. 195-215.
- Kunik, V. et al., 2007. Functional Representation of Enzymes by Specific Peptides. *PLOS Computational Biology*, Volume 3, p. e167.
- Kyte, J. & Doolittle, R., 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1), pp. 105-132.
- La, D., Sutch, B. & Livesay, D., 2005. Predicting protein functional sites with phylogenetic motifs. *Proteins*, Volume 58, pp. 309-320.
- Landeweerd, G. et al., 1983. Binary tree versus single level tree classification of white blood cells. *Pattern Recogn.*, Volume 16, pp. 571-577.
- Landgraf, R., Xenarios, I. & Eisenberg, D., 2001. Three-dimensional Cluster Analysis Identifies Interfaces and Functional Residue Clusters in Proteins. *Journal of Molecular Biology*, 307(5), pp. 1487-1502.
- Laskowski, R., 2005. Structural Quality Assurance. Em: *Structural Bioinformatics*. Hoboken(NJ): John Wiley & Sons, Inc..
- Laurikkala, J., 2001. *Improving Identification of Difficult Small Classes by Balancing Class Distribution*. s.l., s.n., pp. 63-66.
- Lee, B., Park, K. & Kim, D., 2008. Analysis of the residue-residue coevolution network and the

functionally important residues in proteins. *Proteins*, Volume 72, pp. 863-872.

Lee, B. & Richards, F., 1971. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, Volume 55, pp. 379-400.

Lee, K., Fitch, C. & García-Moreno, B., 2002. Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein. *Protein Sci.*, 11(5), pp. 1004-1016.

Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, Volume 10, p. 707.

Lewis, D. D. & Catlett, J., 1994. *Heterogeneous uncertainty sampling for supervised learning*. San Francisco, Morgan Kaufmann, pp. 148-156.

Lichtarge, O., Bourne, H. & Cohen, F., 1996. An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, Volume 257, pp. 342-358.

Lijnzaad, P., Berendsen, H. & Argos, P., 1996. A method for detecting hydrophobic patches on protein surfaces.. *Proteins*, 26(2), pp. 192-203.

Li, M. et al., 2011. A local average connectivity-based method for identifying essential proteins from the network level. *Computational Biology and Chemistry*, 35(3), pp. 143-150.

Lin, C. et al., 2008. Hubba: hub objects analyzer - a framework of interactome hubs identification for network biology. *Nucleic Acids Res.*, 36(Web Server issue), pp. 438-43.

Lovell, S. et al., 2003. Structure validation by Calpha geometry: phi,psi and Cbeta deviation.. *Proteins*, 50(3), pp. 437-50.

Lovell, S., Word, J., Richardson, J. & Richardson, D., 2000. The penultimate rotamer library. *Proteins*, 40(3), pp. 389-408.

Lupyan, D., Leo-Macias, A. & Ortiz, A., 2005. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, 21(15), pp. 3255-63.

Ma, B., Wolfson, H. J. & Nussinov, R., 2001. Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Current Opinion in Structural Biology*, Volume 11, pp. 364-369.

Maciejewski, T. & Stefanowski, J., 2011. Local neighbourhood extension of SMOTE for mining imbalanced data. *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, 15(11), pp. 104-111.

MacQueen, J. B., 1967. *Some Methods for classification and Analysis of Multivariate Observations*. s.l., University of California Press., pp. 281-297.

Madabushi, S. et al., 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology*, Volume 316, pp. 139-154.

Malabanan, M., Amyes, T. & Richard, J., 2010. A role for flexible loops in enzyme catalysis. *Current Opinion in Structural Biology*, Volume 20, pp. 702-710.

Mancini, A. et al., 2004. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces.. *Bioinformatics*, 20(13), pp. 2145-2147.

Marimont, R. & Shapiro, M., 1979. Nearest Neighbour Searches and the Curse of Dimensionality. *IMA J. Appl. Math.*, 24(1), pp. 59-70.

Martin, A., 2004. PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, 20(6), pp. 986-8.

Martin, A., 2005. Mapping PDB chains to UniProtKB entries. *Bioinformatics*, 21(23), pp. 4297-301.

Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T., 2004. Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol Biol Evol.*, Volume 21, pp. 1781-1792.

Mitchell, T., 1997. *Machine Learning*. s.l.:McGraw Hill.

Mitternacht, S. & Berezovsky, I., 2011. A geometry-based generic predictor for catalytic and allosteric sites. *Protein Engineering, Design & Selection*, 24(4), pp. 405-409.

Mohri, M., Rostamizadeh, A. & Talwalkar, A., 2012. *Foundations of Machine Learning*. s.l.:The MIT Press..

- Murthy, S., Kasif, S. & Salzberg, S., 1994. A System for Induction of Oblique Decision Trees. *Journal of Artificial Intelligence Research*, Volume 2, pp. 1-32.
- Murthy, S., Kasif, S., Salzberg, S. & Beigel, R., 1993. *OCI: a randomized induction of oblique decision trees*. s.l., s.n., pp. 322-327.
- Napierala, K. & Stefanowski, J., 2012. BRACID: a comprehensive approach to learning rules from imbalanced data. *Journal of Intelligent Information Systems*, 39(2), pp. 335-373.
- Neshich, G. et al., 2006. The Star STING server: a multiplatform environment for protein structure analysis. *GMR*, Volume 5, pp. 717-722.
- Neshich, G. et al., 2014. Using Structural and Physical–Chemical Parameters to Identify, Classify, and Predict Functional Districts in Proteins—The Role of Electrostatic Potential. Em: W. Rocchia & M. Spaguolo, eds. *Computational Electrostatics for Biological Applications*. s.l.:Springer International Publishing, pp. 227-254.
- Neshich, G. et al., 2004. JavaProtein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Research*, 32(13), pp. 595-601.
- Noh, J. & Rieger, H., 2004. Random walks on complex networks. *Phys Rev Lett.*, 92(11), p. 118701 Epub.
- Ofran, Y., Punta, M., Schneider, R. & Rost, B., 2005. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discovery Today*, pp. 1475-82.
- Ondrechen, M., 2001. THEMATICS: A simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci USA*, Volume 98, pp. 12473-12478.
- Ooi, T., Oobatake, M., Némethy, G. & Scheraga, H., 1987. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci USA.*, 84(10), pp. 3086-3090.
- Ota, M., Kinoshita, K. & Nishikawa, K., 2003. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *Journal of Molecular Biology*, Volume 327, pp. 1053-1064.
- Papadakis, E. & Theocharis, B., 2006. A genetic method for designing TSK models based on objective weighting: Application to classification problems. *Soft. Computiong*, 10(9), pp. 805-824.
- Peng, I., Long, F. & Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1226-1238.
- Petrova, N. & Wu, C., 2006. Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics*, Volume 7, p. 312.
- Pickersgill, R., Smith, D., Worboys, K. & Jenkins, J., 1998. Crystal Structure of Polygalacturonase from *Erwinia carotovora* ssp. *carotovora*. *Journal of Biological Chemistry*, 273(38), pp. 24660-24664.
- Porollo, A. & Meller, J., 2007. Prediction-based fingerprints of protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 66(3), pp. 630-645.
- Porter, C., Bartlett, G. & Thornton, J., 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, Volume 32, pp. 129-133.
- Prati, R., Batista, G. & Monard, M., 2004. *Class imbalance versus class overlapping: an analysis of a learning system behavior*. s.l., s.n., pp. 312-321.
- Prlić, A. et al., 2012. BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, 28(20), pp. 2693-2695.
- Przulj, N., Wigle, D. & Jurisica, I., 2004. Functional topology in a network of protein interactions. *Bioinformatics*, Volume 20, p. 340–348.
- Pugalenthia, G., Kumarb, K. K., Suganthana, P. & Gangalb, R., 2008. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochemical and Biophysical Research Communications*, 367(3), pp. 630-634.
- Quinlan, J., 1986. Induction of Decision Trees. *Journal Machine Learning*, 1(1), pp. 81-106.
- Quinlan, J., 1990. Learning Logical Definitions from Relations. *Machine Learning*, 5(3), pp. 239-266.

- Quinlan, J., 1993. *C4.5: Programs for Machine Learning*. San Francisco(CA): Morgan Kaufmann Publishers Inc.
- Quinlan, J., 1996. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, Volume 4, pp. 77-90.
- Quinlan, R., 2001. *C5.0: An informal tutorial*. [Online] Available at: <http://www.rulequest.com/see5-unix.html> [Acesso em 2014].
- Radzicka, A. & Wolfenden, R., 1988. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, Volume 27, pp. 1664-1670.
- Ribeiro, C. et al., 2010. Analysis of binding properties and specificity through identification of the interface forming residues (IFR) for serine proteases in silico docked to different inhibitors.. *BMC Structural Biologu*, 20 10.pp. 10-36.
- Rissanen, J., 1978. Modeling by shortest data description. *Automatica*, 14(5), pp. 465-471.
- Rocchia, W. & Neshich, G., 2007. Electrostatic Potential Calculation for biomolecules - creating a database of pre-calculated values reported on a per residue basis for all PDB protein structures. *Genetics and Molecular Research*, 6(4), pp. 923-936.
- Rost, B. et al., 2003. Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, 60(12), pp. 2637-2650.
- Rückert, U. & De Raedt, L., 2008. An experimental evaluation of simplicity in rule learning. *Artificial Intelligence*, 2008, 172(1), pp. 19-28.
- Sanjaka, B. & Changhui, Y., 2013. *Prediction of enzyme catalytic sites on protein using a graph kernel method*. s.l., s.n., pp. 31-33.
- Sankararaman, S. & Sjölander, K., 2008. INTREPID—INformation-theoretic TREE traversal for Protein functional site IDENTification. *Bioinformatics*, 24(21), pp. 2445-2452.
- Schapire, R. E., 1990. The strength of weak learnability. *Mach. Learn*, Volume 5, p. 197–227.
- Schnoes, A., Brown, S., Dodevski, I. & Babbit, P. C., 2009. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLOS Computational Biology*.
- Schrauber, H., Eisenhaber, F. & Argos, P., 1993. Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *Proteins*, 77(1), pp. 97-110.
- Schreier, B. & Höcker, B., 2010. Engineering the enolase magnesium II binding site: implications for its evolution. *Biochemistry*, Sep, 49(35), pp. 7582-9.
- Schrödinger, L., 2010. *The {PyMOL} Molecular Graphics System, Version~1.3r1*. s.l.:s.n.
- Segura, J., Jones, P. & Fernandez-Fuentes, N., 2011. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinformatics*, Volume 12, p. 352.
- Seiffert, C., Khoshgoftaar, T., Van Hulse, J. & Napolitano, A., 2009. Rusboost: A hybrid approach to alleviating class imbalance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 40(1), pp. 185-197.
- Shapiro, L. & Harris, T., 2000. Finding function through structural genomics. *Current Opinion in Biotechnology*, pp. 31-5.
- Shapovalov, M. & Dunbrack Jr., R., 2011. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure*, 19(6), pp. 844-858.
- Sharma, K. R., 2009. *Bioinformatics*. Nova Iorque: McGraw-Hill.
- Shatsky, M., Nussinov, R. & Wolfson, H. J., 2002. MultiProt - A Multiple Protein Structural Alignment Algorithm. Em: R. Guigó & D. Gusfield, eds. *Algorithms in Bioinformatics*. Lecture Notes in Computer Science ed. s.l.:Springer Berlin Heidelberg, pp. 235-250.
- Shrake, A. & Rupley, J., 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol*, 79(2), pp. 351-371.

- Sievers, F. et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.. *Molecular Systems Biol*, Volume 7, p. 539.
- Sikonja, M. & Kononenko, I., 2003. Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, Volume 53, pp. 23-69.
- Sikora, R. & Piramuthu, S., 2007. Framework for efficient feature selection in genetic algorithm based data mining. *European Journal of Operational Research*, Volume 180, pp. 723-737.
- Sillitoe, I. et al., 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.. *Nucleic Acids Res.*, Jan, 41(Database issue), pp. 490-8.
- Sobolev, V. et al., 1999. Automated analysis of interatomic contacts in proteins.. *Bioinformatics*, 15(4), pp. 327-332.
- Somarowthu, S., Yang, H., Hildebrand, D. & Ondrechen, M., 2011. High-Performance Prediction of Functional Residues in Proteins with Machine Learning and Computed Input Features. *Biopolymers*, 95(6), pp. 390-400.
- Sridharan, S., Nicholls, A. & Honig, B., 1992. A new vertex algorithm to calculate solvent accessible surface areas. *Biophys*, Volume 61, p. A174.
- Stefanowski, J., 2013. Overlapping, Rare Examples and Class Decomposition in Learning Classifiers from Imbalanced. Em: *Emerging Paradigms in Machine Learning*. s.l.:Springer Berlin Heidelberg, pp. 277-306.
- Tang, J., Alelyani, S. & Liu, H., 2014. Feature Selection for Classification: A Review. Em: C. C. Aggarwal, ed. *Data Classification: Algorithms and Applications*. s.l.:Chapman and Hall/CRC, p. 37.
- Tang, Y., Sheng, Z., Y.Z., C. & Zhang, Z., 2008. An improved prediction of catalytic residues in enzyme structures. *Protein Engineering design & selection*, Volume 21, pp. 295-302.
- The UniProt Consortium, 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, Volume 42, pp. 191-198.
- Tian, W. & Skolnick, J., 2003. How well is enzyme function conserved as a function of pairwise sequence identity?. *Journal of Molecular Biology*, 333(4), pp. 863-882.
- Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser.*, Volume 58, pp. 267-288.
- Tickle, A., Andrews, R., Golea, M. & Diederich, J., 1998a. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Trans. Neural Networks*, Volume 9, pp. 1057-1068.
- Tickle, I., Laskowski, R. & Moss, D., 1998b. Error Estimates of Protein Structure Coordinates and Deviations from Standard Geometry by Full-Matrix Refinement of B- and B2-Crystallin. *Biological Crystallography*, Volume 54, pp. 243-252.
- Todd, A., Orenco, C. & Thornton, J., 2001. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, Volume 307, pp. 1113-1143.
- Tong, W. et al., 2009. Partial Order Optimum Likelihood (POOL): Maximum Likelihood Prediction of Protein Active Site Residues Using 3D Structure and Sequence Properties. *PLOS Computational Biology*, Jan.
- Topf, M., Várnai, P., Schofield, C. & Richards, W., 2002. Molecular dynamics simulations of the acyl-enzyme and the tetrahedral intermediate in the deacylation step of serine proteases. *Proteins*, 47(3), pp. 357-69.
- Tsodikov, O. V., Record, M. T. J. & Sergeev, Y. V., 2002. A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature.. *J. Comput. Chem.*, Volume 23, pp. 600-609.
- Utgoff, P., 1989. Perceptron Trees: a Case Study in Hybrid Concept Representations. *Connection Science*, Volume 1, pp. 377-391.
- Vacic, V., Iakoucheva, L., Lonardi, S. & Radivojac, P., 2010. Graphlet Kernels for Prediction of Functional Residues in Protein Structures. *J Comput Biol.*, 17(1), pp. 55-72.

- Veropoulos, K., C., C. & Cristianini, N., 1999. *Controlling the Sensitivity of Support Vector Machines*. s.l., s.n., pp. 55-60.
- Volkamer, A., Griewel, A., Grombacher, T. & Rarey, M., 2010. (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J Chem Inf Model.*, 50(11), pp. 2041-52.
- Volkamer, A., Kuhn, D., Rippmann, F. & Rarey, M., 2013. Predicting enzymatic function from global binding site descriptors. *Proteins: Structure, Function, and Bioinformatics*, March, 81(3), p. 479–489.
- Volpato, V., Adelfio, A. & Pollastri, G., 2013. Accurate prediction of protein enzymatic class by N-to-1 Neural Networks. *BMC Bioinformatics*, 14(Suppl 1), p. S11.
- Wallace, C. & Boulton, D., 1968. An information measure for classification. *The Computer Journal*, 11(2), pp. 185-194.
- Wang, J., Hartling, J. & Flanagan, J., 1997. The structure of ClpP at 2.3 Å resolution suggests a model for ATP-dependent proteolysis. *Cell*, 91(14), pp. 447-56.
- Wang, K. & Samudrala, R., 2006. measures, Incorporating background frequency improves entropy-based residue conservation. *BMC Bioinformatics*, Volume 7, p. 385.
- Wang, S. & Yao, X., 2009. *Diversity analysis on imbalanced data sets by using ensemble models*. s.l., s.n., pp. 324-331.
- Wang, X. et al., 2007. Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters*, 28(4), pp. 459-471.
- Webb, E., 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*. San Diego: International Union of Biochemistry and Molecular Biology by Academic Press.
- Wedekind, J., Reed, G. & Rayment, I., 1995. Octahedral coordination at the high-affinity metal site in enolase: crystallographic analysis of the MgII--enzyme complex from yeast at 1.9 Å resolution. *Biochemistry*, Apr, 34(13), pp. 4325-30.
- Weiss, G., 2004. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1), pp. 7-19.
- Weissig, H. & Bourne, P., 1999. An analysis of the Protein Data Bank in search of temporal and global trends. *Bioinformatics*, 15(10), pp. 807-831.
- Wei, Y., Ko, J., Murga, L. & Ondrechen, M., 2007. Selective Prediction of Interaction Sites in Protein Structures with THEMATICS. *BMC Bioinformatics*, Volume 8, p. 119.
- Welch, B. L., 1947. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1-2), pp. 28-35.
- Whitley, D., Beveridge, R., Guerra, C. & Graves, C., 1998. *Messy genetic algorithms for subset feature selection*. s.l., s.n., pp. 568-575.
- Wilson, D. & Martinez, T., 2000. Reduction Techniques for Instance-Based Learning Algorithms. *Machine Learning*, Volume 38, pp. 257-286.
- Wu, J., A Xiong, H. & A Chen, J., 2010. COG: local decomposition for rare class analysis. *Data Mining and Knowledge Discovery*, 20(2), pp. 191-220.
- Xin, F. et al., 2010. Structure-based kernels for the prediction of catalytic residues and their involvement in human inherited disease. *Bioinformatics*, 26(16), pp. 1975-1982.
- Yahalom, R. et al., 2011. Structure-based identification of catalytic residues. *Proteins: Structure, Function, and Bioinformatics*, June, 79(6), pp. 1952-1963.
- Yao, H. et al., 2003. An accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures. *Journal of Molecular Biology*, 326(1), pp. 255-261.
- Youn, E., Peters, B., Radivojac, P. & Mooney, S., 2007. Evaluation of features for catalytic residue prediction in novel folds. *Protein Science*, Volume 16, pp. 216-226.
- Yuan, Z., Zhao, J. & Wang, Z., 2003. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Engineering design & selection*, Volume 16, pp. 109-114.
- Yu, H. et al., 2007. The importance of bottlenecks in protein networks: correlation with gene

essentiality and expression dynamics. *PLoS Comput. Biol.*, Volume 3 e59.

Zhang, J., Bala, J. W., Hadjarian, A. & Han, B., 2011. Ranking Cases with Classification Rules. Em: J. Fürnkranz & E. Hüllermeier, eds. *Preference Learning*. s.l.:Springer Berlin Heidelberg, pp. 155-177.

Zhang, T. et al., 2008. Accurate sequence-based prediction of catalytic residues. *Bioinformatics*, 24(20), pp. 2329-2338.

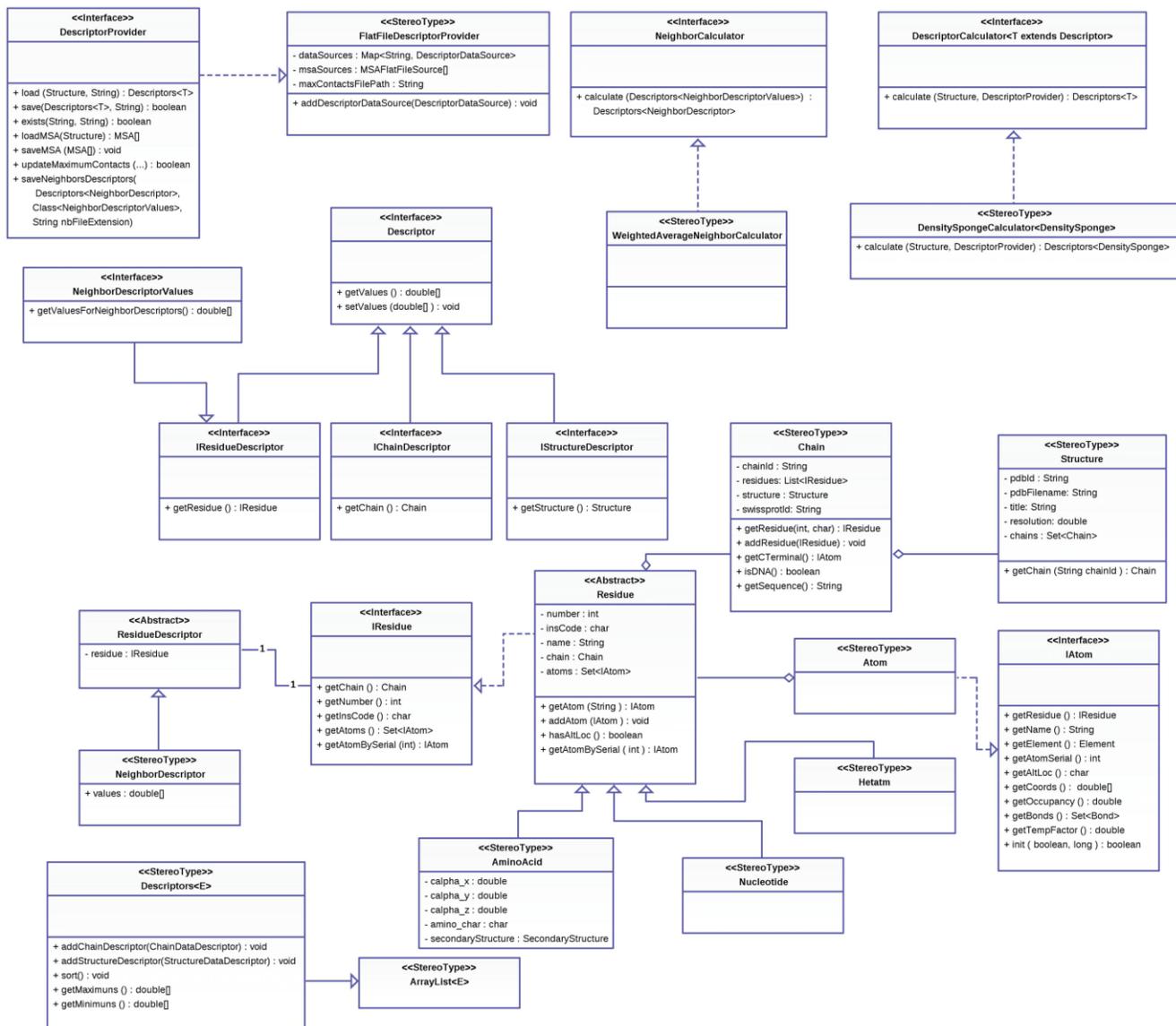
Zhu, S. et al., 2004. Evolutionary trace analysis of scorpion toxins specific for K-channels. 54(2), pp. 361-370.

Zięba, M., Tomczak, J., Lubicz, M. & Świątek, J., 2014. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*, Volume 14 Part A, pp. 99-108.

Zvelebil, M. & Sternberg, M., 1988. Analysis and prediction of the location of catalytic residues in enzymes. *Protein engineering design & selection*, Volume 2, pp. 127-138.

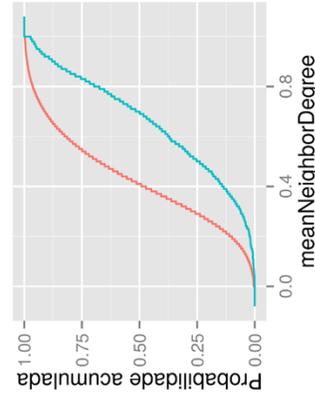
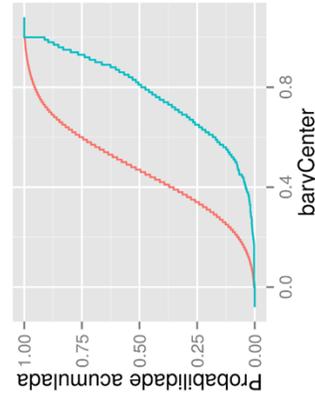
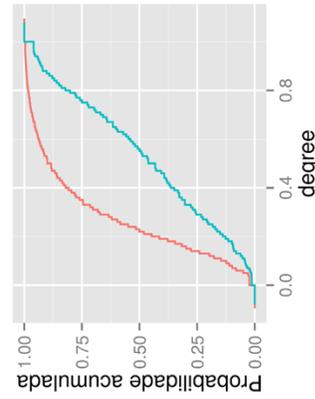
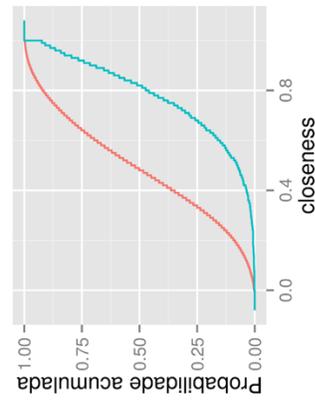
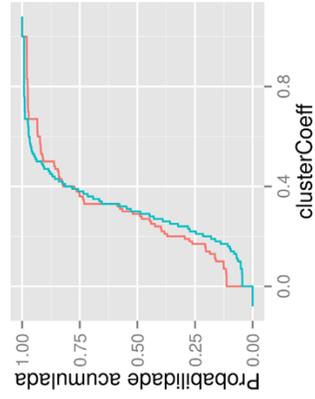
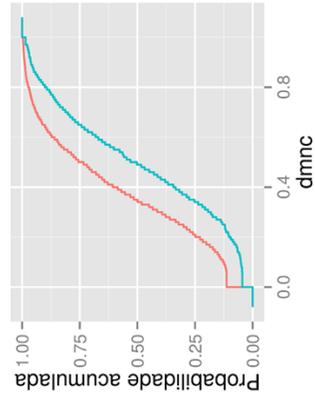
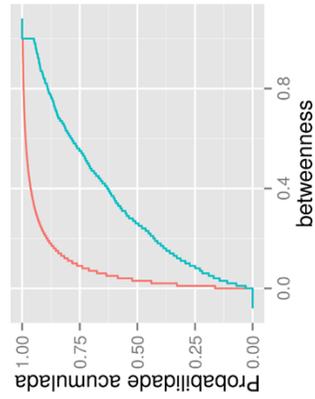
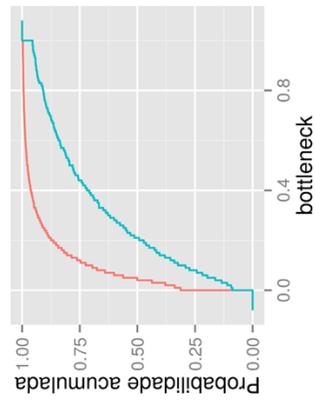
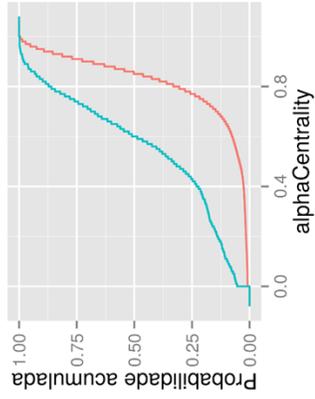
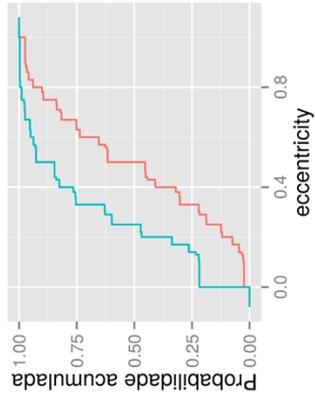
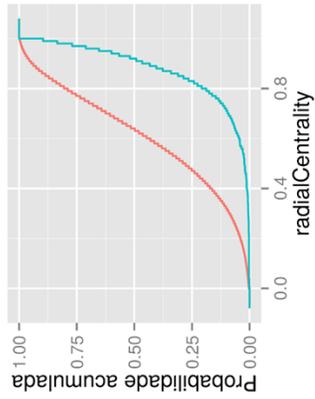
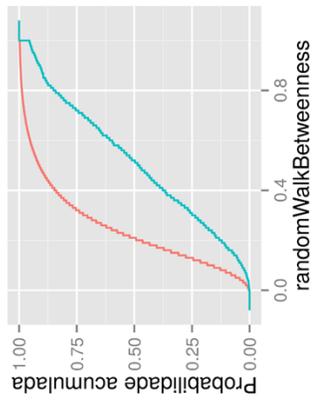
Apêndice B - Modelo UML de Classes Simplificado - STING Descriptor Library (SDL)

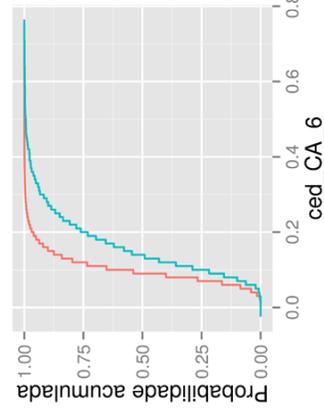
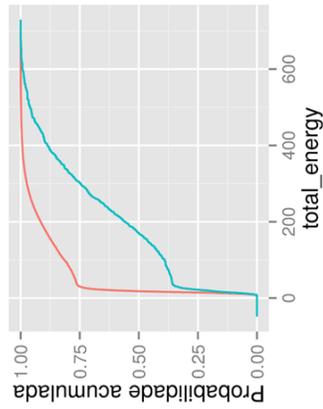
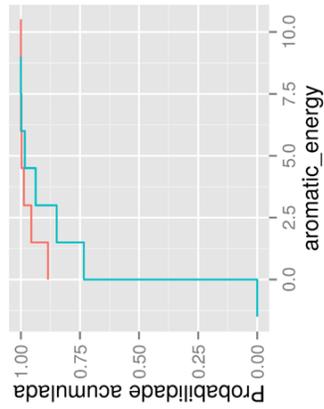
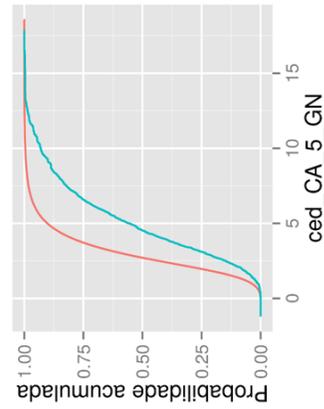
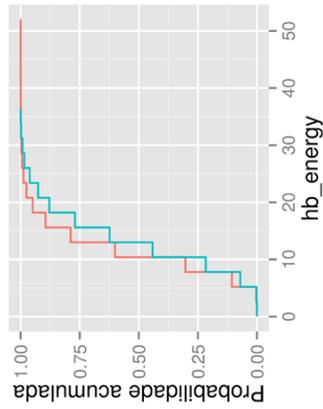
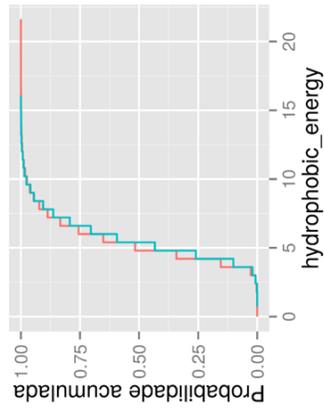
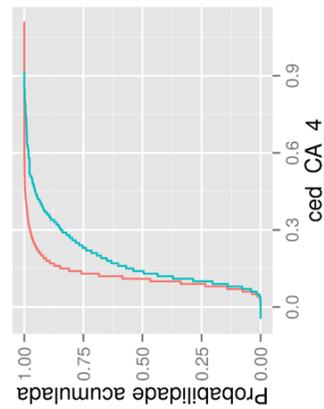
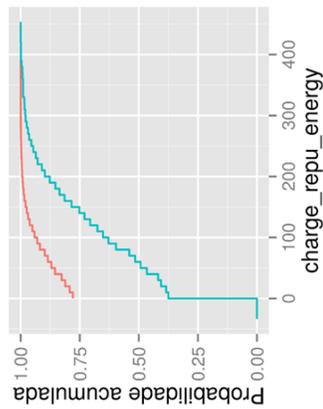
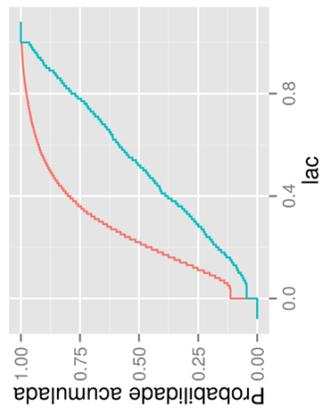
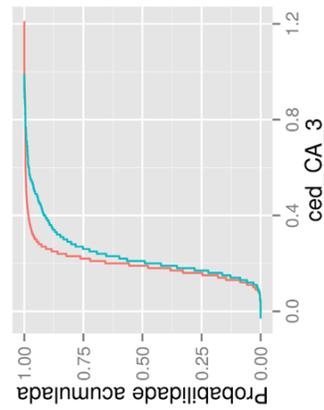
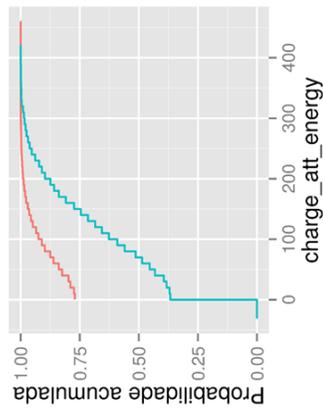
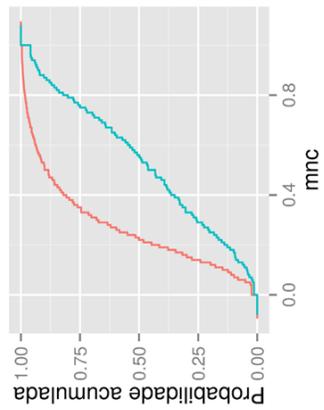
Na figura é apresentado o diagrama UML de Classes de forma simplificada da SDL, desenvolvida neste trabalho.

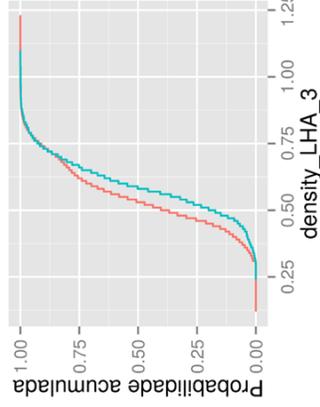
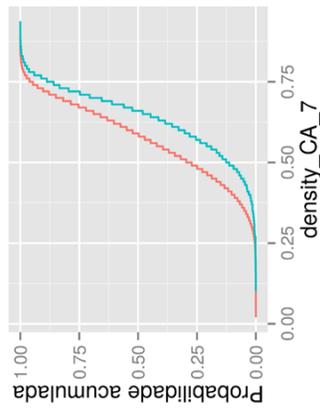
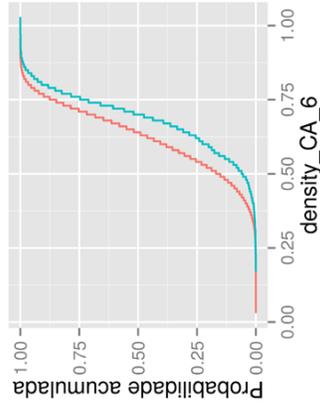
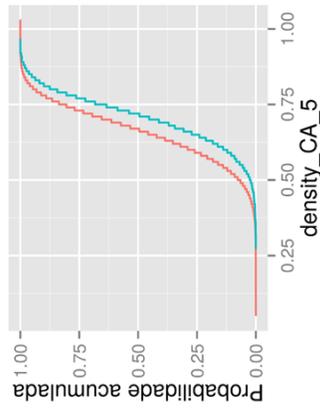
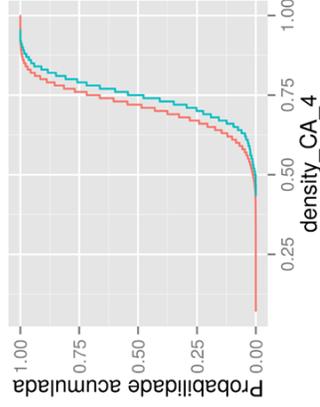
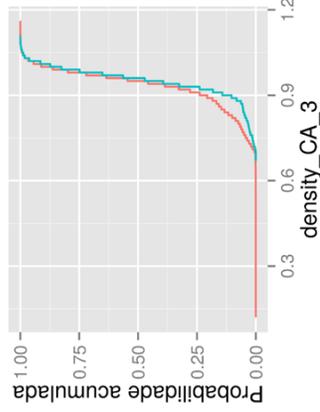
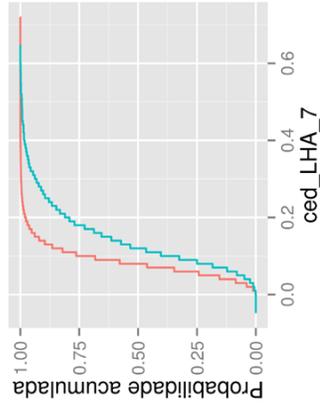
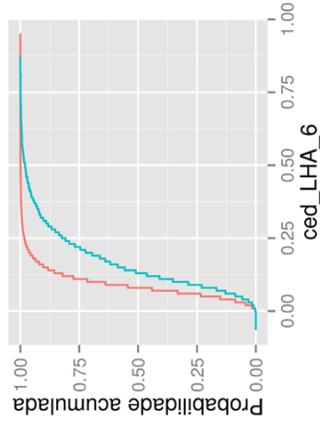
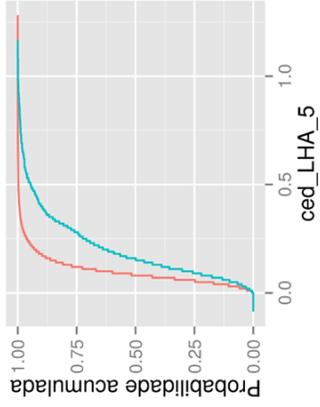
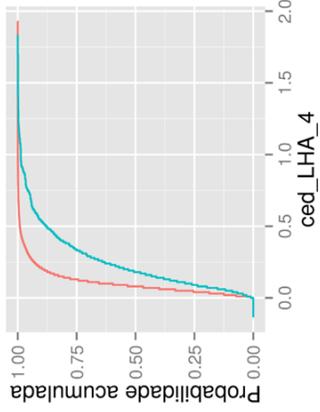
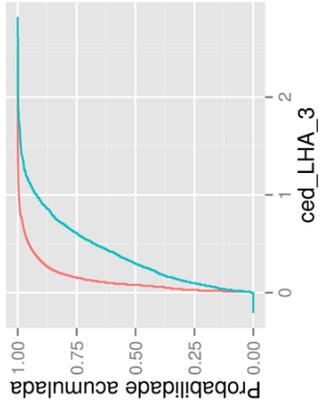
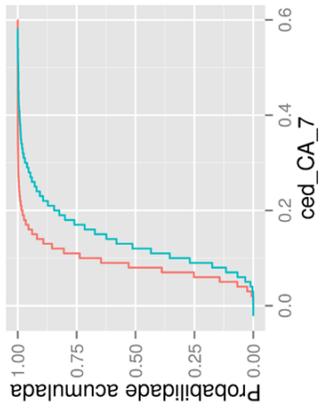


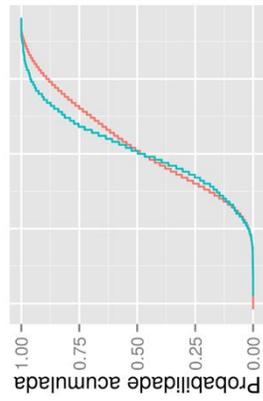
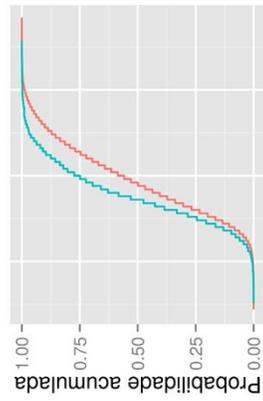
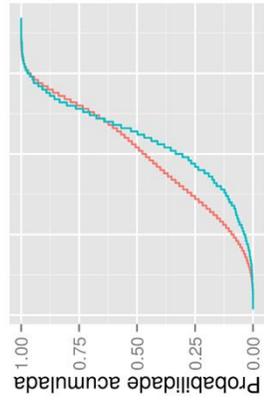
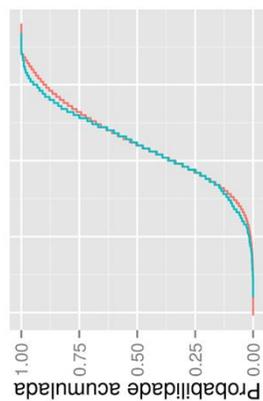
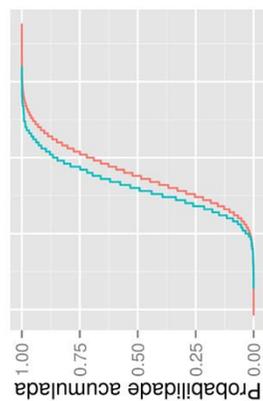
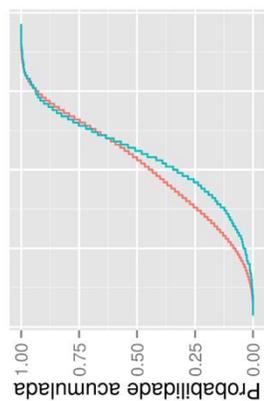
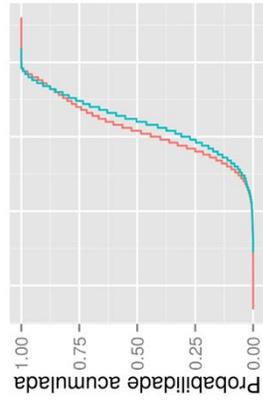
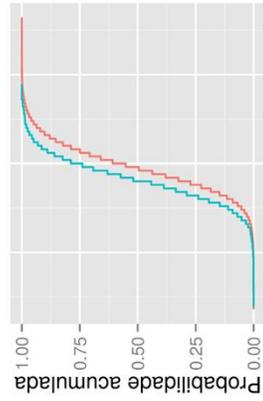
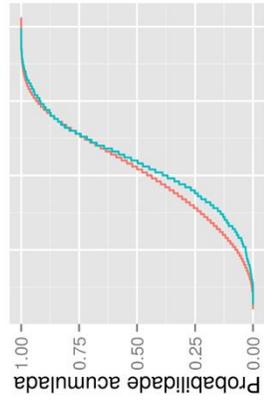
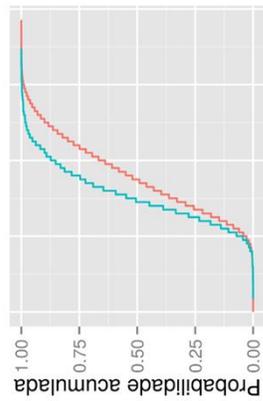
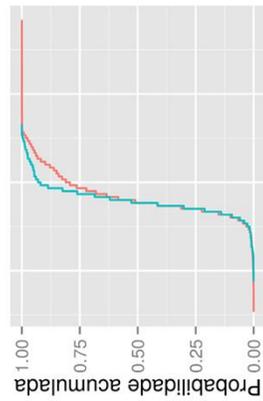
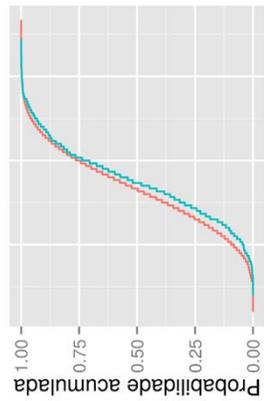
Apêndice C - Funções de distribuição acumulada empírica para os descritores do Blue Star STING

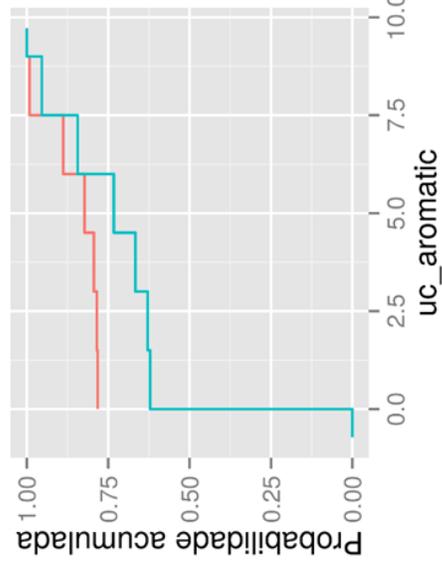
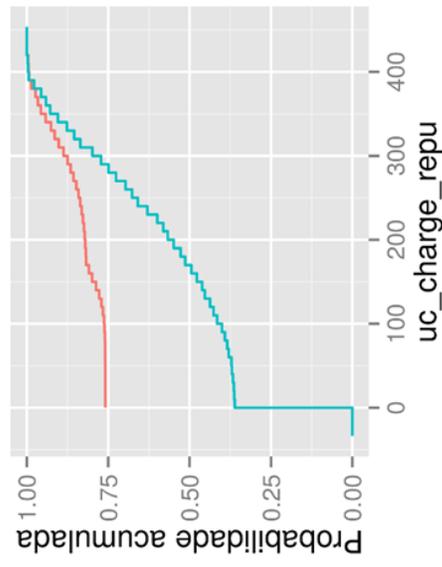
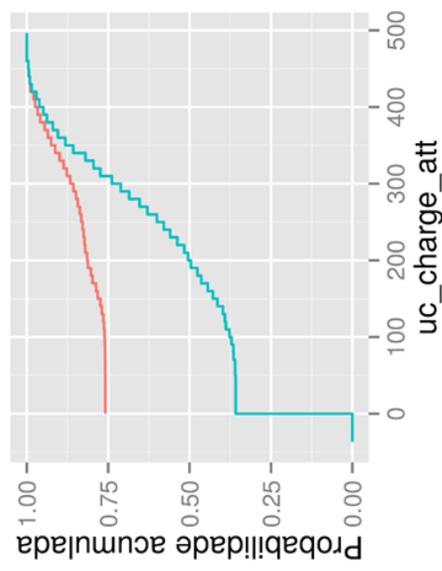
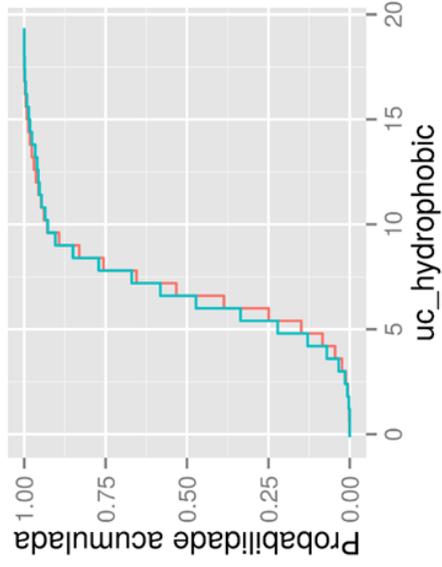
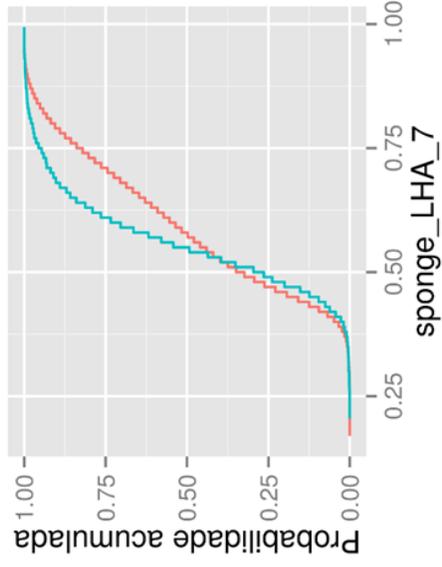
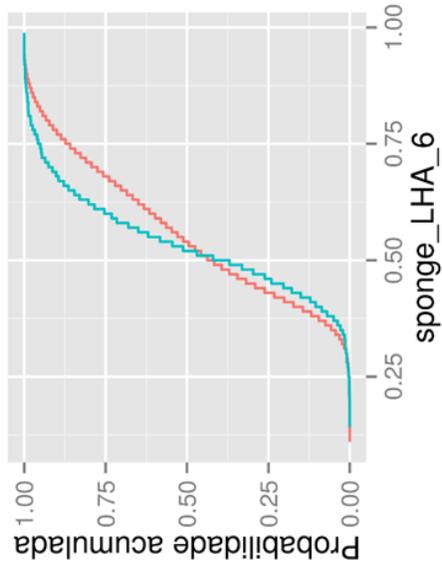
Nos gráficos das figuras a seguir são apresentadas as FDA's para os descritores de proteínas utilizados neste trabalho. Os gráficos ilustram as diferenças entre os FDA's para as duas classes de resíduos de aminoácidos, de forma que seja possível identificar aqueles que apresentam maiores diferenças entre as duas distribuições.

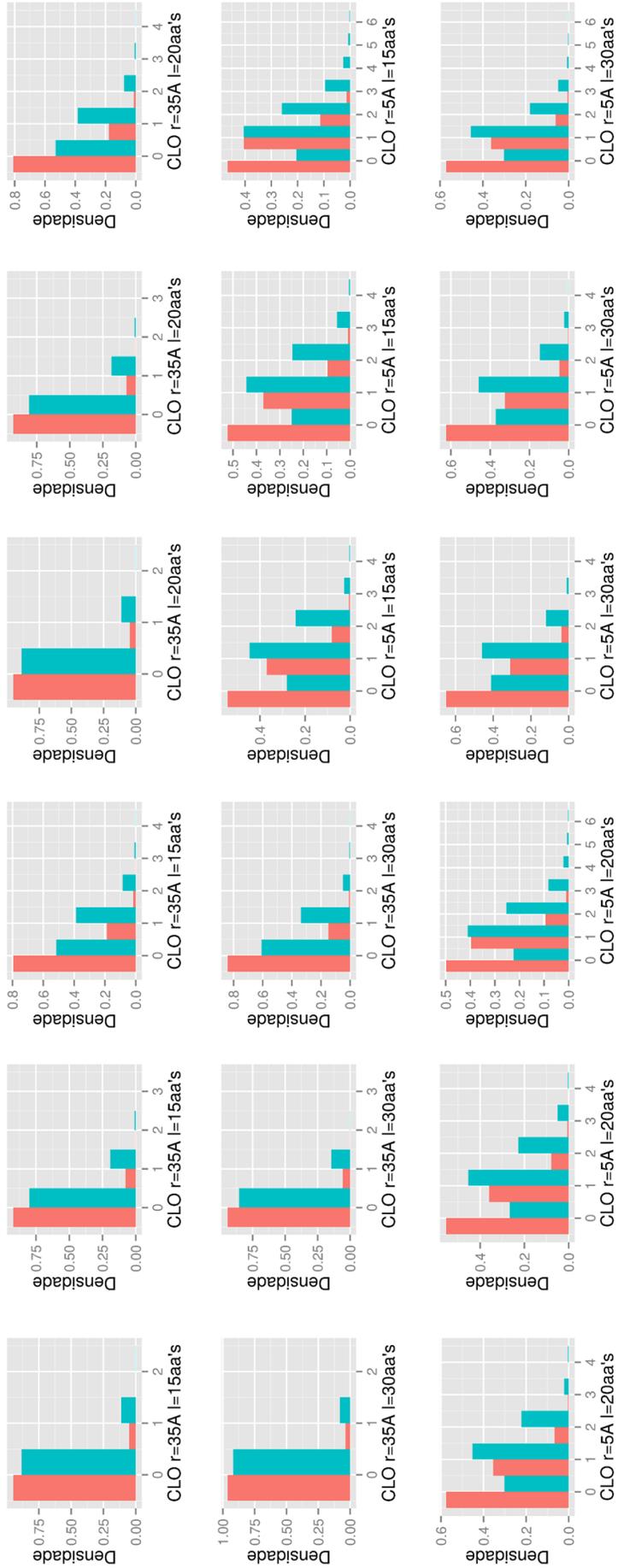


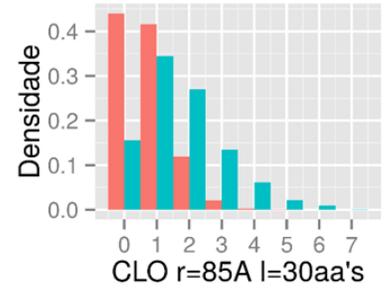
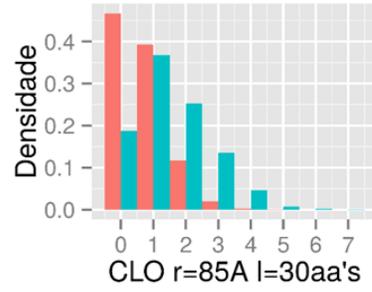
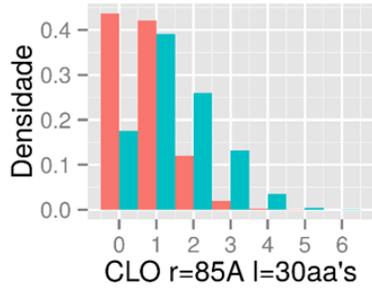
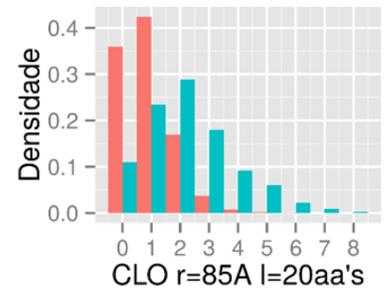
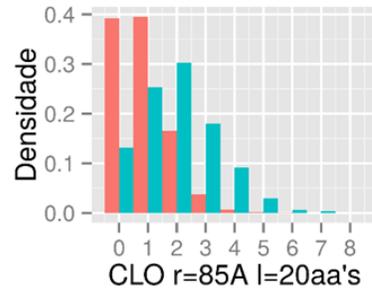
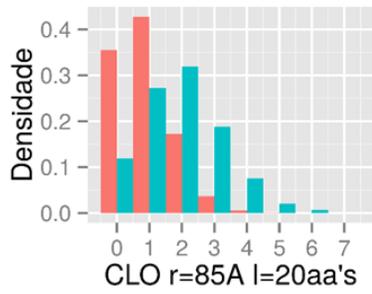
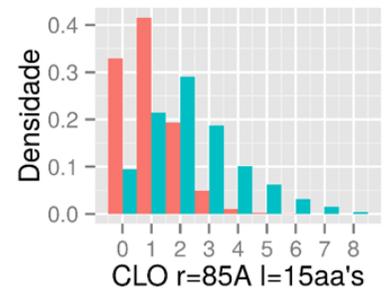
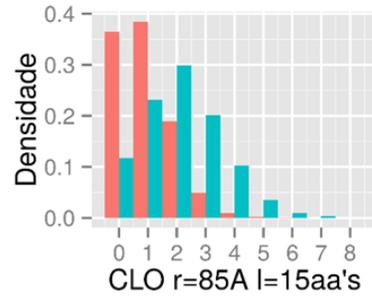
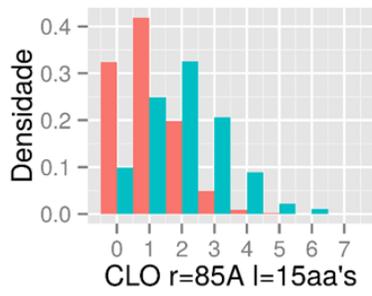


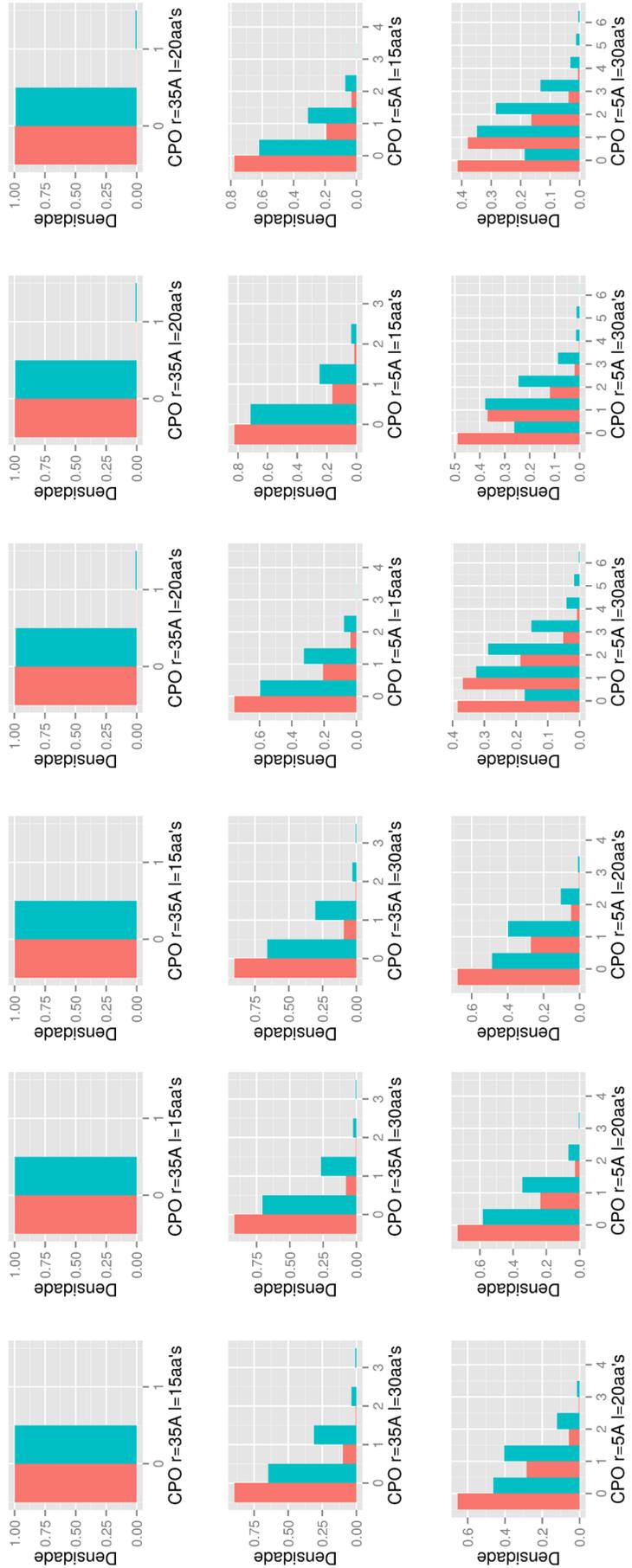


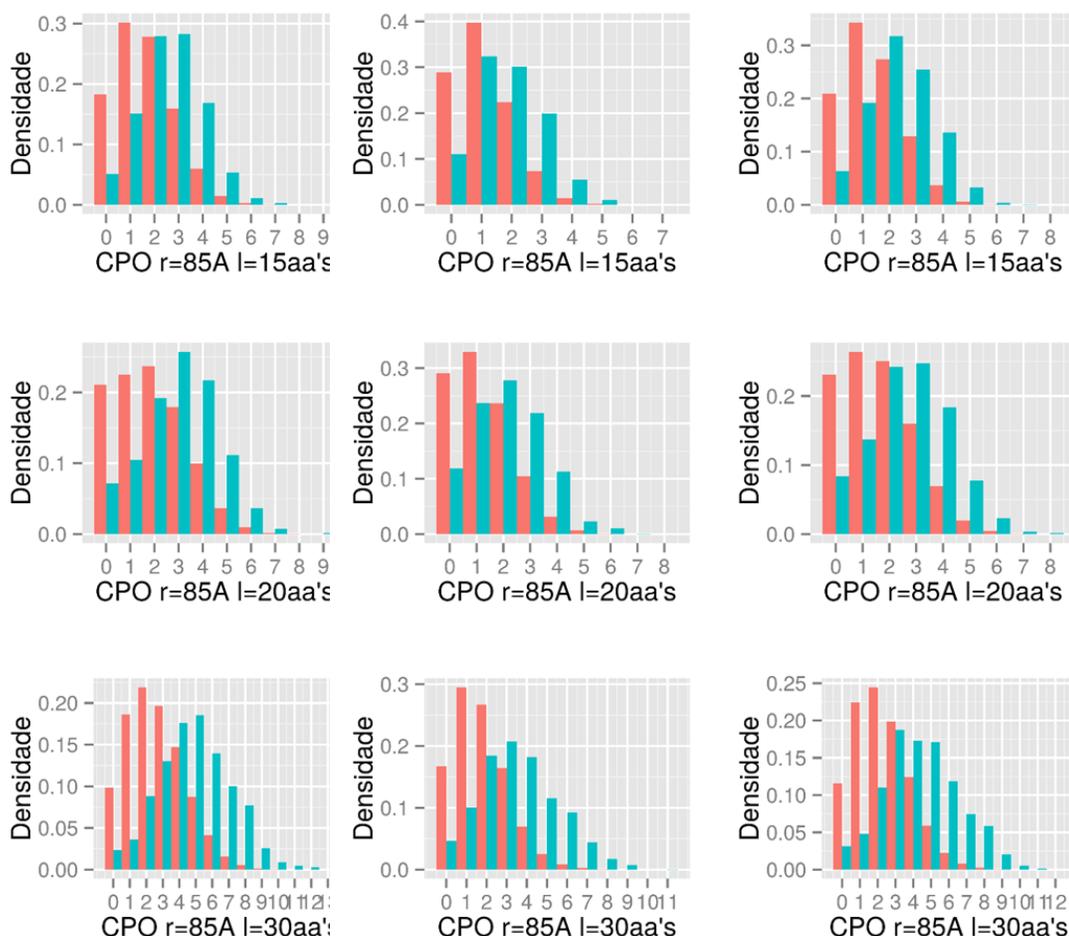












Apêndice D - Lista de regras encontradas utilizando-se JPD para diversas famílias enzimáticas durante os anos 2008 e 2009

Na tabela a seguir são mostrados os resultados do processo manual aplicado para diversas enzimas utilizando-se o ^JPD, obtidos durante análises nos anos de 2008 e 2009. O objetivo é exemplificar de forma simples a prova de conceito que foi usada para iniciar este trabalho.

FAMÍLIAS E PDB'S	EC	REGRAS 2008
ACETYLCHOLINESTERASE-LIKE (1QO9; 2ACE; 2GYU; 1B41)	3.1.1.7	Conservation (Sting): Relative Entropy ≤ 27 ; Evolutionary Pressure ≤ 12 ; Reliability ≥ 50 Hydrofobicity: Constant < -3 ; Complex < 0 Electrostatic Potential@CA - Min.: -40; Máx.: -28; -14; -6; -9.
ACETYLCHOLINESTERASE-LIKE (2ACE; 1F8U; 1P0M; 1DX4; 1C7I)	3.1.1.7	Cross Presence Order: Order@LHA ≥ 4 ; 5 Distances from Center of Gravity < 13 Conservation HSSP: Relative Entropy ≤ 24 ; Evolutionary Pressure ≤ 11 ; Conservation SH2QS: Relative Entropy ≤ 27 ; Relative Entropy 100 ≥ 2 Secondary Structure PDB: Helix; Turn;

<p>CYTRATE SYNTASE (2CTS; 1CSH; 1IXE; 1NXE)</p>	<p>2.3.3.1 2.3.3.16</p> <p>Cross Link Order: Order@CA ≤ 2; Order@CB ≤ 2; Order@LHA ≤ 2 Pocket in Complex: Volume(#water volumes) ≥ 4 Density: @CA Radius = 3; SW = 3; Sliding Window(x100) < 130 Conservation HSSP: Evolutionary Pressure ≤ 8 Conservation SH2QS: Relative Entropy ≤ 25; Evolutionary Pressure ≤ 5 Unused Contacts: Energy Total > 200 Hydrophobic Scale: Constant < 0; Complex < 0; Isolation < 0</p>
<p>CYTRATE SYNTASE (1CSH; 2CTS; 1AJ8; 1A59)</p>	<p>2.3.3.1 2.3.3.16</p> <p>Relative Entropy(HSSP) < 5; 10 Unused contact energy (total) > 397; 440; 255; 340 All Internal contact energy (total) > 20; 40; 15; 8 Ligant Pocket Residue selection Apenas para 1CSH: Curvature in complex < 20</p>
<p>PHAGE LYSOZYME (2LZM; 2L78; 103L; 7LZM)</p>	<p>3.2.1.17</p> <p>Curtavure in Complex: < 10; =<10; =<19; =<12 Pocket in Complex: ≥0; >0 Unused Energy Total >385; ≥385 Surface Accessibility: in Isolation<46</p>
<p>ALPHA AMYLASES (1Z32; 1SMD; 1XCX; 1MFV; 1HNY; 1KGW)</p>	<p>3.2.1.1</p> <p>Conservation HSSP: Evolutionary Pressure ≤ 1 Electrostatic Potential: @LHA Range: [-239, -178] Cross Presence Order: Order@LHA ≥ 6</p>
<p>CCP-LIKE (1CCA; 1CCP; 1BEK; 1BEP)</p>	<p>1.11.1.5</p> <p>Cross Presence Order: Order@LHA ≥ 3;2; Order@CB ≥ 2 Sponge(@CA Radius = 3: Internal(x100) ≥ 84; 86; 86; 84 Conservation SH2QS:Evolutionary Pressure ≤ 2; 3; Reliability (x100) ≥ 37; 37; Relative Entropy 100 ≤ 31 Electrostatic Potencial: @CA:[4,12],[5,16]; @LHA: Min.: -220, Máx.: -102; -96; -87; -80; Average: [-20,20] Contacts Energy: All Contacts; All Contacts; Internal Energy ≥ 13; 18.</p>
<p>SERINE PROTEASES (1MEE; 1TK2; 1LW6; 1SBN; 1TEC)</p>	<p>3.4.21.62 3.4.21.66</p> <p>Conservation HSSP: Relative Entropy ≤ 4, 5, 5, 6, 5; Evolutionary Pressure ≤ 0 Unused Contacts: Total ≥ 55, 69, 52, 45, 69</p>
<p>SERINE PROTEASES (1PPF; 1T32; 1AZZ; 1UTL)</p>	<p>3.4.21.37 3.4.21.20 3.4.21.32 3.4.21.4</p> <p>Conserv.(Sting): Evol. Pressure: < 10; Relative Entropy 100: < 10; Relative Entropy: < 10 Accessibility: Isolation: > 1; Relative:< 0,2 Electrostatic Potencial: @LHA: < -80,00 Cross Presence Order: Order@CA: > 2</p>
<p>ARGINASE-LIKE (1D3V; 1RLA; 1PQ3)</p>	<p>3.5.3.1</p> <p>Unused Contacts Energy ≥ 278, 348; Contacts ≤ 69, 103, 85 Conservation SH2QS: Evolutionary Pressure ≤ 0, 2; Relative Entropy 100 ≤ 2</p>
<p>GLYCOSYL HYDROLSES (6TAA; 1BVN; 1B2Y; 1AMY; 1B1Y)</p>	<p>3.2.1.1 3.2.1.2</p> <p>Cross Presence Order@CA ≤ 3 ; @CB≤4 ; @LHA ≥ 3 Electrostatic Potential @Surface: Min.: -300, Máx.: -1, 1, -8; @CA: [-27, 300] Internal Contacts: Charged Repulsive (and) Hydrophobic Scale ;Constant: <-6 Secondary structure: unselect beta sheet</p>

<p>GLYCOSYL HYDROLASES (2BVW; 1OC6; 1QK0, 1QK2; 1CB2)</p>	<p>3.2.1.91</p>	<p>Conservation (HSSP): Relative Entropy < 1; Reliability: > 41 Conservation – AA-CoEvolution – HSSP: > 2, 3 Geometric – Cross Presence Order@LHA: > 1 Unused Energy Total: > 90, 140</p>
<p>VERTEBRATE PHOPHOLIPASE A2 (1GMZ; 2NOT)</p>	<p>3.1.1.4</p>	<p>Cross Link Order: @CA ≥ 1, 0; @CB ≥ 1, 0; Order@LHA ≥ 1, 0 Distances from N Terminal < 15 Conservation HSSP: Relative Entropy ≤ 100, 75; Relative Entropy 100 ≤ 10, 5; Evolutionary Pressure ≤ 1, 10 Diff Conservation: Relative Entropy < 80; Relative Entropy 100 ≤ 15, 8; Evolutionary Pressure ≤ 2, 5 Secondary Structure (PDB, DSSP, Stride): Helix Electrostatic Potencial@LHA: Range:[-300,25] Hydrophobic Scale: Constant < 0, ≤ 0; Complex < 4; Isolation < 4</p>
<p>VERTEBRATE PHOPHOLIPASE A2 (1FXP; 1A2A; 1PP2; 1G4I; 1ZBL; 1QKU)</p>	<p>3.1.1.4</p>	<p>Unused Contacts: Energy ≥ 175; Contacts ≥ 34 Cross Presence Order@CA ≥ 2</p>
<p>HAL-PAL LIKE (1W27)</p>	<p>4.3.1.24</p>	<p>Pocket in Complex: Volume(#water volumes) ≥ 1 Conservation SH2QS: Relative Entropy 100 ≤ 21; Evolutionary Pressure ≤ 9 Unused Contacts: Energy Total ≥ 70 Protein Ligand Contacts: Selection Mode:OR; Undefined Relevant Sites: Residue Location:Surface</p>
<p>TIROSINA QUINASES (2J0K)</p>	<p>2.7.10.2</p>	<p>Conservation (HSSP): Relative Entropy ≤ 13; Relative Entropy 100 ≤ 0; Pressure Evol. ≤ 2 Structural : Secondary Element PDB: Helix, Turn, Coil Temperature factor: @CA < 75; @LHA < 78 Unused Energy: Contacts Total ≤ 78; Energy Total ≤ 373 Rotamer: % Rotamer Library ≤ 8</p>
<p>CYCLOPHILIN (PEP-TIDYLPROLYL ISOMERASE) (1CWA)</p>	<p>5.2.1.8</p>	<p>Conservation (HSSP): Relative Entropy ≤ 10 Surface Accessibility: In isolation > 5; In complex > 5 Pocket in isolation: Volume (# water volumes) ≥ 0 Electrostatic Potential: Average: Minimum = 0; Maximum = 300</p>
<p>PHOSPHOFRUCTOKINASE (3PFK; 2PFK; 4PFK; 6PFK; 1MTO)</p>	<p>2.7.1.11</p>	<p>Cross Presence Order: Order@CA ≥ 0 ; Order@LHA ≥ 0, 0, 1, 1 Curvature in Complex: Curvature Value (x100) ≤ 27.0 Sponge: @CA Radius = 3 ; SW = 3 ; Internal(x100) ≤ 85 ; Sliding Window(x100) ≤ 83 Electrostatic Potencial: Exclude neutral residues @CA: [-18,18]; [-300,25]; [-9,22] @LHA: [-171,14]; [-189,18]; [-171,25]; [-174,14] @Surface: [-18,2]; [-9,2]; Range:[-9,6]; [-12,6]; [-6,6] @Average: [-54,10]; [-42,22] Unused Contacts: Total ≥ 36 / Energy Total ≥ 74; 86; 82 Hydrophobic Scale: Constant < 1 / Complex < 1 / Isolation < 1 Pocket in Complex: Volume(#water volumes) ≥ 0, 13, 0</p>

CYSTEINE PROTEINASES
(**1PPN**; **1NLN**; **1GEC**;
1PPO; **1AEC**)

3.4.22.2
3.4.22.39
3.4.22.25
3.4.22.30
3.4.22.14

Pocket in Isolation: Volume(#water volumes) $\geq 0, 6$
Conservation HSSP: Relative Entropy $\leq 21, 22, 17$;
Evolutionary Pressure $\leq 4, 4, 5$ / Reliability (x100) ≥ 88
Relative Entropy 100 ≤ 5
Conservation SH2QS: Relative Entropy $\leq 19, 19$
Evolutionary Pressure ≤ 4 / Reliability (x100) $\leq 98, 98$
Secondary Structure PDB: Sheet / Turn / Coil
Residue Property: Type: Hydrophilic

Relevant Sites: Prosite; Residue Location: Surface
Conservation HSSP: Relative Entropy $\leq 7, 7, 9, 9, 9$
Reliability (x100) $\geq 96, 92$
Residue Property: Type: Hydrophilic
Dihedral Angles: Most favoured regions
Contacts Energy All Contacts: Internal Energy ≤ 13 (Except 1NLN)
Cross Link Order: Order@CB $\leq 1, \geq 1$

FAMILIAS E PDB'S	EC	REGRAS 2009
ACETYLCHOLINESTERASE-LIKE (1QO9 ; 2ACE ; 2GYU ; 1B41)	3.1.1.7	Conservation (Sting): Relative Entropy ≤ 27 ; Evolutionary Pressure ≤ 12 ; Reliability ≥ 50 Hydrophobicity: Constant < -3 ; Complex < 0 Electrostatic Potential@CA - Min.: -40; Max.: -28; -14; -6; -9.
ACETYLCHOLINESTERASE-LIKE (2ACE ; 1F8U ; 1P0M ; 1DX4 ; 1C7I)	3.1.1.7	Cross Presence Order : Order@LHA $\geq 4 ; 5, >1$ Distances from Center of Gravity $< 13; 15$ Cross Link Order : Order@CA ≥ 3 Conservation HSSP: Relative Entropy ≤ 24 ; Evolutionary Pressure ≤ 11 ; Conservation SH2QS: Relative Entropy $\leq 27, 26$; Relative Entropy 100 ≥ 2 Secondary Structure PDB: Helix; Turn; Contacts: Unused contacts: Unused Energy Total > 42
CYTRATE SYNTASE (2CTS ; 1CSH ; 1IXE ; 1NXE)	2.3.3.1 2.3.3.16	Cross Link Order: Order@CA ≤ 2 ; Order@CB ≤ 2 ; Order@LHA ≤ 2 Pocket in Complex: Volume(#water volumes) ≥ 4 Density: @CA Radius = 3; SW = 3; Sliding Window(x100) < 130 Conservation HSSP: Evolutionary Pressure ≤ 8 Conservation SH2QS: Relative Entropy ≤ 25 ; Evolutionary Pressure ≤ 5 ; 1 Unused Contacts: Energy Total > 200 Hydrophobic Scale: Constant < 0 ; Complex < 0 ; Isolation < 0
CYTRATE SYNTASE (1CSH ; 2CTS ; 1AJ8 ; 1A59)	2.3.3.1 2.3.3.16	Relative Entropy(HSSP) $< 5; 10; = < 5$ Unused contact energy (total) $> 397; 440; 255; 340$ All Internal contact energy (total) $> 20; 40; 15; < 100$ Ligant Pocket Residue selection Apenas para 1CSH: Curvature in complex < 20
PHAGE LYSOZYME (2LZM ; 2L78 ; 103L ; 7LZM)	3.2.1.17	Curvature in Complex: $< 10; = < 10; = < 19; = < 12$ Pocket in Complex: $\geq 0; > 0$ Unused Energy Total $> 385; \geq 385$ Surface Accessibility: in Isolation < 46 Cross Presence Order@CA = < 0

ALPHA AMYLASES (1Z32; 1SMD; 1XCX; 1MFV; 1HNY; 1KGW)	3.2.1.1	Conservation HSSP: Evolutionary Pressure ≤ 1 Electrostatic Potential: @LHA Range: [-239, -178] Cross Presence Order: Order@LHA ≥ 6
CCP-LIKE (1CCA; 1CCP; 1BEK; 1BEP)	1.11.1.5	Cross Presence Order: Order@LHA ≥ 3 ; 2; Order@CB ≥ 2 Sponge(@CA Radius = 3: Internal(x100) ≥ 84 ; 86; 86; 84 Conservation SH2QS: Evolutionary Pressure ≤ 10 ; 9; 10; 7; Reliability (x100) ≥ 37 ; 37; Relative Entropy 100 ≤ 31 Conservation HSSP: Evolutionary Pressure ≤ 0 Electrostatic Potential: @CA: [4, 12], [1, 16]; @LHA: Min.: -220, Max.: - 102; -96; -87; -80; Average: [-35, 20] Contacts Energy: All Contacts; All Contacts; Internal Energy ≥ 13 ; 18.
SERINE PROTEASES (1MEE; 1TK2; 1LW6; 1SBN; 1TEC)	3.4.21.62 3.4.21.66	Conservation HSSP: Relative Entropy ≤ 4 , 5, 5, 3, 5; Evolutionary Pressure ≤ 0 Unused Contacts: Total ≥ 55 , 69, 52, 45, 69 Surface accessibility in complex < 1 Hydrophobicity: Complex ≤ 0
SERINE PROTEASES (1PPF; 1T32; 1AZZ; 1UTL)	3.4.21.37 3.4.21.20 3.4.21.32 3.4.21.4	Conserv.(Sting): Evol. Pressure: < 10 ; Relative Entropy 100: < 10 ; Relative Entropy: < 10 , 1 Accessibility: Isolation: > 1 ; Relative: $\leq 0, 2$ Electrostatic Potential: @LHA: $< -80, 00$
ARGINASE-LIKE (1D3V; 1RLA; 1PQ3)	3.5.3.1	Unused Contacts Energy ≥ 278 , 348; Contacts ≤ 69 , ≥ 103 , ≤ 85 Conservation SH2QS: Evolutionary Pressure ≤ 0 , 2; Relative Entropy 100 ≤ 2
GLYCOSYL HYDROLASES (6TAA; 1BVN; 1B2Y; 1AMY; 1B1Y)	3.2.1.1 3.2.1.2	Cross Presence Order@CA ≤ 3 ; @CB ≤ 4 ; @LHA ≥ 4 , 5, 3, 6, 6 Electrostatic Potential @Surface: Min.: -300, Max.: -1, 1, -8; @CA: [- 27, 300] Internal Contacts: Charged Repulsive (and) Hydrophobic Scale ;Constant: < -6 Secondary structure: unselect beta sheet
GLYCOSYL HYDROLASES (2BVW; 1OC6; 1QK0, 1QK2; 1CB2)	3.2.1.91	Conservation (HSSP): Relative Entropy < 1 , ≤ 1 ; Reliability: > 41 , ≥ 99 , ≤ 49 , ≤ 49 Geometric – Cross Presence Orde@LHA: > 1 Unused Energy Total: > 90 , 140, ≥ 154 , < 255 Electrostatic Potential: @CA: Min = -39 Max = -24
VERTEBRATE PHOSPHOLIPASE A2 (1GMZ; 2NOT)	3.1.1.4	Conservation HSSP: Relative Entropy ≤ 5 , 6; E. Pressure ≤ 10 ; Secondary Structure (PDB): Helix Hydrofobic Scale: Constant < 1 ; Complex < 1 ; Isolation < 1 Unused Contacts: Contacts ≤ 54 Temp Factor@maximum < 6
VERTEBRATE PHOSPHOLIPASE A2 (1A2A; 1PP2; 1G4I)	3.1.1.4	Unused Contacts: Energy ≥ 175 ; Contacts ≥ 34 Cross Presence Order@CA ≥ 2

<p>HAL-PAL LIKE (1W27)</p>	<p>4.3.1.24</p> <p>Pocket in Complex: Volume(#water volumes) ≥ 1 Conservation SH2QS: Relative Entropy $100 \leq 21$; Evolutionary Pressure ≤ 9 Unused Contacts: Energy Total ≥ 70 Protein Ligand Contacts: Selection Mode:OR; Undefined Relevant Sites: Residue Location:Surface</p>
<p>TIROSINA QUINASES (2J0K)</p>	<p>2.7.10.2</p> <p>Conservation (HSSP): Relative Entropy ≤ 13; Relative Entropy $100 \leq 0$; Pressure Evol. ≤ 2 Structural : Secondary Element PDB: Helix, Turn, Coil Temperature factor: @CA<75; @LHA<78 Unused Energy: Contacts Total ≤ 78; Energy Total ≤ 373 Rotamer: % Rotamer Library ≤ 8</p>
<p>CYCLOPHILIN (PEP-TIDYLPROLYL ISOMERASE) (1CWA)</p>	<p>5.2.1.8</p> <p>Conservation (HSSP): Relative Entropy ≤ 9 Surface Accessibility: In isolation >5; In complex > 5 Pocket in isolation: Volume (# water volumes) ≥ 0 Electrostatic Potential: Average: Minimum=0; Maximum= 300 Conservation - SH2QS: Evolutionary Pressure ≤ 8</p>
<p>PHOSPHOFRUCTOKINASE (3PFK; 2PFK; 4PFK; 6PFK; 1MTO)</p>	<p>2.7.1.11</p> <p>Cross Presence Order: Order@CA $\geq 0, \leq 4$; Order@LHA $\geq 0, 0, 1, 1$ Curvature in Complex: Curvature Value (x100) ≤ 27.0 Sponge: @CA Radius = 3 ; SW = 3 ; Internal(x100) ≤ 85 ; Sliding Window(x100) ≤ 83 Electrostatic Potencial@CA: Exclude neutral residues ; Range:[-18,18]; [-300,25]; [-9,22] @LHA: Exclude neutral residues ; Range:[-171,14]; [-189,18]; [-171,25]; [-174,14] @Surface: Exclude neutral residues ; Range:[-18,2]; [-9,2]; Exclude neutral residues / Range:[-9,6]; Exclude neutral residues / Range:[-12,6]; [-6,6] @Average: Range:[-54,10]; [-42,22] Unused Contacts: Total ≥ 36 / Energy Total ≥ 74; 86; 82 Hydrophobic Scale: Constant < 1 / Complex < 1 / Isolation < 1 Pocket in Complex: Volume(#water volumes) $\geq 0, 13, 0$</p>
<p>CYSTEINE PROTEINASES (1PPN; 1NLN; 1GEC; 1PPO; 1AEC)</p>	<p>3.4.22.2 3.4.22.39 3.4.22.25 3.4.22.30 3.4.22.14</p> <p>Pocket in Isolation: Volume(#water volumes) $\geq 0, 6$ Conservation HSSP: Relative Entropy $\leq 21, 17, \geq 33$ / Evolutionary Pressure $\leq 4, 4, 5$ / Reliability (x100) ≥ 88 / Relative Entropy $100 \leq 5$ Conservation SH2QS: Relative Entropy $\leq 17, 19$ / Evolutionary Pressure ≤ 4 / Reliability (x100) $\leq 98, \geq 98$ Secondary Structure PDB: Sheet / Turn / Coil (Except 3PFK, 2PFK) Residue Property: Type:Hydrophilic (Just for 4PFK, 6PFK)</p>
<p>CYSTEINE PROTEINASES (1PPN; 1NLN; 1GEC; 1PPO; 1AEC)</p>	<p>Relevant Sites: Prosite; Residue Location:Surface Conservation HSSP: Relative Entropy $\leq 7, 9$; Reliability (x100) $\geq 96, 92$; Evol. Pressure ≤ 3 Residue Property: Type:Hydrophilic Dihedral Angles: Most favoured regions Contacts Energy All Contacts: Internal Energy ≤ 13 (Except 1NLN) Cross Link Order: Order@CB $\leq 1, \geq 1$ Rotamer: Percent ≤ 23 Unused Contacts > 121</p>

Apêndice E - Regras para caracterização de resíduos de aminoácidos catalíticos para todas as sub-subclasses estudadas

Nas subseções seguintes são apresentadas em tabelas os conjuntos de regras e o respectivo número de acertos e erros de cada regra no conjunto de indução.

EC. 1.1.1

Regras	Sup./Err.
radialCentrality ≥ 0.9 and pocket_score ≥ 33.98 and density_CA_r=7_WNADist ≥ 3.59 and clo_r=5_l=15_LHA_VD ≥ 2.79 and cpo_r=5_l=20_CA_WNASurf ≥ 0.24	12/0
localCloseness ≥ 0.67 and alphaCentrality ≤ 0.35 and hydroKDI ≤ -0.05 and ced_CA_r=6 ≥ 0.13 and sponge_CA_r=3_VD ≥ 0.46 and hbmwws_energy ≤ 2.6	14/0
radialCentrality ≥ 0.87 and pocket_polarityScore ≥ 13 and charge_att_WNADist ≥ 318.23 and pageRank ≥ 0.74 and charge_att_energy ≤ 170	7/0
distance_CG ≤ 11.63 and uc_hb_ms_VD ≥ 18.62 and clo_r=5_l=20_CB_WNASurf ≤ 0.14 and uc_hb_ss_VD ≥ 11.76 and cpo_r=85_l=30_CA ≥ 5	7/0
distance_CG ≤ 12.91 and density_LHA_r=7_WNADist ≤ 3.06 and cpo_r=35_l=30_CA_VD ≥ 0.34 and dmnc ≤ 0.22 and cpo_r=5_l=15_CA_VD ≤ 0.14	6/0

EC. 2.1.1

Regras	Sup./Err.
(apolarASA ≥ 177.52) and (cpo_85_30_CB_VD ≥ 5.36) and (uc_hbmwm_WNADist ≤ 47.03) and (clo_35_15_LHA_WNASurf ≥ 0.3) and (density_CA_7 ≥ 0.55)	12/0
(apolarASA ≥ 177.52) and (cpo_5_20_LHA_GN ≥ 8.41) and (clo_35_30_CA_GN ≤ 0.32) and (density_LHA_3_WNADist ≥ 3.26) and (clo_85_30_LHA_VD ≥ 1.86) and (charge_repu_energy ≤ 140)	8/0
(alphaCentrality ≤ 0.68) and (meanNeighborDegree ≥ 0.63) and (uc_hbmws_WNADist ≤ 60.59) and (uc_charge_repu_VD ≥ 318.44) and (density_CA_5_WNADist ≥ 3.87)	5/0

EC. 2.3.1

Regras	Sup./Err.
(distance_CG ≤ 10.99) and (total_energy ≥ 29.6) and (score ≥ 30.86) and (cpo_5_20_CB_VD ≥ 0.57) and (cpo_5_15_CA_WNADist ≤ 4.96)	14/0
(distance_CG ≤ 10.99) and (clo_85_20_CA_VD ≥ 3.46) and (flexibility ≥ 0.15) and	8/0

(clo_5_15_CB_VD \geq 3.29) and (sponge_CA_7_WNADist \geq 2.63)

(distance_CG \leq 14.08) and (cpo_35_30_CA_WNADist \geq 1.9) and (polarASA \geq 123.35) and
(clo_5_20_CB_VD \geq 2.7) 13/5

(distance_CG \leq 14.08) and (hb_mm_WNADist \leq 54.12) and (uc_ss_bond_WNADist \geq 22.8) and
(distance_CG \leq 7.71) and (sponge_CA_4_VD \leq 0.93) and (uc_hbmws_VD \leq 19.18) 9/0

EC. 2.4.2

Regras

Sup./Err.

(localCloseness \geq 0.68) and (score \geq 18.45) and (hbmws_VD \leq 0.27) and
(clo_5_30_LHA_WNADist \leq 4.86) and (uc_hbsws_VD \geq 15.6) 9/0

(localCloseness \geq 0.7) and (hydrophobicityScore \geq 9.76) and (uc_aromatic_WNADist \leq 5.25) and
(density_CA_5_VD \geq 1.48) and (total_energy_WNADist \leq 527.1) 10/0

(localCloseness \geq 0.65) and (score \geq 13.39) and (uc_hbmws_WNADist \geq 77.59) and
(ced_LHA_4_VD \leq 0.23) and (cpo_35_30_CA_VD \geq 0.16) and (accIsolation \leq 123.16) 7/0

(score \geq 18.45) and (hb_ss_VD \geq 0.81) and (distance_CG \leq 7.75) and (cpo_35_20_CA_GN \geq
0.22) and (clo_5_15_CA \geq 1) 7/0

(charge_att_energy \geq 90) and (sponge_LHA_6_WNADist \leq 3.12) and (ced_CA_6_VD \leq 0.22)
and (aromatic_VD \geq 1.96) 5/0

EC. 2.5.1

Regras

Sup./Err.

(localCloseness \geq 0.7) and (uc_hb_mm_GN \geq 436.56) and (density_LHA_3 \leq 0.52) and
(clo_85_15_LHA_VD \leq 3.59) and (charge_repu_energy \leq 150) 8/0

(localCloseness \geq 0.7) and (clo_85_20_CB_VD \geq 5.87) and (clo_5_30_CB_WNASurf \geq 0.57)
and (hbms_energy \leq 5.2) 8/0

(localCloseness \geq 0.7) and (uc_aromatic_VD \geq 8.91) and (hb_mm_WNASurf \leq 15.95) and
(accIsolation \geq 23.14) 6/0

(alphaDensity \geq 5.61) and (uc_ss_bond_VD \geq 0.92) and (density_CA_5_WNADist \leq 3.9) and
(clo_85_15_LHA_WNASurf \leq 1.86) 6/0

(distance_CG \leq 14.15) and (sponge_LHA_5 \geq 0.56) and (sponge_LHA_4_WNADist \leq 3) and
(hbsws_GN \leq 1.11) 5/0

EC. 2.7.1

Regras

Sup./Err.

(localCloseness \geq 0.8) and (score \geq 30.3) and (cpo_85_20_LHA_WNASurf \leq 1.3) and 12/0

(sponge_LHA_5_WNADist \geq 2.42)	
(alphaCentrality \leq 0.67) and (radialCentrality \geq 0.95) and (hbmws_GN \geq 24.71) and (hbmwm_GN \leq 12.99) and (density_CA_5 \geq 0.63)	8/0
(proportionPolarAtoms \geq 49.52) and (eccentricity \leq 0.17) and (uc_hbmwm_VD \geq 22.31) and (cpo_5_15_CB_WNASurf \geq 0.33) and (hbmws_WNASurf \leq 1.09)	9/0
(randomWalkBetweenness \geq 0.82) and (efficiency \leq 0.64) and (ced_CA_3_WNADist \leq 1.03)	6/0

EC. 3.1.1

Regras	Sup./Err.
(radialCentrality \geq 0.83) and (cpo_35_30_LHA_VD \geq 1.04) and (baryCenter \geq 0.9) and (ced_CA_4_VD \geq 0.47) and (ced_CA_6 \leq 0.28)	20/0
(radialCentrality \geq 0.83) and (cpo_35_30_CA_VD \geq 1.08) and (clo_35_15_CB_WNADist \geq 1.12) and (sponge_LHA_5 \geq 0.48) and (cpo_85_15_LHA_VD \geq 5.45)	16/0
(cpo_85_30_LHA_VD \geq 8.21) and (drugScore \geq 0.45) and (hydroRI \leq -0.07)	18/7

EC. 3.1.3

Regras	Sup./Err.
(localCloseness \geq 0.82) and (ced_LHA_5_VD \geq 0.43) and (uc_charge_att_WNADist \geq 547.72) and (charge_repu_WNASurf \leq 50.23) and (ced_CA_3 \leq 0.55)	19/0
(meanNeighborDegree \geq 0.77) and (volumeScore \geq 4) and (uc_charge_att \geq 240) and (hbmws_energy \leq 0)	7/0
(meanNeighborDegree \geq 0.64) and (uc_hydrophobic_WNADist \leq 34.93) and (hydrophobic_energy \leq 4.2) and (hbsws_GN \leq 3.9)	7/0
(baryCenter \geq 0.8) and (diceSimilarity \geq 0.77) and (sponge_LHA_4 \geq 0.6) and (clo_5_15_CB_WNASurf \leq 1) and (clo_5_20_CB \leq 1)	7/0
(alphaCentrality \leq 0.73) and (sponge_LHA_4_WNADist \leq 2.57) and (hydrophobic_GN \leq 140.34) and (cpo_85_15_LHA_VD \geq 3.89) and (clo_35_15_CA_VD \geq 0.03)	8/0
(ced_LHA_5 \geq 0.33) and (aromatic_VD \geq 1.5) and (hb_mm_WNADist \leq 43.52) and (clo_5_30_LHA_VD \leq 3.69) and (hydrophobic_energy \geq 5.4)	7/0

EC. 3.2.1

Regras	Sup./Err.
(localCloseness \geq 0.84) and (totalASA \geq 237.85) and (uc_hbmwws_VD \leq 22.87) and (distance_CG \leq 10.61) and (uc_charge_repu_GN \geq 4650.87) and (cpo_5_20_LHA_WNADist \leq	29/0

5.78)	
(localCloseness \geq 0.79) and (localCloseness \geq 0.91) and (polarASA \geq 145.09) and (uc_aromatic \leq 1.5) and (uc_hbmwvm_VD \geq 17.36) and (density_CA_7_WNADist \geq 3.73)	19/0
(localCloseness \geq 0.79) and (uc_aromatic \leq 0) and (polarASA \geq 94.82) and (sponge_CA_6_WNADist \leq 2.13) and (hbmwvm_GN \leq 23.97) and (density_LHA_3_VD \geq 1.11)	17/0
(localCloseness \geq 0.73) and (localCloseness \geq 0.91) and (hydroKDI \leq -0.11) and (ced_CA_5_WNADist \geq 0.83) and (cpo_85_15_LHA_GN \leq 98.59) and (hbsws_WNASurf \leq 0.43)	11/0
(localCloseness \geq 0.65) and (cpo_85_30_CA_VD \geq 10.93) and (hydroKDI \leq -0.06) and (volume \geq 865.04) and (clo_35_20_CA_GN \geq 4.22) and (clo_5_20_CB_WNADist \leq 6.76) and (hbmm_energy \leq 10.4)	11/0

EC. 3.4.21

Regras	Sup./Err.
(cpo_35_30_LHA_VD \geq 1.09) and (hb_ss_VD \geq 2.84) and (uc_hbmwvm_VD \geq 20.16) and (hb_ss_WNASurf \leq 1.53)	18/0
(distance_CG \leq 13.67) and (ced_LHA_3_WNADist \geq 1.09) and (apolarAlphaProportion \geq 0.22) and (hb_ss_WNADist \geq 4.4) and (charge_repu_WNASurf \leq 61.52)	8/0
(distance_CG \leq 13.67) and (uc_charge_att_WNADist \geq 810.18) and (cpo_5_15_CB_VD \geq 0.35) and (hbmwvm_GN \geq 10.06)	9/0

EC. 4.2.1

Regras	Sup./Err.
(baryCenter \geq 0.84) and (cpo_85_30_CB_VD \geq 8.95) and (cpo_85_20_CB_VD \leq 5.88) and (ced_CA_4 \leq 0.16) and (clo_5_15_CB \geq 1)	15/0
(localCloseness \geq 0.82) and (score \geq 20.76) and (clo_85_30_LHA_WNADist \leq 5.66) and (clo_5_20_CA_WNADist \geq 2.03) and (sponge_CA_3_VD \geq 0.47) and (sponge_CA_7_WNADist \geq 2.89)	10/0
(ced_LHA_5 \geq 0.15) and (charge_repu_WNASurf \leq 20.18) and (cpo_5_20_CB_GN \leq 5.82) and (sponge_LHA_6 \geq 0.46) and (accIsolation \leq 55.67)	7/0
(ced_CA_5_VD \geq 0.3) and (distance_CG \leq 10.64) and (sponge_LHA_3_VD \geq 1.14) and (hb_ms_GN \leq 33.36) and (distance_NTerminal \geq 11.49)	6/0

EC. 4.1.1

Regras	Sup./Err.
(localCloseness \geq 0.8) and (cpo_85_30_CA_VD \geq 9.12) and (density_LHA_7 \leq 0.53) and (aggregateConstraints \geq 0.08) and (hydroRI \leq -0.07)	13/0
(localCloseness \geq 0.9) and (clo_85_15_LHA_VD \geq 7.79) and (charge_repu_energy \geq 130)	9/0
(betweenness \geq 0.06) and (score \geq 17.88) and (clo_85_30_CB_VD \leq 0.66) and (density_CA_7_WNADist \geq 3.57) and (cpo_85_20_CB_VD \leq 0.9) and (hydrophobic_energy \geq 4.2)	6/0
(betweenness \geq 0.2) and (sponge_LHA_4_WNADist \leq 2.86) and (uc_hydrophobic \geq 7.8) and (hb_mm_WNADist \leq 42.27) and (sponge_LHA_3 \geq 0.52) and (hydroRI \leq 0.25)	7/0

Apêndice F - Testes estatísticos para comparação entre classificadores

A seguir são apresentados os resultados do emprego do teste estatístico de Welch para comparação entre o desempenho dos diversos classificadores avaliados. Em vermelho são destacados os casos onde houve diferenças significativas entre os classificadores comparados (rejeição da hipótese nula do teste de Welch).

i. RIPPER vs. C4.5

SUB-SUBCLASSE	WELCH-T	VALOR-P
1.1.1	0.321	0.752
3.1.1	0.652	0.523
2.3.1	0.812	0.428
4.2.1	-0.309	0.761
2.4.2	0.587	0.565
4.1.1	2.713	0.014
2.1.1	0.888	0.387
2.5.1	-0.993	0.334
3.1.3	-1.422	0.173
2.7.1	1.367	0.190
3.4.21	1.695	0.107
3.2.1	1.949	0.069

ii. RUS, ROS, RWOS vs. RIPPER sem pré-processamento

SUB-SUBCLASSE	RIPPER				C4.5	
	Amostragem	Proporção	Welch-t	Valor-p	Welch-t	Valor-p
1.1.1	RUS	1:1	3.31762	0.00722	3.05422	0.013254
1.1.1	ROS	1:1	-0.30729	0.762233	-0.31684	0.755706
1.1.1	RWOS	1:1	1.009006	0.330104	-0.22784	0.823117
1.1.1	RUS	1:2	1.903211	0.082754	2.350509	0.038722
1.1.1	ROS	1:2	-0.57377	0.573698	0.188091	0.853106
1.1.1	RWOS	1:2	0.394804	0.699127	-0.28634	0.778095
1.1.1	RUS	1:6	0.711743	0.487048	0.939869	0.361834

1.1.1	ROS	1:6	0.565139	0.579146	-0.37594	0.711424
1.1.1	RWOS	1:6	-0.40574	0.689718	0.317738	0.75448
1.1.1	RUS	1:10	0.48264	0.635558	0.403491	0.693841
1.1.1	ROS	1:10	-0.64113	0.53271	-0.67679	0.507297
1.1.1	RWOS	1:10	-0.6912	0.498414	0.521791	0.608417
1.1.1	RUS	1:25	-1.13188	0.27346	0.270249	0.790304
1.1.1	ROS	1:25	0.418192	0.680821	0.324066	0.749845
1.1.1	RWOS	1:25	-0.7276	0.476395	-0.08467	0.933456
1.1.1	RUS	1:50	-0.56622	0.578603	-0.56337	0.5802
1.1.1	ROS	1:50	-1.19794	0.25013	0.412224	0.685054
1.1.1	RWOS	1:50	-0.54397	0.594069	1.325915	0.203782
3.1.1	RUS	1:1	6.011584	5.88E-05	4.825431	0.000722
3.1.1	ROS	1:1	0.652537	0.522432	2.49615	0.026251
3.1.1	RWOS	1:1	4.184897	0.001454	2.135965	0.047367
3.1.1	RUS	1:2	5.635156	0.000114	3.729418	0.003248
3.1.1	ROS	1:2	1.326115	0.204461	0.362451	0.721411
3.1.1	RWOS	1:2	1.887842	0.07558	0.971211	0.344322
3.1.1	RUS	1:6	3.690843	0.002616	2.56876	0.024565
3.1.1	ROS	1:6	-0.7785	0.44639	0.733906	0.473273
3.1.1	RWOS	1:6	0.241473	0.812336	1.029419	0.317425
3.1.1	RUS	1:10	0.104692	0.917815	0.943547	0.359582
3.1.1	ROS	1:10	0.786518	0.442194	0.985725	0.338844
3.1.1	RWOS	1:10	1.063395	0.301997	0.746845	0.465183
3.1.1	RUS	1:25	-0.69324	0.497037	0.028544	0.97762
3.1.1	ROS	1:25	0.01128	0.991126	0.110536	0.91328
3.1.1	RWOS	1:25	1.101259	0.286059	-0.57041	0.575785
3.1.1	RUS	1:50	0.203125	0.841316	0.004359	0.996573
3.1.1	ROS	1:50	0.037001	0.970894	-0.42517	0.67595
3.1.1	RWOS	1:50	3.902241	0.001066	1.511976	0.14825
2.3.1	RUS	1:1	2.692028	0.024185	2.423427	0.036544
2.3.1	ROS	1:1	0.308498	0.761957	-0.51637	0.611925
2.3.1	RWOS	1:1	2.202162	0.053603	0.216998	0.830773
2.3.1	RUS	1:2	2.310845	0.045307	1.673519	0.125009
2.3.1	ROS	1:2	-0.42798	0.673823	-0.21516	0.832192
2.3.1	RWOS	1:2	1.167454	0.268113	-0.58048	0.569175
2.3.1	RUS	1:6	1.24044	0.240348	0.505689	0.62289
2.3.1	ROS	1:6	-0.22062	0.827988	-0.74768	0.464702
2.3.1	RWOS	1:6	0.261484	0.798163	-1.33805	0.197726
2.3.1	RUS	1:10	1.104484	0.293585	0.479861	0.63931
2.3.1	ROS	1:10	0.573424	0.57428	0.163797	0.871877
2.3.1	RWOS	1:10	0.547447	0.5926	-0.64167	0.529229
2.3.1	RUS	1:25	0.23636	0.816032	0.037756	0.97031
2.3.1	ROS	1:25	0.497022	0.62542	-1.94146	0.068399
2.3.1	RWOS	1:25	0.831159	0.417644	-0.40697	0.68893
2.3.1	RUS	1:50	0.92682	0.368304	-0.28058	0.78273
2.3.1	ROS	1:50	1.078668	0.299588	-0.85542	0.404574
2.3.1	RWOS	1:50	0.499001	0.625413	0.102021	0.919934

4.2.1	RUS	1:1	2.912457	0.014045	3.800995	0.003726
4.2.1	ROS	1:1	0.523785	0.606888	1.405409	0.178357
4.2.1	RWOS	1:1	0.719726	0.480973	0.877508	0.392525
4.2.1	RUS	1:2	2.305073	0.041415	3.218348	0.009136
4.2.1	ROS	1:2	0.677713	0.507709	1.776506	0.094093
4.2.1	RWOS	1:2	0.691788	0.499377	0.816976	0.425187
4.2.1	RUS	1:6	1.11318	0.287567	1.831164	0.091567
4.2.1	ROS	1:6	-0.12872	0.899038	0.398173	0.69519
4.2.1	RWOS	1:6	0.019406	0.984743	1.811571	0.088824
4.2.1	RUS	1:10	0.46896	0.646495	1.022585	0.321576
4.2.1	ROS	1:10	0.620411	0.543344	1.455629	0.165437
4.2.1	RWOS	1:10	0.717324	0.48461	2.068854	0.053806
4.2.1	RUS	1:25	0.202212	0.842285	0.071535	0.943763
4.2.1	ROS	1:25	0.108328	0.91495	0.838691	0.413651
4.2.1	RWOS	1:25	-0.58598	0.565555	2.019522	0.058729
4.2.1	RUS	1:50	-0.16618	0.870683	1.524579	0.145497
4.2.1	ROS	1:50	0.601997	0.55472	2.631175	0.019278
4.2.1	RWOS	1:50	0.055598	0.956373	0.348852	0.731598
2.4.2	RUS	1:1	1.202662	0.255905	0.698019	0.500755
2.4.2	ROS	1:1	-1.40556	0.179092	-1.50491	0.149697
2.4.2	RWOS	1:1	-0.34424	0.736028	-1.2605	0.223605
2.4.2	RUS	1:2	0.313109	0.760286	0.629805	0.541833
2.4.2	ROS	1:2	-0.70977	0.487094	-1.82105	0.087468
2.4.2	RWOS	1:2	-0.72644	0.477105	-1.54786	0.139693
2.4.2	RUS	1:6	-0.1202	0.906095	-0.95497	0.358503
2.4.2	ROS	1:6	-0.75616	0.459536	-0.38407	0.705425
2.4.2	RWOS	1:6	-1.44868	0.164949	-2.2033	0.040951
2.4.2	RUS	1:10	-1.52965	0.143503	-1.06156	0.308391
2.4.2	ROS	1:10	-1.38362	0.183397	-1.02045	0.322899
2.4.2	RWOS	1:10	-0.67087	0.510916	-0.59447	0.560576
2.4.2	RUS	1:25	-2.08101	0.052111	-1.27948	0.223402
2.4.2	ROS	1:25	-0.99736	0.331845	0.433399	0.669947
2.4.2	RWOS	1:25	-0.80304	0.43469	0.444755	0.6619
2.4.2	RUS	1:50	-0.72668	0.476837	-1.53188	0.142939
2.4.2	ROS	1:50	-0.93007	0.365006	1.060945	0.304571
2.4.2	RWOS	1:50	-0.69417	0.496939	0.174599	0.863347
4.1.1	RUS	1:1	4.171368	0.00165	1.047128	0.32162
4.1.1	ROS	1:1	1.037412	0.314006	-1.71366	0.103778
4.1.1	RWOS	1:1	2.76647	0.013279	-1.66966	0.119033
4.1.1	RUS	1:2	3.665794	0.003797	0.530471	0.607609
4.1.1	ROS	1:2	1.264073	0.222339	-2.35881	0.031901
4.1.1	RWOS	1:2	1.186103	0.251424	-1.88238	0.076865
4.1.1	RUS	1:6	1.924769	0.079185	-0.31815	0.756334
4.1.1	ROS	1:6	1.653334	0.115711	-2.42924	0.025996
4.1.1	RWOS	1:6	2.541902	0.021455	-2.42441	0.026255
4.1.1	RUS	1:10	1.518916	0.154883	-0.75503	0.461433
4.1.1	ROS	1:10	0.671906	0.510365	-0.52329	0.607239

4.1.1	RWOS	1:10	0.524691	0.606211	-0.856	0.403346
4.1.1	RUS	1:25	1.925786	0.070202	-1.81034	0.087242
4.1.1	ROS	1:25	1.479429	0.15843	-1.97519	0.063795
4.1.1	RWOS	1:25	1.52678	0.144201	0.037046	0.970921
4.1.1	RUS	1:50	0.562064	0.581685	-1.84557	0.083567
4.1.1	ROS	1:50	2.122687	0.047911	-0.2485	0.806561
4.1.1	RWOS	1:50	1.867164	0.07859	-0.55229	0.58794
2.1.1	RUS	1:1	1.014867	0.33294	-0.03839	0.970082
2.1.1	ROS	1:1	-0.10329	0.918979	-1.55531	0.139686
2.1.1	RWOS	1:1	0.307223	0.763796	-0.99693	0.332028
2.1.1	RUS	1:2	-0.07359	0.942448	-0.10229	0.920404
2.1.1	ROS	1:2	-0.6743	0.508702	-1.62557	0.122747
2.1.1	RWOS	1:2	-0.8388	0.413673	-0.90323	0.37834
2.1.1	RUS	1:6	-1.59703	0.127868	-1.10609	0.284933
2.1.1	ROS	1:6	-1.06866	0.299591	-1.20096	0.246243
2.1.1	RWOS	1:6	-0.74393	0.466529	-0.91752	0.371004
2.1.1	RUS	1:10	-1.32853	0.200737	-1.18545	0.258795
2.1.1	ROS	1:10	-2.14968	0.046519	-0.06654	0.947691
2.1.1	RWOS	1:10	0.128205	0.899523	0.00326	0.997435
2.1.1	RUS	1:25	-1.27961	0.217558	-1.53702	0.14196
2.1.1	ROS	1:25	-1.10596	0.283523	-0.2996	0.768001
2.1.1	RWOS	1:25	-1.21063	0.241865	-0.91985	0.36983
2.1.1	RUS	1:50	-1.94265	0.073253	-0.56113	0.581644
2.1.1	ROS	1:50	-1.70521	0.105784	-1.22606	0.236471
2.1.1	RWOS	1:50	-0.31323	0.758495	-0.49064	0.629613
2.5.1	RUS	1:1	1.196318	0.261244	3.144362	0.011668
2.5.1	ROS	1:1	-0.82611	0.420198	0.90658	0.377139
2.5.1	RWOS	1:1	-0.56643	0.581513	0.880228	0.391047
2.5.1	RUS	1:2	0.311073	0.762057	2.723795	0.022002
2.5.1	ROS	1:2	-1.06535	0.302351	0.972573	0.34389
2.5.1	RWOS	1:2	-0.61153	0.548604	0.1395	0.890664
2.5.1	RUS	1:6	-0.83428	0.415321	0.932905	0.365101
2.5.1	ROS	1:6	-0.92769	0.366481	1.147377	0.266295
2.5.1	RWOS	1:6	-1.31162	0.208548	1.003433	0.329022
2.5.1	RUS	1:10	-1.13324	0.275371	0.646161	0.526821
2.5.1	ROS	1:10	-2.16252	0.04599	1.02703	0.321
2.5.1	RWOS	1:10	-0.43197	0.671095	1.52416	0.146629
2.5.1	RUS	1:25	-0.05978	0.952994	1.026781	0.318618
2.5.1	ROS	1:25	-1.25238	0.227436	0.742471	0.467391
2.5.1	RWOS	1:25	-0.84737	0.408254	1.123562	0.276426
2.5.1	RUS	1:50	-1.07836	0.295707	-0.66776	0.514042
2.5.1	ROS	1:50	-0.38014	0.708315	1.278853	0.217203
2.5.1	RWOS	1:50	-0.304	0.764612	-0.11321	0.911126
3.1.3	RUS	1:1	3.056047	0.011428	6.216774	8.29E-05
3.1.3	ROS	1:1	0.163901	0.871637	1.285568	0.214928
3.1.3	RWOS	1:1	0.391818	0.699793	2.204944	0.042566
3.1.3	RUS	1:2	1.761159	0.098002	4.361534	0.000491

3.1.3	ROS	1:2	-0.85262	0.412187	1.254573	0.225682
3.1.3	RWOS	1:2	-0.11887	0.906917	2.682081	0.015501
3.1.3	RUS	1:6	1.403029	0.185695	2.833162	0.011736
3.1.3	ROS	1:6	1.939129	0.068325	1.945987	0.067447
3.1.3	RWOS	1:6	-0.14851	0.883704	1.516135	0.147246
3.1.3	RUS	1:10	0.216105	0.831704	2.070184	0.05347
3.1.3	ROS	1:10	1.516288	0.151894	1.496213	0.152548
3.1.3	RWOS	1:10	0.807219	0.430341	2.500536	0.022286
3.1.3	RUS	1:25	0.353004	0.729508	1.908456	0.072736
3.1.3	ROS	1:25	0.445084	0.661588	2.751455	0.0154
3.1.3	RWOS	1:25	-0.50296	0.621104	2.261858	0.036457
3.1.3	RUS	1:50	-0.41957	0.679781	1.336882	0.198891
3.1.3	ROS	1:50	1.444347	0.173251	0.533757	0.600698
3.1.3	RWOS	1:50	4.883626	0.000328	5.998265	1.18E-05
2.7.1	RUS	1:1	6.0667	3.44E-05	3.47391	0.006284
2.7.1	ROS	1:1	-0.26279	0.796685	0.096583	0.924281
2.7.1	RWOS	1:1	2.328507	0.033352	0.330073	0.745265
2.7.1	RUS	1:2	4.781442	0.000416	2.979669	0.01211
2.7.1	ROS	1:2	1.011684	0.327042	-0.4252	0.67588
2.7.1	RWOS	1:2	1.78211	0.092674	-0.84043	0.412104
2.7.1	RUS	1:6	3.928915	0.001359	1.256994	0.22814
2.7.1	ROS	1:6	1.177457	0.2571	0.942416	0.358585
2.7.1	RWOS	1:6	1.38874	0.182615	0.234706	0.817346
2.7.1	RUS	1:10	2.258522	0.036687	0.733199	0.474509
2.7.1	ROS	1:10	0.579255	0.572665	0.105357	0.917308
2.7.1	RWOS	1:10	0.780946	0.446747	-0.15162	0.881205
2.7.1	RUS	1:25	1.095763	0.287669	1.423449	0.173056
2.7.1	ROS	1:25	0.549085	0.590874	0.884258	0.38822
2.7.1	RWOS	1:25	0.906561	0.3814	2.270986	0.036738
2.7.1	RUS	1:50	-0.22561	0.824353	0.614556	0.546672
2.7.1	ROS	1:50	0.798777	0.435515	1.396911	0.180342
2.7.1	RWOS	1:50	1.86442	0.079473	-0.04778	0.962423
3.4.21	RUS	1:1	4.686582	0.000736	2.428719	0.034794
3.4.21	ROS	1:1	0.353709	0.728021	0.143243	0.887752
3.4.21	RWOS	1:1	2.625671	0.017451	-0.49045	0.629742
3.4.21	RUS	1:2	4.603991	0.000834	1.769149	0.105706
3.4.21	ROS	1:2	0.625808	0.539297	-0.71414	0.484812
3.4.21	RWOS	1:2	2.053537	0.054924	-0.39023	0.700961
3.4.21	RUS	1:6	3.398934	0.004173	0.873894	0.395618
3.4.21	ROS	1:6	-0.13839	0.89156	-0.99921	0.330948
3.4.21	RWOS	1:6	0.273148	0.788391	-1.67678	0.111105
3.4.21	RUS	1:10	1.022069	0.320295	0.161893	0.873582
3.4.21	ROS	1:10	0.474276	0.641006	-0.36169	0.723684
3.4.21	RWOS	1:10	0.579511	0.569967	0.046159	0.963692
3.4.21	RUS	1:25	-0.16533	0.870914	-0.40002	0.69605
3.4.21	ROS	1:25	-0.03532	0.972263	-0.99237	0.33537
3.4.21	RWOS	1:25	1.001335	0.332044	0.397281	0.695928

3.4.21	RUS	1:50	-0.60864	0.550757	-1.99664	0.063261
3.4.21	ROS	1:50	0.209043	0.83697	-0.0601	0.952798
3.4.21	RWOS	1:50	2.13516	0.04862	0.770568	0.451394
3.2.1	RUS	1:1	16.03982	9.97E-09	10.12183	1.58E-06
3.2.1	ROS	1:1	0.591265	0.562238	0.99758	0.333035
3.2.1	RWOS	1:1	6.774767	3.19E-06	1.940752	0.070695
3.2.1	RUS	1:2	13.07594	5.09E-09	7.63027	3.16E-06
3.2.1	ROS	1:2	1.288861	0.217012	1.278894	0.219595
3.2.1	RWOS	1:2	4.539259	0.000327	1.783573	0.092944
3.2.1	RUS	1:6	9.831123	2.71E-08	4.672689	0.000472
3.2.1	ROS	1:6	0.379518	0.708744	1.364349	0.191055
3.2.1	RWOS	1:6	2.917593	0.00932	1.948937	0.067173
3.2.1	RUS	1:10	4.95977	0.000102	4.458315	0.000474
3.2.1	ROS	1:10	0.282477	0.782456	0.923003	0.368409
3.2.1	RWOS	1:10	1.679764	0.114625	0.506084	0.620414
3.2.1	RUS	1:25	2.704649	0.014513	0.253304	0.802963
3.2.1	ROS	1:25	0.214705	0.833262	0.466762	0.647813
3.2.1	RWOS	1:25	1.266396	0.221593	-0.27204	0.788982
3.2.1	RUS	1:50	0.223608	0.825964	-0.04983	0.96086
3.2.1	ROS	1:50	2.477613	0.024307	1.169583	0.258732
3.2.1	RWOS	1:50	1.201239	0.24549	0.112871	0.911561

iii. *EA-TSS vs. RIPPER sem pré-processamento*

SUB-SUBCLASSE	WELCH-T	VALOR-P
1.1.1	1.1924	0.2504
2.1.1	1.9275	0.0699
2.3.1	-0.0925	0.9279
2.4.2	3.029	0.007823
2.5.1	0.7726	0.4528
2.7.1	-0.4281	0.6737
3.1.1	-0.0866	0.9322
3.1.3	0.1638	0.8728
3.2.1	0.7927	0.4384
3.4.21	1.133	0.2722
4.1.1	-0.1517	0.8811
4.2.1	1.2584	0.2313

iv. *RUSBoost+RIPPER vs. RIPPER sem pré-processamento*

SUB-SUBCLASSE	RIPPER			C4.5	
	Proporção	Welch-t	Valor-p	Welch-t	Valor-p
1.1.1	1:1	2.892225	0.01438	1.954253	0.07771
1.1.1	1:2	1.137211	0.271909	0.766241	0.459156
1.1.1	1:6	-1.01952	0.323803	-0.34915	0.732349
1.1.1	1:10	-0.50867	0.617256	-1.64326	0.117867

1.1.1	1:25	-0.74623	0.465709	-0.51538	0.612567
1.1.1	1:50	-0.45589	0.653949	-0.62546	0.539561
3.1.1	1:1	3.563068	0.002341	3.338574	0.007268
3.1.1	1:2	1.896044	0.07462	1.106148	0.290176
3.1.1	1:6	-0.08838	0.930553	-0.85636	0.404133
3.1.1	1:10	-0.63872	0.531652	-0.91927	0.37396
3.1.1	1:25	-1.18072	0.253173	-0.48944	0.630439
3.1.1	1:50	-0.73284	0.473143	0.157091	0.876968
2.3.1	1:1	1.864947	0.093499	1.145124	0.275865
2.3.1	1:2	1.193191	0.259571	0.061765	0.951886
2.3.1	1:6	-0.37148	0.715416	-1.72383	0.101984
2.3.1	1:10	-1.34395	0.195668	-2.08412	0.052326
2.3.1	1:25	0.177472	0.861322	-2.51308	0.022692
2.3.1	1:50	1.828822	0.084811	-2.42116	0.027902
4.2.1	1:1	2.092075	0.056293	2.593446	0.02706
4.2.1	1:2	1.179872	0.254107	1.056893	0.312739
4.2.1	1:6	0.406797	0.68976	0.257038	0.800162
4.2.1	1:10	-1.4146	0.175542	0.094628	0.925658
4.2.1	1:25	-0.33215	0.743686	1.682425	0.110165
4.2.1	1:50	0.26663	0.792801	1.339182	0.197333
2.4.2	1:1	0.283168	0.782633	0.003738	0.997084
2.4.2	1:2	1.1108	0.282655	-1.61759	0.125037
2.4.2	1:6	-0.77209	0.450083	-2.35265	0.030257
2.4.2	1:10	-1.15418	0.263842	-3.26069	0.004452
2.4.2	1:25	-1.04315	0.310831	-1.25887	0.227123
2.4.2	1:50	-1.1622	0.263895	-1.18876	0.252232
4.1.1	1:1	4.448715	0.00091	0.283647	0.782479
4.1.1	1:2	2.557412	0.021731	-1.22398	0.242994
4.1.1	1:6	-0.03289	0.974147	-2.65564	0.016928
4.1.1	1:10	-0.36974	0.716253	-2.50052	0.0225
4.1.1	1:25	0.336884	0.741023	-1.82603	0.084517
4.1.1	1:50	2.309748	0.033025	-1.95618	0.066652
2.1.1	1:1	0.196176	0.847344	-1.41206	0.187152
2.1.1	1:2	-1.30404	0.209561	-2.0662	0.05351
2.1.1	1:6	-1.06021	0.304076	-1.28091	0.217162
2.1.1	1:10	-2.55277	0.021477	-2.29836	0.033888
2.1.1	1:25	-1.06374	0.306168	-4.09487	0.000727
2.1.1	1:50	1.586517	0.132795	-0.27338	0.787782
2.5.1	1:1	0.162029	0.87412	1.193784	0.255521
2.5.1	1:2	-1.12718	0.275049	-0.05991	0.953214
2.5.1	1:6	-1.08422	0.293053	-1.04612	0.309359
2.5.1	1:10	-1.53929	0.14546	-2.4892	0.022866
2.5.1	1:25	-1.70368	0.108414	-0.26897	0.7912
2.5.1	1:50	-0.91523	0.374433	0.092063	0.927855
3.1.3	1:1	1.645813	0.124819	3.857377	0.002446

3.1.3	1:2	0.553757	0.586687	2.737964	0.016259
3.1.3	1:6	-1.13958	0.269498	-0.33355	0.742658
3.1.3	1:10	-0.0517	0.95936	1.026767	0.318899
3.1.3	1:25	-0.50376	0.620561	0.427694	0.674903
3.1.3	1:50	0.789659	0.440868	2.013605	0.061625
2.7.1	1:1	5.290211	5.89E-05	2.099649	0.056946
2.7.1	1:2	2.335758	0.031394	1.576969	0.13702
2.7.1	1:6	2.095962	0.052644	-0.08897	0.930095
2.7.1	1:10	-0.47829	0.639459	-1.14127	0.268835
2.7.1	1:25	0.047023	0.963172	0.220625	0.827892
2.7.1	1:50	0.625474	0.543189	-0.31237	0.758347
3.4.21	1:1	3.726859	0.002209	0.987648	0.340975
3.4.21	1:2	1.105514	0.283523	-0.23199	0.820407
3.4.21	1:6	0.020702	0.983713	-1.3121	0.206295
3.4.21	1:10	-0.8619	0.403405	-1.36614	0.191879
3.4.21	1:25	0.296096	0.771263	-1.25204	0.228256
3.4.21	1:50	0.870759	0.397474	-0.58667	0.56618
3.2.1	1:1	13.59891	1.39E-08	7.559453	5.87E-06
3.2.1	1:2	7.161272	2.72E-06	5.83108	0.000104
3.2.1	1:6	4.591909	0.000293	1.544906	0.142891
3.2.1	1:10	3.043856	0.00699	0.246345	0.808236
3.2.1	1:25	-1.30158	0.209658	-1.74764	0.099634
3.2.1	1:50	-0.94788	0.362032	-1.62282	0.123671

v. *SMOTE+FS vs. RIPPER sem pré-processamento*

SUB-SUBCLASSE	CLASSIFICADOR	TAXA SUBREAMOSTRAGEM	WELCH-T	VALOR-P
1.1.1	RIPPER	-	-0.79024	0.440149
1.1.1	SMOTEBagging	-	-2.45716	0.024803
1.1.1	FS+SMOTE	100	-1.25773	0.228054
1.1.1	SMOTEBoost	100	-1.31575	0.204906
1.1.1	FS+SMOTE	200	-1.44482	0.165695
1.1.1	SMOTEBoost	200	-2.4516	0.025241
1.1.1	FS+SMOTE	300	-1.85566	0.084741
1.1.1	SMOTEBoost	300	-1.89786	0.076794
1.1.1	FS+SMOTE	400	-1.01589	0.32384
1.1.1	SMOTEBoost	400	-1.40668	0.178959
1.1.1	FS+SMOTE	500	-1.14408	0.26758
1.1.1	SMOTEBoost	500	-2.72037	0.014149
1.1.1	FS+SMOTE	600	-0.48467	0.633759
1.1.1	SMOTEBoost	600	-1.30245	0.212975
1.1.1	FS+SMOTE	700	0.136063	0.893518
1.1.1	SMOTEBoost	700	-2.37465	0.030487
1.1.1	FS+SMOTE	800	-1.05381	0.306658
1.1.1	SMOTEBoost	800	-2.13055	0.047288

1.1.1	FS+SMOTE	900	-0.66907	0.513038
1.1.1	SMOTEBoost	900	-0.90396	0.378609
1.1.1	FS+SMOTE	1000	-0.72586	0.478558
1.1.1	SMOTEBoost	1000	-1.15287	0.265707
3.1.1	RIPPER	-	-3.08414	0.007007
3.1.1	SMOTEBagging	-	-3.44575	0.002884
3.1.1	FS+SMOTE	100	-2.49274	0.023514
3.1.1	SMOTEBoost	100	-2.25755	0.037657
3.1.1	FS+SMOTE	200	-1.92254	0.070906
3.1.1	SMOTEBoost	200	-1.65989	0.114286
3.1.1	FS+SMOTE	300	-3.17957	0.005319
3.1.1	SMOTEBoost	300	-2.15195	0.048007
3.1.1	FS+SMOTE	400	-3.22694	0.005094
3.1.1	SMOTEBoost	400	-1.86621	0.07854
3.1.1	FS+SMOTE	500	-1.485	0.15531
3.1.1	SMOTEBoost	500	-3.38976	0.003285
3.1.1	FS+SMOTE	600	-1.82887	0.084119
3.1.1	SMOTEBoost	600	-3.32679	0.003754
3.1.1	FS+SMOTE	700	-1.6993	0.106664
3.1.1	SMOTEBoost	700	-1.85526	0.080962
3.1.1	FS+SMOTE	800	-1.77015	0.094076
3.1.1	SMOTEBoost	800	-3.18381	0.005586
3.1.1	FS+SMOTE	900	-1.43711	0.167844
3.1.1	SMOTEBoost	900	-2.59563	0.018299
3.1.1	FS+SMOTE	1000	-1.65887	0.114609
3.1.1	SMOTEBoost	1000	-1.90665	0.075657
2.3.1	RIPPER	-	0.424026	0.677181
2.3.1	SMOTEBagging	-	-0.95615	0.352594
2.3.1	FS+SMOTE	100	-0.33999	0.737867
2.3.1	SMOTEBoost	100	-0.46559	0.65058
2.3.1	FS+SMOTE	200	-0.17683	0.861717
2.3.1	SMOTEBoost	200	-0.39846	0.695479
2.3.1	FS+SMOTE	300	0.250101	0.805419
2.3.1	SMOTEBoost	300	0.298369	0.768979
2.3.1	FS+SMOTE	400	0.011081	0.991282
2.3.1	SMOTEBoost	400	0.245731	0.808824
2.3.1	FS+SMOTE	500	-0.20796	0.838311
2.3.1	SMOTEBoost	500	0.300496	0.767887
2.3.1	FS+SMOTE	600	0.039307	0.96931
2.3.1	SMOTEBoost	600	0.790562	0.441292
2.3.1	FS+SMOTE	700	0.230849	0.820409
2.3.1	SMOTEBoost	700	-0.33206	0.74368
2.3.1	FS+SMOTE	800	0.324019	0.749991
2.3.1	SMOTEBoost	800	-0.15906	0.875531
2.3.1	FS+SMOTE	900	0.795627	0.441443

2.3.1	SMOTEBoost	900	0.516702	0.615088
2.3.1	FS+SMOTE	1000	0.967863	0.350596
2.3.1	SMOTEBoost	1000	0.306471	0.762935
4.2.1	RIPPER	-	0.772697	0.450474
4.2.1	SMOTEBagging	-	-0.88644	0.388853
4.2.1	FS+SMOTE	100	1.100189	0.285878
4.2.1	SMOTEBoost	100	0.770024	0.451411
4.2.1	FS+SMOTE	200	0.363761	0.720542
4.2.1	SMOTEBoost	200	0.242868	0.810855
4.2.1	FS+SMOTE	300	-0.07016	0.944842
4.2.1	SMOTEBoost	300	-0.08264	0.935178
4.2.1	FS+SMOTE	400	-0.04618	0.963826
4.2.1	SMOTEBoost	400	-0.02455	0.980685
4.2.1	FS+SMOTE	500	-0.77731	0.447089
4.2.1	SMOTEBoost	500	-0.3914	0.700105
4.2.1	FS+SMOTE	600	1.145518	0.267606
4.2.1	SMOTEBoost	600	0.528826	0.603391
4.2.1	FS+SMOTE	700	0.262851	0.795798
4.2.1	SMOTEBoost	700	0.123427	0.903495
4.2.1	FS+SMOTE	800	-0.12806	0.899628
4.2.1	SMOTEBoost	800	0.437061	0.667401
4.2.1	FS+SMOTE	900	0.041195	0.967619
4.2.1	SMOTEBoost	900	-0.59383	0.561684
4.2.1	FS+SMOTE	1000	-0.11981	0.905976
4.2.1	SMOTEBoost	1000	-0.44826	0.659397
2.4.2	RIPPER	-	-1.14953	0.268539
2.4.2	SMOTEBagging	-	-0.20094	0.843034
2.4.2	FS+SMOTE	100	-0.0329	0.974137
2.4.2	SMOTEBoost	100	1.456181	0.165317
2.4.2	FS+SMOTE	200	-0.16583	0.87017
2.4.2	SMOTEBoost	200	1.506786	0.150596
2.4.2	FS+SMOTE	300	0.621419	0.542138
2.4.2	SMOTEBoost	300	1.390589	0.183476
2.4.2	FS+SMOTE	400	-0.38112	0.708153
2.4.2	SMOTEBoost	400	0.982476	0.338942
2.4.2	FS+SMOTE	500	0.242581	0.811101
2.4.2	SMOTEBoost	500	0.694673	0.49632
2.4.2	FS+SMOTE	600	0.604648	0.553082
2.4.2	SMOTEBoost	600	0.750796	0.46365
2.4.2	FS+SMOTE	700	1.113093	0.285654
2.4.2	SMOTEBoost	700	1.85594	0.083236
2.4.2	FS+SMOTE	800	-0.65259	0.524691
2.4.2	SMOTEBoost	800	0.338201	0.739135
2.4.2	FS+SMOTE	900	0.946174	0.359067
2.4.2	SMOTEBoost	900	0.025663	0.979815

2.4.2	FS+SMOTE	1000	0.227185	0.82289
2.4.2	SMOTEBoost	1000	-0.3241	0.749924
4.1.1	RIPPER	-	-1.0007	0.337859
4.1.1	SMOTEBagging	-	-0.83019	0.421874
4.1.1	FS+SMOTE	100	-1.28069	0.226222
4.1.1	SMOTEBoost	100	-1.9204	0.079066
4.1.1	FS+SMOTE	200	-1.40405	0.184298
4.1.1	SMOTEBoost	200	-1.05437	0.310274
4.1.1	FS+SMOTE	300	-1.68336	0.116309
4.1.1	SMOTEBoost	300	-2.55728	0.021411
4.1.1	FS+SMOTE	400	-1.40238	0.183707
4.1.1	SMOTEBoost	400	-1.6411	0.124283
4.1.1	FS+SMOTE	500	-3.36022	0.003589
4.1.1	SMOTEBoost	500	-1.80557	0.090267
4.1.1	FS+SMOTE	600	-0.77595	0.451297
4.1.1	SMOTEBoost	600	-1.22957	0.241619
4.1.1	FS+SMOTE	700	-0.92857	0.369433
4.1.1	SMOTEBoost	700	-2.00063	0.067565
4.1.1	FS+SMOTE	800	-1.19868	0.250411
4.1.1	SMOTEBoost	800	-1.31197	0.214643
4.1.1	FS+SMOTE	900	-2.01072	0.06055
4.1.1	SMOTEBoost	900	-0.96701	0.351188
4.1.1	FS+SMOTE	1000	-1.0151	0.329495
4.1.1	SMOTEBoost	1000	-0.36993	0.717476
2.1.1	RIPPER	-	-1.15334	0.268252
2.1.1	SMOTEBagging	-	-1.00728	0.331656
2.1.1	FS+SMOTE	100	-0.47772	0.640066
2.1.1	SMOTEBoost	100	-1.28035	0.221496
2.1.1	FS+SMOTE	200	0.29681	0.770176
2.1.1	SMOTEBoost	200	-0.40348	0.691527
2.1.1	FS+SMOTE	300	-0.24797	0.807313
2.1.1	SMOTEBoost	300	-1.04107	0.31616
2.1.1	FS+SMOTE	400	-0.71023	0.488116
2.1.1	SMOTEBoost	400	-0.72061	0.484068
2.1.1	FS+SMOTE	500	-1.65112	0.120816
2.1.1	SMOTEBoost	500	-1.70889	0.112124
2.1.1	FS+SMOTE	600	-1.4525	0.169227
2.1.1	SMOTEBoost	600	-0.00751	0.994093
2.1.1	FS+SMOTE	700	-0.85031	0.41062
2.1.1	SMOTEBoost	700	-0.9251	0.369554
2.1.1	FS+SMOTE	800	-1.53999	0.14346
2.1.1	SMOTEBoost	800	0.310017	0.760293
2.1.1	FS+SMOTE	900	-0.54421	0.594333
2.1.1	SMOTEBoost	900	-1.45888	0.165815
2.1.1	FS+SMOTE	1000	-0.45573	0.65465

2.1.1	SMOTEBoost	1000	-1.29795	0.213861
2.5.1	RIPPER	-	0.387464	0.70297
2.5.1	SMOTEBagging	-	-1.33142	0.199683
2.5.1	FS+SMOTE	100	-0.40164	0.69273
2.5.1	SMOTEBoost	100	-0.54563	0.593094
2.5.1	FS+SMOTE	200	-2.10509	0.049685
2.5.1	SMOTEBoost	200	-0.0301	0.976315
2.5.1	FS+SMOTE	300	-0.89152	0.385016
2.5.1	SMOTEBoost	300	0.659059	0.51833
2.5.1	FS+SMOTE	400	0.267422	0.792584
2.5.1	SMOTEBoost	400	-0.34633	0.73313
2.5.1	FS+SMOTE	500	-1.32428	0.203027
2.5.1	SMOTEBoost	500	-0.08467	0.933504
2.5.1	FS+SMOTE	600	-0.42274	0.67768
2.5.1	SMOTEBoost	600	0.936332	0.364939
2.5.1	FS+SMOTE	700	0.024918	0.98042
2.5.1	SMOTEBoost	700	-0.78409	0.443653
2.5.1	FS+SMOTE	800	-1.95078	0.067689
2.5.1	SMOTEBoost	800	-0.03693	0.97095
2.5.1	FS+SMOTE	900	-1.86975	0.077876
2.5.1	SMOTEBoost	900	-0.37853	0.709507
2.5.1	FS+SMOTE	1000	-0.40089	0.69353
2.5.1	SMOTEBoost	1000	-1.04607	0.309374
3.1.3	RIPPER	-	-0.54438	0.592974
3.1.3	SMOTEBagging	-	0.783568	0.443483
3.1.3	FS+SMOTE	100	-1.5736	0.133712
3.1.3	SMOTEBoost	100	0.701472	0.492279
3.1.3	FS+SMOTE	200	-0.42586	0.675792
3.1.3	SMOTEBoost	200	0.638932	0.531109
3.1.3	FS+SMOTE	300	-0.92708	0.367327
3.1.3	SMOTEBoost	300	0.423665	0.676989
3.1.3	FS+SMOTE	400	0.462873	0.649319
3.1.3	SMOTEBoost	400	0.819478	0.423324
3.1.3	FS+SMOTE	500	-0.24363	0.810455
3.1.3	SMOTEBoost	500	0.266062	0.794487
3.1.3	FS+SMOTE	600	0.034036	0.97328
3.1.3	SMOTEBoost	600	1.672722	0.111675
3.1.3	FS+SMOTE	700	-0.37205	0.714355
3.1.3	SMOTEBoost	700	0.137561	0.892214
3.1.3	FS+SMOTE	800	0.344957	0.734155
3.1.3	SMOTEBoost	800	1.479414	0.15822
3.1.3	FS+SMOTE	900	-0.61811	0.547779
3.1.3	SMOTEBoost	900	-0.02198	0.982711
3.1.3	FS+SMOTE	1000	-0.30198	0.766133
3.1.3	SMOTEBoost	1000	0.15258	0.880451

2.7.1	RIPPER	-	-0.35688	0.727544
2.7.1	SMOTEBagging	-	-1.09295	0.293441
2.7.1	FS+SMOTE	100	-1.60866	0.127977
2.7.1	SMOTEBoost	100	-0.46396	0.649435
2.7.1	FS+SMOTE	200	-1.70154	0.110543
2.7.1	SMOTEBoost	200	-0.10733	0.91607
2.7.1	FS+SMOTE	300	-0.53518	0.600191
2.7.1	SMOTEBoost	300	0.407881	0.689792
2.7.1	FS+SMOTE	400	-0.45435	0.656292
2.7.1	SMOTEBoost	400	0.569563	0.577634
2.7.1	FS+SMOTE	500	-1.45997	0.163
2.7.1	SMOTEBoost	500	0.356811	0.727103
2.7.1	FS+SMOTE	600	-0.67068	0.514455
2.7.1	SMOTEBoost	600	0.68848	0.501613
2.7.1	FS+SMOTE	700	-0.5782	0.570953
2.7.1	SMOTEBoost	700	-0.61924	0.546469
2.7.1	FS+SMOTE	800	0.079667	0.937558
2.7.1	SMOTEBoost	800	5.67E-16	1
2.7.1	FS+SMOTE	900	0.681305	0.505781
2.7.1	SMOTEBoost	900	0.321986	0.751446
2.7.1	FS+SMOTE	1000	0.654373	0.521298
2.7.1	SMOTEBoost	1000	-0.09222	0.927687
3.4.21	RIPPER	-	-2.36331	0.032641
3.4.21	SMOTEBagging	-	-4.30045	0.000431
3.4.21	FS+SMOTE	100	-3.78122	0.001768
3.4.21	SMOTEBoost	100	-2.27909	0.036185
3.4.21	FS+SMOTE	200	-2.32722	0.034395
3.4.21	SMOTEBoost	200	-2.40581	0.031667
3.4.21	FS+SMOTE	300	-1.53127	0.145935
3.4.21	SMOTEBoost	300	-2.11512	0.053759
3.4.21	FS+SMOTE	400	-1.77896	0.097978
3.4.21	SMOTEBoost	400	-3.68851	0.001807
3.4.21	FS+SMOTE	500	-3.16585	0.005786
3.4.21	SMOTEBoost	500	-3.31674	0.004279
3.4.21	FS+SMOTE	600	-1.90667	0.076073
3.4.21	SMOTEBoost	600	-4.66016	0.000199
3.4.21	FS+SMOTE	700	-1.83734	0.085806
3.4.21	SMOTEBoost	700	-2.48076	0.02495
3.4.21	FS+SMOTE	800	-2.48953	0.023272
3.4.21	SMOTEBoost	800	-2.17769	0.049578
3.4.21	FS+SMOTE	900	-1.79405	0.092557
3.4.21	SMOTEBoost	900	-2.59454	0.020549
3.4.21	FS+SMOTE	1000	-2.39917	0.028225
3.4.21	SMOTEBoost	1000	-3.14449	0.006013
3.2.1	RIPPER	-	-3.27934	0.006441

3.2.1	SMOTEBagging	-	-6.0774	2.24E-05
3.2.1	FS+SMOTE	100	-2.72819	0.0174
3.2.1	SMOTEBoost	100	-3.44341	0.003625
3.2.1	FS+SMOTE	200	-2.83407	0.011563
3.2.1	SMOTEBoost	200	-2.86741	0.01345
3.2.1	FS+SMOTE	300	-2.58357	0.019805
3.2.1	SMOTEBoost	300	-3.06803	0.009309
3.2.1	FS+SMOTE	400	-2.61735	0.017456
3.2.1	SMOTEBoost	400	-4.59188	0.000364
3.2.1	FS+SMOTE	500	-2.10266	0.054413
3.2.1	SMOTEBoost	500	-3.18907	0.007567
3.2.1	FS+SMOTE	600	-3.14872	0.006661
3.2.1	SMOTEBoost	600	-3.89309	0.001617
3.2.1	FS+SMOTE	700	-2.48558	0.024021
3.2.1	SMOTEBoost	700	-3.28084	0.006073
3.2.1	FS+SMOTE	800	-2.28233	0.036506
3.2.1	SMOTEBoost	800	-4.20183	0.000754
3.2.1	FS+SMOTE	900	-1.9121	0.076256
3.2.1	SMOTEBoost	900	-7.28474	1.25E-06
3.2.1	FS+SMOTE	1000	-2.58615	0.018648
3.2.1	SMOTEBoost	1000	-3.12031	0.006866

Apêndice G - Desempenhos dos classificadores após pré-processamento (subamostragem e sobreamostragem aleatórias)

A seguir são apresentados os desempenhos dos classificadores treinados após o pré-processamento dos dados para todos as sub-subclasses e tipos de amostragem.

i. Subamostragem aleatória (RUS)

Na Tabela 7-10, é apresentado o mesmo resultado para os classificadores treinados utilizando-se o algoritmo para construção de árvores de decisão C4.5.

Para todas as proporções entre classes consideradas, observou-se que os maiores valores de *F-measure* são obtidos quanto a subamostragem é menor, ou seja, para proporções de 1:25 e 1:50. Isto porque, para outras proporções, há uma grande remoção de amostras negativas e uma queda da precisão dos classificadores, mesmo que seguido de aumento da sensibilidade. Dessa forma, a drástica redução do número de amostras da classe majoritária ocasiona a remoção de amostras importantes para o aprendizado, sendo que reduções em menor escala resultam em melhor desempenho.

Tabela 7-9 na página nº 120.

ii. Sobreamostragem aleatória (ROS)

RIPPER								
Proporção entre classes			1:1	1:2	1:6	1:10	1:25	1:50
EC.1.1.1	S	0.20(0.10)	0.26(0.16)	0.17(0.11)	0.25(0.12)	0.19(0.16)	0.29(0.10)	
	P	0.20(0.15)	0.19(0.14)	0.16(0.11)	0.18(0.03)	0.15(0.09)	0.20(0.08)	
	F	0.19(0.12)	0.21(0.13)	0.15(0.09)	0.20(0.05)	0.16(0.12)	0.22(0.06)	
EC.2.1.1	S	0.09(0.13)	0.13(0.12)	0.14(0.10)	0.18(0.11)	0.17(0.19)	0.24(0.21)	
	P	0.07(0.10)	0.09(0.08)	0.11(0.09)	0.17(0.15)	0.11(0.10)	0.11(0.08)	
	F	0.08(0.11)	0.10(0.08)	0.12(0.09)	0.17(0.11)	0.12(0.09)	0.14(0.10)	
EC.2.3.1	S	0.16(0.08)	0.24(0.15)	0.23(0.14)	0.20(0.15)	0.18(0.13)	0.16(0.10)	
	P	0.18(0.11)	0.17(0.09)	0.17(0.09)	0.14(0.09)	0.14(0.11)	0.11(0.06)	
	F	0.16(0.08)	0.20(0.11)	0.19(0.10)	0.15(0.09)	0.15(0.10)	0.13(0.07)	
EC.2.4.2	S	0.19(0.13)	0.17(0.15)	0.15(0.10)	0.21(0.16)	0.25(0.14)	0.16(0.10)	
	P	0.15(0.07)	0.11(0.07)	0.13(0.08)	0.14(0.09)	0.12(0.10)	0.14(0.10)	
	F	0.15(0.07)	0.13(0.08)	0.13(0.08)	0.16(0.10)	0.14(0.10)	0.14(0.08)	
EC.2.5.1	S	0.17(0.18)	0.18(0.10)	0.18(0.16)	0.26(0.14)	0.19(0.11)	0.12(0.11)	
	P	0.11(0.11)	0.11(0.06)	0.13(0.13)	0.14(0.07)	0.12(0.08)	0.12(0.15)	
	F	0.13(0.13)	0.13(0.07)	0.14(0.13)	0.17(0.07)	0.14(0.08)	0.10(0.09)	
EC.2.7.1	S	0.25(0.16)	0.22(0.13)	0.23(0.14)	0.26(0.23)	0.25(0.19)	0.25(0.13)	
	P	0.26(0.16)	0.16(0.10)	0.16(0.14)	0.18(0.16)	0.21(0.14)	0.18(0.10)	
	F	0.24(0.14)	0.19(0.11)	0.18(0.12)	0.20(0.16)	0.20(0.11)	0.20(0.09)	
EC.3.1.1	S	0.29(0.14)	0.30(0.11)	0.37(0.12)	0.32(0.16)	0.34(0.12)	0.33(0.15)	
	P	0.25(0.10)	0.24(0.07)	0.30(0.09)	0.22(0.09)	0.27(0.11)	0.28(0.10)	
	F	0.26(0.11)	0.24(0.06)	0.32(0.09)	0.25(0.11)	0.29(0.10)	0.29(0.10)	
EC.3.1.3	S	0.24(0.17)	0.33(0.10)	0.17(0.15)	0.21(0.12)	0.26(0.14)	0.20(0.11)	
	P	0.26(0.12)	0.28(0.10)	0.12(0.15)	0.21(0.13)	0.22(0.15)	0.23(0.12)	
	F	0.23(0.14)	0.28(0.04)	0.13(0.13)	0.17(0.07)	0.22(0.12)	0.18(0.06)	
EC.3.2.1	S	0.37(0.10)	0.38(0.09)	0.42(0.10)	0.41(0.15)	0.46(0.13)	0.38(0.09)	
	P	0.36(0.06)	0.32(0.11)	0.35(0.09)	0.34(0.12)	0.32(0.10)	0.26(0.07)	
	F	0.36(0.07)	0.34(0.08)	0.37(0.05)	0.37(0.13)	0.37(0.10)	0.31(0.07)	
EC.3.4.21	S	0.23(0.14)	0.30(0.19)	0.35(0.23)	0.27(0.13)	0.35(0.20)	0.29(0.11)	
	P	0.24(0.18)	0.19(0.11)	0.22(0.11)	0.20(0.09)	0.21(0.15)	0.21(0.07)	
	F	0.22(0.14)	0.21(0.10)	0.25(0.13)	0.22(0.10)	0.24(0.15)	0.23(0.08)	
EC.4.1.1	S	0.14(0.06)	0.14(0.11)	0.11(0.10)	0.17(0.10)	0.16(0.09)	0.12(0.12)	
	P	0.12(0.08)	0.11(0.09)	0.09(0.08)	0.13(0.09)	0.10(0.05)	0.06(0.06)	
	F	0.13(0.07)	0.11(0.09)	0.10(0.09)	0.14(0.07)	0.11(0.06)	0.08(0.08)	
EC.4.2.1	S	0.15(0.13)	0.19(0.12)	0.18(0.08)	0.17(0.11)	0.22(0.12)	0.16(0.12)	
	P	0.15(0.11)	0.13(0.07)	0.18(0.11)	0.13(0.07)	0.14(0.10)	0.13(0.08)	
	F	0.14(0.11)	0.14(0.07)	0.17(0.09)	0.14(0.08)	0.16(0.09)	0.14(0.10)	
Médias	S	0.21(0.13)	0.24(0.13)	0.23(0.13)	0.24(0.14)	0.25(0.14)	0.23(0.12)	

	P	0.20(0.11)	0.18(0.09)	0.18(0.11)	0.18(0.10)	0.18(0.11)	0.17(0.09)
	F	0.19(0.11)	0.19(0.09)	0.19(0.10)	0.20(0.10)	0.19(0.10)	0.18(0.08)
C4.5							
Proporção entre classes		1:01	1:02	1:06	1:10	1:25	1:50
EC.1.1.1	S	0.25(0.16)	0.18(0.09)	0.20(0.14)	0.21(0.13)	0.15(0.11)	0.14(0.11)
	P	0.16(0.07)	0.15(0.09)	0.20(0.19)	0.20(0.11)	0.16(0.07)	0.17(0.17)
	F	0.18(0.07)	0.15(0.08)	0.18(0.12)	0.19(0.10)	0.15(0.09)	0.14(0.12)
EC.2.1.1	S	0.16(0.17)	0.13(0.10)	0.12(0.10)	0.06(0.09)	0.09(0.12)	0.11(0.13)
	P	0.08(0.09)	0.09(0.09)	0.09(0.11)	0.05(0.08)	0.05(0.06)	0.09(0.11)
	F	0.11(0.10)	0.10(0.09)	0.09(0.09)	0.05(0.06)	0.05(0.06)	0.09(0.08)
EC.2.3.1	S	0.19(0.12)	0.18(0.12)	0.19(0.10)	0.16(0.12)	0.26(0.09)	0.21(0.17)
	P	0.15(0.09)	0.13(0.08)	0.17(0.10)	0.11(0.07)	0.20(0.10)	0.17(0.14)
	F	0.16(0.09)	0.14(0.08)	0.16(0.08)	0.13(0.08)	0.22(0.09)	0.18(0.14)
EC.2.4.2	S	0.16(0.09)	0.23(0.17)	0.13(0.12)	0.20(0.18)	0.12(0.16)	0.07(0.09)
	P	0.13(0.12)	0.14(0.11)	0.09(0.09)	0.11(0.10)	0.06(0.06)	0.07(0.11)
	F	0.13(0.08)	0.16(0.12)	0.09(0.08)	0.12(0.12)	0.06(0.07)	0.04(0.05)
EC.2.5.1	S	0.13(0.12)	0.15(0.13)	0.10(0.11)	0.16(0.12)	0.14(0.14)	0.08(0.10)
	P	0.10(0.10)	0.09(0.11)	0.08(0.07)	0.09(0.06)	0.09(0.08)	0.10(0.11)
	F	0.10(0.07)	0.09(0.08)	0.09(0.08)	0.10(0.05)	0.10(0.09)	0.08(0.09)
EC.2.7.1	S	0.22(0.20)	0.25(0.18)	0.17(0.13)	0.20(0.13)	0.15(0.08)	0.13(0.16)
	P	0.16(0.15)	0.18(0.10)	0.12(0.11)	0.20(0.20)	0.15(0.15)	0.10(0.10)
	F	0.17(0.14)	0.20(0.11)	0.13(0.11)	0.17(0.12)	0.14(0.09)	0.11(0.12)
EC.3.1.1	S	0.15(0.06)	0.26(0.08)	0.23(0.08)	0.26(0.11)	0.27(0.09)	0.30(0.12)
	P	0.17(0.06)	0.26(0.12)	0.23(0.10)	0.20(0.09)	0.27(0.15)	0.28(0.09)
	F	0.16(0.06)	0.24(0.09)	0.23(0.08)	0.22(0.08)	0.26(0.09)	0.28(0.09)
EC.3.1.3	S	0.31(0.14)	0.27(0.12)	0.24(0.12)	0.25(0.13)	0.27(0.17)	0.30(0.16)
	P	0.23(0.10)	0.28(0.13)	0.27(0.22)	0.33(0.29)	0.20(0.06)	0.34(0.24)
	F	0.26(0.11)	0.26(0.11)	0.23(0.10)	0.24(0.13)	0.21(0.06)	0.29(0.15)
EC.3.2.1	S	0.38(0.13)	0.32(0.11)	0.30(0.10)	0.33(0.09)	0.33(0.11)	0.31(0.11)
	P	0.24(0.11)	0.25(0.13)	0.26(0.11)	0.28(0.10)	0.31(0.17)	0.28(0.11)
	F	0.29(0.10)	0.27(0.11)	0.27(0.10)	0.29(0.08)	0.30(0.13)	0.28(0.10)
EC.3.4.21	S	0.24(0.23)	0.24(0.20)	0.24(0.12)	0.22(0.09)	0.25(0.09)	0.15(0.09)
	P	0.12(0.10)	0.19(0.13)	0.22(0.16)	0.15(0.04)	0.19(0.08)	0.26(0.27)
	F	0.15(0.14)	0.20(0.14)	0.21(0.11)	0.17(0.05)	0.20(0.08)	0.16(0.08)
EC.4.1.1	S	0.16(0.10)	0.24(0.18)	0.19(0.11)	0.11(0.11)	0.15(0.10)	0.10(0.11)
	P	0.11(0.09)	0.15(0.11)	0.13(0.06)	0.06(0.06)	0.13(0.10)	0.06(0.07)
	F	0.13(0.09)	0.18(0.13)	0.15(0.08)	0.08(0.08)	0.14(0.09)	0.07(0.08)
EC.4.2.1	S	0.13(0.08)	0.13(0.10)	0.23(0.21)	0.14(0.08)	0.18(0.11)	0.08(0.06)
	P	0.12(0.08)	0.10(0.07)	0.14(0.08)	0.13(0.10)	0.13(0.06)	0.08(0.07)

	F	0.12(0.08)	0.11(0.08)	0.16(0.10)	0.12(0.07)	0.15(0.08)	0.08(0.06)
Médias	S	0.21(0.13)	0.22(0.13)	0.20(0.12)	0.19(0.12)	0.20(0.11)	0.17(0.12)
	P	0.15(0.10)	0.17(0.11)	0.17(0.12)	0.16(0.11)	0.16(0.10)	0.17(0.13)
	F	0.16(0.09)	0.17(0.11)	0.17(0.09)	0.16(0.09)	0.17(0.09)	0.15(0.10)

iii. *Sobreamostragem via atribuição de pesos às amostras positivas (RWOS)*

RIPPER							
Proporção entre classes		1:01	1:02	1:06	1:10	1:25	1:50
EC.1.1.1	S	0.62(0.30)	0.53(0.23)	0.41(0.25)	0.41(0.24)	0.32(0.21)	0.25(0.18)
	P	0.09(0.04)	0.10(0.04)	0.14(0.08)	0.16(0.08)	0.16(0.08)	0.19(0.14)
	F	0.14(0.06)	0.16(0.05)	0.20(0.10)	0.21(0.09)	0.21(0.12)	0.21(0.15)
EC.2.1.1	S	0.50(0.30)	0.34(0.25)	0.22(0.15)	0.14(0.14)	0.18(0.15)	0.09(0.15)
	P	0.04(0.02)	0.06(0.04)	0.09(0.12)	0.05(0.05)	0.18(0.30)	0.09(0.13)
	F	0.07(0.04)	0.10(0.06)	0.10(0.08)	0.07(0.06)	0.12(0.09)	0.09(0.14)
EC.2.3.1	S	0.65(0.20)	0.50(0.25)	0.39(0.16)	0.28(0.19)	0.18(0.14)	0.12(0.07)
	P	0.05(0.01)	0.07(0.03)	0.11(0.03)	0.12(0.07)	0.11(0.08)	0.25(0.14)
	F	0.08(0.02)	0.12(0.04)	0.16(0.05)	0.15(0.07)	0.13(0.10)	0.15(0.07)
EC.2.4.2	S	0.60(0.18)	0.34(0.20)	0.33(0.22)	0.26(0.18)	0.18(0.18)	0.15(0.15)
	P	0.07(0.03)	0.11(0.11)	0.13(0.12)	0.09(0.08)	0.13(0.15)	0.15(0.13)
	F	0.11(0.05)	0.13(0.08)	0.17(0.11)	0.13(0.11)	0.15(0.16)	0.13(0.12)
EC.2.5.1	S	0.53(0.22)	0.33(0.18)	0.31(0.19)	0.21(0.20)	0.15(0.14)	0.08(0.09)
	P	0.06(0.03)	0.07(0.06)	0.10(0.05)	0.07(0.06)	0.14(0.14)	0.14(0.14)
	F	0.11(0.04)	0.11(0.09)	0.14(0.07)	0.11(0.08)	0.13(0.12)	0.10(0.10)
EC.2.7.1	S	0.45(0.27)	0.43(0.26)	0.38(0.21)	0.38(0.24)	0.27(0.24)	0.19(0.12)
	P	0.10(0.12)	0.11(0.07)	0.12(0.08)	0.15(0.10)	0.14(0.13)	0.15(0.08)
	F	0.13(0.11)	0.16(0.09)	0.18(0.09)	0.19(0.11)	0.18(0.16)	0.16(0.09)
EC.3.1.1	S	0.68(0.08)	0.45(0.19)	0.43(0.12)	0.37(0.14)	0.30(0.18)	0.08(0.05)
	P	0.09(0.02)	0.15(0.09)	0.22(0.06)	0.21(0.10)	0.23(0.12)	0.57(0.43)
	F	0.16(0.03)	0.21(0.11)	0.28(0.06)	0.25(0.08)	0.24(0.12)	0.14(0.08)
EC.3.1.3	S	0.50(0.28)	0.66(0.14)	0.48(0.20)	0.30(0.21)	0.35(0.15)	0.01(0.04)
	P	0.15(0.11)	0.16(0.07)	0.19(0.10)	0.17(0.08)	0.24(0.14)	0.07(0.21)
	F	0.22(0.13)	0.25(0.09)	0.25(0.10)	0.20(0.11)	0.27(0.14)	0.02(0.06)
EC.3.2.1	S	0.77(0.16)	0.68(0.15)	0.60(0.14)	0.59(0.09)	0.49(0.11)	0.41(0.08)
	P	0.12(0.05)	0.17(0.08)	0.21(0.05)	0.23(0.10)	0.28(0.05)	0.31(0.07)
	F	0.20(0.07)	0.25(0.07)	0.31(0.06)	0.32(0.09)	0.35(0.05)	0.35(0.06)
EC.3.4.21	S	0.56(0.32)	0.40(0.24)	0.36(0.07)	0.30(0.13)	0.24(0.19)	0.10(0.14)
	P	0.09(0.08)	0.11(0.07)	0.17(0.06)	0.17(0.06)	0.16(0.14)	0.20(0.25)
	F	0.13(0.09)	0.15(0.10)	0.23(0.07)	0.22(0.08)	0.18(0.16)	0.12(0.15)

EC.4.1.1	S	0.40(0.24)	0.37(0.24)	0.19(0.21)	0.27(0.17)	0.16(0.14)	0.10(0.09)
	P	0.04(0.04)	0.08(0.05)	0.06(0.04)	0.11(0.08)	0.08(0.07)	0.10(0.07)
	F	0.07(0.06)	0.12(0.07)	0.08(0.06)	0.14(0.08)	0.10(0.08)	0.10(0.07)
EC.4.2.1	S	0.47(0.19)	0.43(0.21)	0.35(0.22)	0.20(0.12)	0.31(0.18)	0.15(0.17)
	P	0.09(0.08)	0.08(0.04)	0.11(0.05)	0.11(0.05)	0.15(0.08)	0.26(0.20)
	F	0.13(0.10)	0.14(0.07)	0.17(0.08)	0.14(0.06)	0.19(0.08)	0.16(0.16)
Médias	S	0.56(0.23)	0.46(0.21)	0.37(0.18)	0.31(0.17)	0.26(0.17)	0.14(0.11)
	P	0.08(0.05)	0.11(0.06)	0.14(0.07)	0.14(0.08)	0.17(0.12)	0.21(0.17)
	F	0.13(0.07)	0.16(0.08)	0.19(0.08)	0.18(0.09)	0.19(0.12)	0.14(0.10)
C4.5							
Proporção entre classes		1:01	1:02	1:06	1:10	1:25	1:50
EC.1.1.1	S	0.31(0.12)	0.27(0.17)	0.18(0.08)	0.15(0.11)	0.18(0.12)	0.12(0.09)
	P	0.12(0.05)	0.15(0.08)	0.14(0.12)	0.13(0.08)	0.17(0.13)	0.10(0.07)
	F	0.17(0.06)	0.17(0.09)	0.15(0.09)	0.14(0.09)	0.17(0.11)	0.11(0.07)
EC.2.1.1	S	0.15(0.12)	0.14(0.14)	0.12(0.12)	0.06(0.07)	0.10(0.10)	0.07(0.07)
	P	0.05(0.06)	0.05(0.05)	0.06(0.06)	0.04(0.07)	0.07(0.08)	0.06(0.08)
	F	0.08(0.07)	0.07(0.07)	0.07(0.07)	0.04(0.07)	0.07(0.07)	0.06(0.07)
EC.2.3.1	S	0.19(0.13)	0.21(0.06)	0.23(0.15)	0.19(0.10)	0.21(0.17)	0.12(0.09)
	P	0.09(0.06)	0.16(0.16)	0.18(0.09)	0.15(0.09)	0.15(0.10)	0.21(0.20)
	F	0.12(0.08)	0.16(0.08)	0.20(0.11)	0.16(0.09)	0.15(0.12)	0.13(0.08)
EC.2.4.2	S	0.26(0.19)	0.21(0.15)	0.23(0.14)	0.16(0.18)	0.08(0.13)	0.08(0.08)
	P	0.09(0.08)	0.12(0.10)	0.13(0.07)	0.09(0.10)	0.05(0.07)	0.07(0.10)
	F	0.12(0.08)	0.14(0.10)	0.16(0.09)	0.10(0.12)	0.06(0.09)	0.07(0.08)
EC.2.5.1	S	0.21(0.15)	0.25(0.18)	0.18(0.17)	0.10(0.10)	0.12(0.10)	0.11(0.08)
	P	0.07(0.05)	0.10(0.06)	0.07(0.07)	0.07(0.06)	0.08(0.07)	0.18(0.16)
	F	0.10(0.07)	0.12(0.07)	0.09(0.08)	0.08(0.06)	0.09(0.07)	0.13(0.09)
EC.2.7.1	S	0.28(0.15)	0.30(0.14)	0.23(0.16)	0.20(0.09)	0.12(0.11)	0.16(0.09)
	P	0.12(0.10)	0.17(0.06)	0.14(0.12)	0.19(0.11)	0.09(0.08)	0.23(0.17)
	F	0.16(0.11)	0.21(0.08)	0.16(0.13)	0.18(0.08)	0.09(0.07)	0.18(0.11)
EC.3.1.1	S	0.21(0.12)	0.30(0.16)	0.21(0.11)	0.28(0.14)	0.33(0.06)	0.12(0.09)
	P	0.14(0.08)	0.17(0.09)	0.23(0.11)	0.20(0.07)	0.28(0.13)	0.38(0.30)
	F	0.16(0.09)	0.21(0.11)	0.21(0.09)	0.23(0.09)	0.29(0.09)	0.18(0.13)
EC.3.1.3	S	0.33(0.09)	0.29(0.15)	0.26(0.16)	0.24(0.14)	0.21(0.14)	0.04(0.08)
	P	0.19(0.07)	0.17(0.08)	0.29(0.18)	0.18(0.10)	0.22(0.15)	0.07(0.16)
	F	0.23(0.07)	0.20(0.09)	0.24(0.12)	0.20(0.11)	0.20(0.12)	0.05(0.10)
EC.3.2.1	S	0.44(0.07)	0.38(0.13)	0.31(0.05)	0.35(0.10)	0.33(0.10)	0.33(0.10)
	P	0.20(0.05)	0.20(0.09)	0.24(0.10)	0.29(0.14)	0.35(0.11)	0.32(0.12)
	F	0.27(0.05)	0.26(0.10)	0.26(0.08)	0.30(0.12)	0.33(0.10)	0.32(0.11)
EC.3.4.21	S	0.26(0.17)	0.29(0.22)	0.26(0.09)	0.16(0.11)	0.16(0.16)	0.08(0.11)

	P	0.19(0.19)	0.15(0.11)	0.27(0.17)	0.17(0.13)	0.13(0.13)	0.27(0.34)
	F	0.18(0.11)	0.18(0.12)	0.24(0.10)	0.16(0.11)	0.14(0.13)	0.12(0.14)
EC.4.1.1	S	0.19(0.08)	0.27(0.18)	0.23(0.13)	0.11(0.11)	0.08(0.09)	0.09(0.08)
	P	0.08(0.04)	0.10(0.08)	0.13(0.09)	0.08(0.09)	0.05(0.05)	0.08(0.08)
	F	0.11(0.04)	0.14(0.11)	0.16(0.10)	0.09(0.09)	0.06(0.06)	0.08(0.07)
EC.4.2.1	S	0.27(0.16)	0.26(0.15)	0.16(0.13)	0.13(0.15)	0.11(0.12)	0.15(0.14)
	P	0.10(0.06)	0.10(0.06)	0.08(0.06)	0.07(0.07)	0.08(0.08)	0.22(0.23)
	F	0.14(0.08)	0.15(0.08)	0.11(0.07)	0.09(0.09)	0.09(0.10)	0.16(0.14)
Médias	S	0.26(0.13)	0.26(0.15)	0.22(0.12)	0.18(0.12)	0.17(0.11)	0.12(0.09)
	P	0.12(0.07)	0.14(0.09)	0.16(0.12)	0.14(0.11)	0.14(0.10)	0.18(0.17)
	F	0.15(0.08)	0.17(0.09)	0.17(0.09)	0.15(0.09)	0.15(0.09)	0.13(0.10)

Apêndice H - Desempenho ensemble de classificadores (*boosting* e *bagging*)

O desempenho dos *ensembles* de classificadores é apresentado nas subseções a seguir para todas as sub-subclasses EC, incluindo as diferentes taxas de amostragem (proporção entre classes).

i. *RUSBoost*

RIPPER							
Proporção entre classes		1:1	1:2	1:6	1:10	1:25	1:50
EC.1.1.1	S	0.63(0.30)	0.44(0.25)	0.54(0.16)	0.36(0.18)	0.22(0.12)	0.19(0.13)
	P	0.04(0.02)	0.10(0.08)	0.15(0.07)	0.14(0.07)	0.24(0.20)	0.24(0.11)
	F	0.08(0.04)	0.13(0.07)	0.22(0.07)	0.20(0.09)	0.22(0.13)	0.20(0.11)
EC.2.1.1	S	0.54(0.30)	0.46(0.20)	0.24(0.20)	0.30(0.21)	0.16(0.20)	0.02(0.03)
	P	0.04(0.03)	0.07(0.05)	0.08(0.07)	0.15(0.09)	0.13(0.15)	0.07(0.14)
	F	0.07(0.04)	0.12(0.06)	0.12(0.11)	0.19(0.12)	0.13(0.16)	0.03(0.05)
EC.2.3.1	S	0.72(0.19)	0.49(0.16)	0.32(0.17)	0.29(0.17)	0.15(0.10)	0.06(0.09)
	P	0.05(0.01)	0.08(0.05)	0.18(0.13)	0.24(0.13)	0.20(0.10)	0.25(0.33)
	F	0.10(0.02)	0.12(0.03)	0.19(0.08)	0.25(0.12)	0.17(0.09)	0.08(0.10)
EC.2.4.2	S	0.65(0.12)	0.31(0.34)	0.29(0.22)	0.21(0.15)	0.14(0.08)	0.12(0.12)
	P	0.05(0.01)	0.03(0.04)	0.10(0.08)	0.12(0.07)	0.20(0.20)	0.35(0.35)
	F	0.09(0.03)	0.06(0.07)	0.13(0.10)	0.15(0.08)	0.15(0.11)	0.17(0.16)
EC.2.5.1	S	0.60(0.30)	0.24(0.18)	0.20(0.16)	0.27(0.25)	0.17(0.16)	0.09(0.11)
	P	0.04(0.02)	0.14(0.19)	0.13(0.10)	0.15(0.15)	0.30(0.32)	0.37(0.44)
	F	0.08(0.04)	0.15(0.13)	0.13(0.09)	0.19(0.18)	0.19(0.16)	0.14(0.16)
EC.2.7.1	S	0.55(0.40)	0.50(0.24)	0.31(0.25)	0.40(0.21)	0.20(0.12)	0.18(0.17)
	P	0.04(0.03)	0.10(0.07)	0.11(0.10)	0.20(0.14)	0.29(0.20)	0.22(0.18)

	F	0.07(0.06)	0.15(0.08)	0.14(0.11)	0.25(0.12)	0.23(0.14)	0.19(0.17)
EC.3.1.1	S	0.60(0.20)	0.60(0.22)	0.45(0.21)	0.34(0.16)	0.25(0.08)	0.22(0.10)
	P	0.14(0.20)	0.13(0.08)	0.24(0.07)	0.33(0.09)	0.56(0.21)	0.64(0.09)
	F	0.16(0.08)	0.20(0.11)	0.29(0.09)	0.31(0.07)	0.34(0.10)	0.32(0.10)
EC.3.1.3	S	0.69(0.28)	0.60(0.29)	0.51(0.19)	0.36(0.19)	0.27(0.17)	0.14(0.13)
	P	0.10(0.04)	0.15(0.10)	0.25(0.13)	0.20(0.10)	0.33(0.16)	0.32(0.30)
	F	0.17(0.06)	0.21(0.11)	0.31(0.12)	0.25(0.11)	0.27(0.12)	0.19(0.18)
EC.3.2.1	S	0.89(0.08)	0.71(0.26)	0.67(0.07)	0.55(0.11)	0.47(0.08)	0.42(0.17)
	P	0.07(0.01)	0.09(0.05)	0.18(0.03)	0.23(0.07)	0.37(0.07)	0.44(0.09)
	F	0.14(0.02)	0.16(0.08)	0.29(0.04)	0.31(0.05)	0.41(0.05)	0.42(0.13)
EC.3.4.21	S	0.60(0.27)	0.50(0.23)	0.32(0.21)	0.37(0.18)	0.22(0.25)	0.12(0.11)
	P	0.06(0.03)	0.17(0.21)	0.23(0.15)	0.29(0.23)	0.38(0.36)	0.47(0.40)
	F	0.10(0.06)	0.19(0.11)	0.24(0.11)	0.30(0.19)	0.22(0.17)	0.19(0.16)
EC.4.1.1	S	0.34(0.30)	0.36(0.26)	0.32(0.24)	0.26(0.18)	0.15(0.19)	0.04(0.06)
	P	0.02(0.02)	0.06(0.07)	0.12(0.07)	0.25(0.29)	0.23(0.30)	0.22(0.25)
	F	0.04(0.03)	0.08(0.05)	0.16(0.11)	0.18(0.11)	0.14(0.14)	0.07(0.09)
EC.4.2.1	S	0.63(0.26)	0.37(0.29)	0.27(0.16)	0.30(0.16)	0.16(0.12)	0.11(0.07)
	P	0.05(0.03)	0.07(0.05)	0.12(0.05)	0.21(0.13)	0.23(0.14)	0.44(0.36)
	F	0.09(0.05)	0.12(0.08)	0.15(0.07)	0.24(0.14)	0.18(0.12)	0.15(0.10)
C4.5							
Proporção entre classes		1:01	1:02	1:06	1:10	1:25	1:50
EC.1.1.1	S	0.74(0.17)	0.72(0.17)	0.46(0.20)	0.41(0.19)	0.25(0.19)	0.17(0.11)
	P	0.05(0.02)	0.07(0.03)	0.11(0.04)	0.18(0.10)	0.16(0.08)	0.29(0.17)
	F	0.09(0.03)	0.13(0.04)	0.18(0.06)	0.24(0.10)	0.19(0.11)	0.19(0.12)
EC.2.1.1	S	0.75(0.18)	0.46(0.17)	0.17(0.12)	0.17(0.10)	0.17(0.10)	0.04(0.07)
	P	0.04(0.01)	0.06(0.05)	0.06(0.04)	0.09(0.05)	0.35(0.30)	0.07(0.12)
	F	0.08(0.02)	0.11(0.07)	0.08(0.06)	0.11(0.06)	0.19(0.09)	0.05(0.09)
EC.2.3.1	S	0.69(0.14)	0.61(0.08)	0.39(0.18)	0.32(0.15)	0.23(0.13)	0.19(0.12)
	P	0.05(0.02)	0.07(0.02)	0.15(0.07)	0.19(0.08)	0.35(0.18)	0.54(0.26)
	F	0.09(0.04)	0.13(0.03)	0.21(0.09)	0.22(0.08)	0.27(0.14)	0.27(0.15)
EC.2.4.2	S	0.74(0.12)	0.65(0.18)	0.35(0.12)	0.32(0.14)	0.14(0.13)	0.12(0.10)
	P	0.04(0.01)	0.07(0.04)	0.11(0.06)	0.16(0.08)	0.16(0.16)	0.21(0.18)
	F	0.08(0.03)	0.13(0.06)	0.16(0.08)	0.20(0.09)	0.14(0.13)	0.13(0.12)
EC.2.5.1	S	0.64(0.16)	0.52(0.16)	0.34(0.21)	0.29(0.15)	0.13(0.11)	0.11(0.14)
	P	0.05(0.02)	0.08(0.02)	0.12(0.07)	0.20(0.08)	0.17(0.14)	0.24(0.31)
	F	0.09(0.04)	0.13(0.04)	0.17(0.09)	0.23(0.09)	0.14(0.11)	0.12(0.14)
EC.2.7.1	S	0.78(0.19)	0.62(0.24)	0.46(0.22)	0.35(0.14)	0.16(0.11)	0.15(0.09)
	P	0.06(0.02)	0.07(0.03)	0.12(0.07)	0.18(0.10)	0.19(0.15)	0.31(0.15)
	F	0.11(0.04)	0.12(0.05)	0.18(0.09)	0.23(0.11)	0.17(0.11)	0.19(0.10)

EC.3.1.1	S	0.77(0.14)	0.66(0.16)	0.49(0.21)	0.38(0.10)	0.23(0.12)	0.16(0.07)
	P	0.07(0.02)	0.13(0.03)	0.25(0.07)	0.26(0.10)	0.43(0.18)	0.65(0.28)
	F	0.13(0.03)	0.22(0.05)	0.30(0.08)	0.30(0.06)	0.29(0.12)	0.25(0.10)
EC.3.1.3	S	0.79(0.25)	0.76(0.19)	0.59(0.22)	0.40(0.14)	0.28(0.17)	0.15(0.13)
	P	0.10(0.02)	0.13(0.04)	0.25(0.08)	0.23(0.09)	0.35(0.19)	0.32(0.25)
	F	0.18(0.04)	0.21(0.06)	0.33(0.09)	0.27(0.08)	0.29(0.17)	0.20(0.16)
EC.3.2.1	S	0.88(0.08)	0.84(0.09)	0.71(0.12)	0.61(0.14)	0.48(0.14)	0.36(0.12)
	P	0.07(0.02)	0.10(0.02)	0.18(0.03)	0.22(0.05)	0.35(0.12)	0.45(0.14)
	F	0.14(0.03)	0.18(0.03)	0.28(0.05)	0.32(0.06)	0.39(0.10)	0.39(0.10)
EC.3.4.2.1	S	0.76(0.21)	0.58(0.18)	0.35(0.18)	0.37(0.25)	0.21(0.16)	0.15(0.15)
	P	0.07(0.03)	0.10(0.03)	0.18(0.10)	0.20(0.17)	0.35(0.29)	0.51(0.40)
	F	0.12(0.06)	0.17(0.05)	0.23(0.13)	0.25(0.18)	0.23(0.15)	0.20(0.18)
EC.4.1.1	S	0.53(0.20)	0.49(0.16)	0.29(0.12)	0.24(0.15)	0.10(0.06)	0.11(0.11)
	P	0.03(0.01)	0.05(0.02)	0.10(0.05)	0.13(0.07)	0.28(0.31)	0.42(0.38)
	F	0.05(0.02)	0.10(0.04)	0.15(0.06)	0.16(0.10)	0.13(0.09)	0.14(0.10)
EC.4.2.1	S	0.65(0.16)	0.54(0.14)	0.31(0.19)	0.24(0.16)	0.09(0.08)	0.09(0.10)
	P	0.05(0.01)	0.08(0.03)	0.12(0.06)	0.14(0.09)	0.15(0.11)	0.18(0.19)
	F	0.09(0.02)	0.14(0.04)	0.17(0.09)	0.18(0.11)	0.11(0.09)	0.11(0.12)

ii. *SMOTEBoost*

		Taxa de sobreamostragem em %									
Proporção entre classes		100	200	300	400	500	600	700	800	900	1000
EC.1.1.1	S	0.18(0.07)	0.29(0.16)	0.26(0.18)	0.23(0.14)	0.29(0.12)	0.25(0.20)	0.32(0.21)	0.26(0.11)	0.26(0.21)	0.24(0.16)
	P	0.38(0.18)	0.40(0.13)	0.43(0.24)	0.34(0.21)	0.39(0.23)	0.28(0.15)	0.40(0.24)	0.34(0.20)	0.25(0.15)	0.27(0.14)
	F	0.24(0.09)	0.31(0.13)	0.29(0.16)	0.26(0.15)	0.31(0.12)	0.26(0.17)	0.31(0.15)	0.28(0.11)	0.23(0.13)	0.24(0.15)
EC.2.1.1	S	0.10(0.11)	0.06(0.07)	0.10(0.12)	0.09(0.15)	0.13(0.13)	0.06(0.09)	0.08(0.09)	0.05(0.08)	0.12(0.15)	0.11(0.11)
	P	0.31(0.34)	0.21(0.32)	0.19(0.23)	0.18(0.24)	0.30(0.31)	0.17(0.24)	0.26(0.35)	0.16(0.32)	0.23(0.18)	0.21(0.23)
	F	0.14(0.15)	0.09(0.10)	0.13(0.16)	0.12(0.17)	0.18(0.18)	0.07(0.10)	0.12(0.13)	0.06(0.10)	0.15(0.14)	0.14(0.13)
EC.2.3.1	S	0.16(0.06)	0.16(0.08)	0.14(0.11)	0.15(0.10)	0.14(0.07)	0.13(0.09)	0.18(0.12)	0.17(0.09)	0.14(0.06)	0.17(0.12)
	P	0.28(0.05)	0.27(0.11)	0.28(0.29)	0.18(0.11)	0.25(0.20)	0.24(0.29)	0.22(0.15)	0.22(0.12)	0.19(0.08)	0.15(0.10)
	F	0.19(0.04)	0.19(0.09)	0.16(0.10)	0.16(0.10)	0.16(0.08)	0.14(0.08)	0.19(0.13)	0.18(0.10)	0.15(0.05)	0.16(0.10)
EC.2.4.2	S	0.04(0.06)	0.03(0.05)	0.04(0.05)	0.04(0.06)	0.08(0.11)	0.06(0.06)	0.05(0.09)	0.07(0.08)	0.15(0.20)	0.12(0.06)
	P	0.08(0.12)	0.08(0.13)	0.11(0.18)	0.16(0.32)	0.08(0.09)	0.11(0.10)	0.03(0.06)	0.12(0.16)	0.10(0.12)	0.13(0.11)
	F	0.05(0.06)	0.04(0.07)	0.05(0.07)	0.06(0.09)	0.07(0.08)	0.07(0.07)	0.03(0.06)	0.09(0.10)	0.10(0.11)	0.11(0.07)
EC.2.5.1	S	0.09(0.12)	0.07(0.10)	0.05(0.08)	0.08(0.10)	0.07(0.10)	0.04(0.04)	0.11(0.11)	0.07(0.07)	0.08(0.07)	0.10(0.08)
	P	0.27(0.35)	0.22(0.32)	0.09(0.13)	0.21(0.31)	0.19(0.32)	0.10(0.12)	0.18(0.19)	0.12(0.13)	0.19(0.20)	0.28(0.29)
	F	0.12(0.16)	0.09(0.11)	0.06(0.09)	0.10(0.11)	0.09(0.13)	0.05(0.06)	0.13(0.13)	0.09(0.09)	0.10(0.09)	0.14(0.10)

EC.2.7.1	S	0.23(0.13)	0.21(0.14)	0.20(0.13)	0.18(0.11)	0.18(0.13)	0.18(0.10)	0.23(0.15)	0.23(0.14)	0.22(0.13)	0.22(0.10)
	P	0.31(0.18)	0.28(0.18)	0.28(0.23)	0.27(0.19)	0.29(0.29)	0.25(0.19)	0.33(0.22)	0.25(0.19)	0.25(0.10)	0.26(0.14)
	F	0.25(0.12)	0.23(0.13)	0.21(0.14)	0.20(0.12)	0.21(0.16)	0.20(0.12)	0.26(0.15)	0.23(0.14)	0.22(0.10)	0.23(0.11)
EC.3.1.1	S	0.31(0.10)	0.30(0.14)	0.33(0.14)	0.29(0.09)	0.35(0.10)	0.38(0.11)	0.32(0.11)	0.37(0.10)	0.35(0.10)	0.35(0.14)
	P	0.64(0.24)	0.61(0.23)	0.56(0.17)	0.54(0.17)	0.59(0.13)	0.53(0.15)	0.47(0.14)	0.60(0.23)	0.50(0.15)	0.48(0.17)
	F	0.40(0.12)	0.36(0.09)	0.41(0.15)	0.37(0.10)	0.43(0.09)	0.43(0.09)	0.38(0.12)	0.45(0.13)	0.40(0.10)	0.40(0.15)
EC.3.1.3	S	0.18(0.15)	0.18(0.15)	0.18(0.11)	0.17(0.13)	0.22(0.10)	0.13(0.12)	0.23(0.17)	0.16(0.11)	0.24(0.14)	0.22(0.11)
	P	0.40(0.29)	0.27(0.22)	0.39(0.28)	0.31(0.18)	0.31(0.15)	0.20(0.21)	0.27(0.21)	0.21(0.10)	0.30(0.12)	0.32(0.23)
	F	0.20(0.11)	0.20(0.15)	0.22(0.11)	0.20(0.12)	0.23(0.06)	0.14(0.13)	0.23(0.17)	0.17(0.09)	0.24(0.11)	0.23(0.12)
EC.3.2.1	S	0.41(0.09)	0.41(0.10)	0.43(0.09)	0.47(0.09)	0.48(0.14)	0.47(0.09)	0.46(0.11)	0.48(0.09)	0.51(0.07)	0.45(0.10)
	P	0.62(0.11)	0.64(0.16)	0.62(0.18)	0.61(0.11)	0.57(0.09)	0.57(0.13)	0.57(0.14)	0.56(0.13)	0.57(0.07)	0.52(0.13)
	F	0.49(0.08)	0.49(0.11)	0.50(0.12)	0.53(0.09)	0.51(0.12)	0.51(0.09)	0.51(0.11)	0.51(0.08)	0.53(0.04)	0.47(0.08)
EC.3.4.21	S	0.27(0.12)	0.35(0.22)	0.32(0.19)	0.34(0.15)	0.34(0.14)	0.37(0.12)	0.38(0.21)	0.35(0.20)	0.35(0.17)	0.36(0.12)
	P	0.71(0.24)	0.68(0.24)	0.54(0.24)	0.69(0.15)	0.60(0.14)	0.65(0.20)	0.45(0.13)	0.54(0.30)	0.54(0.25)	0.51(0.21)
	F	0.36(0.14)	0.42(0.21)	0.39(0.20)	0.43(0.13)	0.43(0.14)	0.44(0.09)	0.39(0.16)	0.42(0.24)	0.40(0.17)	0.41(0.14)
EC.4.1.1	S	0.24(0.19)	0.17(0.13)	0.22(0.12)	0.22(0.18)	0.20(0.13)	0.20(0.17)	0.25(0.19)	0.22(0.19)	0.17(0.15)	0.15(0.14)
	P	0.52(0.35)	0.41(0.33)	0.46(0.11)	0.39(0.21)	0.50(0.33)	0.34(0.26)	0.42(0.20)	0.35(0.28)	0.34(0.24)	0.26(0.27)
	F	0.30(0.20)	0.22(0.16)	0.29(0.13)	0.26(0.17)	0.25(0.13)	0.24(0.19)	0.29(0.18)	0.26(0.21)	0.22(0.17)	0.19(0.18)
EC.4.2.1	S	0.09(0.09)	0.11(0.07)	0.14(0.17)	0.14(0.12)	0.16(0.09)	0.10(0.08)	0.13(0.05)	0.12(0.10)	0.15(0.06)	0.17(0.14)
	P	0.31(0.31)	0.36(0.28)	0.25(0.18)	0.27(0.17)	0.26(0.18)	0.29(0.20)	0.27(0.14)	0.20(0.17)	0.28(0.12)	0.24(0.14)
	F	0.13(0.12)	0.16(0.10)	0.17(0.15)	0.17(0.12)	0.18(0.11)	0.14(0.11)	0.16(0.06)	0.14(0.12)	0.19(0.06)	0.19(0.12)

iii. *SMOTEBagging e RIPPER com redução de dimensionalidade e SMOTE*

Sub-subclasse		RIPPER	SMOTEBagging
EC.1.1.1	S	0.17(0.10)	0.39(0.26)
	P	0.37(0.28)	0.28(0.06)
	F	0.22(0.13)	0.30(0.13)
EC.2.1.1	S	0.09(0.10)	0.10(0.13)
	P	0.35(0.39)	0.23(0.27)
	F	0.14(0.15)	0.13(0.16)
EC.2.3.1	S	0.11(0.08)	0.28(0.18)
	P	0.46(0.32)	0.21(0.10)
	F	0.15(0.09)	0.22(0.10)
EC.2.4.2	S	0.11(0.11)	0.10(0.09)
	P	0.48(0.43)	0.17(0.22)
	F	0.17(0.16)	0.11(0.11)
EC.2.5.1	S	0.05(0.07)	0.12(0.09)
	P	0.26(0.42)	0.24(0.19)

	F	0.07(0.10)	0.15(0.11)
EC.2.7.1	S	0.22(0.21)	0.31(0.19)
	P	0.40(0.30)	0.32(0.18)
	F	0.25(0.19)	0.28(0.14)
EC.3.1.1	S	0.39(0.11)	0.38(0.09)
	P	0.58(0.22)	0.53(0.15)
	F	0.45(0.13)	0.43(0.09)
EC.3.1.3	S	0.21(0.14)	0.21(0.17)
	P	0.57(0.32)	0.24(0.17)
	F	0.28(0.15)	0.20(0.13)
EC.3.2.1	S	0.47(0.14)	0.59(0.12)
	P	0.59(0.10)	0.57(0.12)
	F	0.52(0.12)	0.57(0.09)
EC.3.4.21	S	0.34(0.19)	0.44(0.13)
	P	0.54(0.22)	0.46(0.14)
	F	0.39(0.18)	0.44(0.10)
EC.4.1.1	S	0.21(0.22)	0.18(0.17)
	P	0.35(0.36)	0.32(0.25)
	F	0.24(0.23)	0.22(0.18)
EC.4.2.1	S	0.08(0.09)	0.20(0.18)
	P	0.32(0.38)	0.32(0.27)
	F	0.12(0.14)	0.22(0.16)

iv. FS+SMOTE+RIPPER

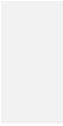
Proporção entre classes		100	200	300	400	500	600	700	800	900	1000
EC.1.1.1	S	0.27(0.20)	0.32(0.18)	0.37(0.13)	0.37(0.22)	0.35(0.15)	0.32(0.20)	0.36(0.18)	0.42(0.19)	0.39(0.11)	0.41(0.11)
	P	0.27(0.16)	0.21(0.07)	0.20(0.05)	0.18(0.06)	0.18(0.10)	0.15(0.08)	0.12(0.04)	0.17(0.09)	0.14(0.06)	0.15(0.07)
	F	0.26(0.17)	0.24(0.10)	0.25(0.06)	0.22(0.08)	0.23(0.10)	0.20(0.10)	0.17(0.07)	0.22(0.08)	0.20(0.07)	0.21(0.07)
EC.2.1.1	S	0.08(0.11)	0.06(0.12)	0.09(0.13)	0.10(0.12)	0.12(0.12)	0.12(0.11)	0.11(0.16)	0.15(0.14)	0.11(0.14)	0.11(0.15)
	P	0.16(0.24)	0.09(0.17)	0.08(0.11)	0.14(0.17)	0.28(0.26)	0.29(0.35)	0.15(0.19)	0.14(0.11)	0.10(0.14)	0.11(0.13)
	F	0.10(0.14)	0.06(0.10)	0.08(0.12)	0.11(0.12)	0.16(0.15)	0.15(0.16)	0.12(0.17)	0.14(0.12)	0.10(0.13)	0.09(0.12)
EC.2.3.1	S	0.17(0.12)	0.21(0.14)	0.29(0.27)	0.27(0.19)	0.30(0.14)	0.37(0.14)	0.37(0.25)	0.36(0.25)	0.31(0.15)	0.34(0.23)
	P	0.25(0.16)	0.20(0.12)	0.12(0.08)	0.14(0.10)	0.14(0.05)	0.12(0.04)	0.11(0.05)	0.11(0.06)	0.10(0.05)	0.09(0.04)
	F	0.19(0.11)	0.18(0.10)	0.16(0.11)	0.17(0.11)	0.18(0.07)	0.17(0.05)	0.16(0.09)	0.16(0.09)	0.14(0.05)	0.13(0.06)
EC.2.4.2	S	0.07(0.08)	0.10(0.11)	0.05(0.06)	0.16(0.09)	0.15(0.16)	0.08(0.08)	0.13(0.12)	0.17(0.07)	0.10(0.08)	0.14(0.17)
	P	0.23(0.32)	0.15(0.14)	0.18(0.33)	0.10(0.08)	0.07(0.07)	0.08(0.12)	0.06(0.08)	0.11(0.06)	0.05(0.05)	0.07(0.09)
	F	0.10(0.12)	0.11(0.11)	0.07(0.10)	0.11(0.07)	0.09(0.09)	0.08(0.08)	0.06(0.05)	0.12(0.05)	0.07(0.06)	0.09(0.11)
EC.2.5.1	S	0.09(0.10)	0.19(0.12)	0.17(0.14)	0.07(0.07)	0.17(0.12)	0.11(0.12)	0.08(0.08)	0.23(0.13)	0.22(0.16)	0.14(0.11)

	P	0.16(0.18)	0.23(0.14)	0.11(0.08)	0.10(0.10)	0.17(0.16)	0.13(0.16)	0.10(0.11)	0.14(0.07)	0.15(0.10)	0.09(0.07)
	F	0.11(0.11)	0.19(0.11)	0.12(0.08)	0.08(0.07)	0.16(0.13)	0.11(0.12)	0.09(0.07)	0.17(0.08)	0.17(0.10)	0.10(0.08)
EC.2.7.1	S	0.26(0.11)	0.34(0.17)	0.29(0.13)	0.35(0.16)	0.38(0.09)	0.42(0.30)	0.37(0.22)	0.32(0.17)	0.34(0.11)	0.34(0.14)
	P	0.46(0.27)	0.30(0.11)	0.26(0.15)	0.21(0.13)	0.27(0.17)	0.22(0.15)	0.22(0.10)	0.21(0.14)	0.20(0.12)	0.16(0.05)
	F	0.30(0.11)	0.31(0.13)	0.25(0.11)	0.25(0.12)	0.28(0.10)	0.26(0.16)	0.25(0.10)	0.22(0.12)	0.21(0.05)	0.21(0.06)
EC.3.1.1	S	0.33(0.08)	0.37(0.11)	0.41(0.07)	0.41(0.10)	0.37(0.11)	0.40(0.12)	0.39(0.13)	0.39(0.14)	0.38(0.12)	0.41(0.11)
	P	0.52(0.15)	0.40(0.13)	0.43(0.12)	0.45(0.12)	0.38(0.14)	0.36(0.12)	0.38(0.17)	0.40(0.14)	0.33(0.09)	0.34(0.13)
	F	0.38(0.07)	0.38(0.11)	0.41(0.08)	0.41(0.07)	0.36(0.11)	0.37(0.10)	0.37(0.10)	0.37(0.11)	0.35(0.09)	0.36(0.10)
EC.3.1.3	S	0.31(0.14)	0.25(0.17)	0.36(0.12)	0.34(0.23)	0.40(0.20)	0.42(0.19)	0.46(0.17)	0.40(0.20)	0.51(0.17)	0.47(0.22)
	P	0.44(0.26)	0.31(0.21)	0.29(0.14)	0.19(0.09)	0.21(0.10)	0.18(0.08)	0.19(0.08)	0.16(0.10)	0.21(0.08)	0.20(0.12)
	F	0.35(0.16)	0.27(0.19)	0.29(0.10)	0.22(0.10)	0.26(0.10)	0.24(0.09)	0.26(0.11)	0.22(0.12)	0.27(0.06)	0.26(0.13)
EC.3.2.1	S	0.53(0.13)	0.54(0.10)	0.53(0.12)	0.58(0.09)	0.58(0.09)	0.61(0.10)	0.58(0.11)	0.60(0.09)	0.57(0.12)	0.56(0.07)
	P	0.46(0.12)	0.41(0.08)	0.41(0.11)	0.37(0.08)	0.38(0.11)	0.41(0.09)	0.37(0.09)	0.36(0.07)	0.37(0.09)	0.37(0.08)
	F	0.48(0.11)	0.46(0.07)	0.45(0.07)	0.44(0.05)	0.45(0.10)	0.48(0.09)	0.45(0.07)	0.45(0.08)	0.44(0.09)	0.44(0.05)
EC.3.4.21	S	0.42(0.18)	0.40(0.21)	0.40(0.22)	0.43(0.26)	0.48(0.18)	0.40(0.22)	0.39(0.21)	0.43(0.17)	0.40(0.22)	0.42(0.20)
	P	0.57(0.17)	0.42(0.19)	0.30(0.14)	0.34(0.20)	0.37(0.12)	0.37(0.19)	0.34(0.18)	0.35(0.13)	0.32(0.14)	0.36(0.10)
	F	0.47(0.16)	0.38(0.17)	0.33(0.16)	0.37(0.20)	0.41(0.14)	0.36(0.17)	0.35(0.16)	0.37(0.13)	0.35(0.16)	0.37(0.13)
EC.4.1.1	S	0.22(0.22)	0.21(0.17)	0.22(0.17)	0.21(0.15)	0.26(0.14)	0.16(0.14)	0.23(0.23)	0.26(0.21)	0.23(0.11)	0.27(0.28)
	P	0.45(0.35)	0.35(0.23)	0.43(0.26)	0.33(0.20)	0.43(0.14)	0.34(0.30)	0.28(0.21)	0.25(0.15)	0.33(0.19)	0.29(0.30)
	F	0.26(0.24)	0.25(0.18)	0.27(0.17)	0.24(0.17)	0.30(0.10)	0.21(0.16)	0.22(0.16)	0.23(0.15)	0.25(0.11)	0.23(0.19)
EC.4.2.1	S	0.10(0.09)	0.15(0.14)	0.21(0.14)	0.19(0.08)	0.24(0.13)	0.18(0.15)	0.20(0.12)	0.24(0.13)	0.22(0.10)	0.26(0.19)
	P	0.17(0.13)	0.15(0.14)	0.17(0.12)	0.16(0.06)	0.18(0.08)	0.09(0.06)	0.15(0.10)	0.14(0.06)	0.14(0.08)	0.16(0.12)
	F	0.12(0.09)	0.15(0.13)	0.17(0.10)	0.17(0.06)	0.20(0.10)	0.12(0.08)	0.16(0.08)	0.17(0.08)	0.16(0.08)	0.17(0.11)

Apêndice I - Relação dos descritores em cada sub-subclasse EC após redução de dimensionalidade

A tabela a seguir apresenta uma listagem dos descritores selecionados pelo método de seleção de atributos adotado para as diferentes sub-subclasses EC.

EC	1.1.1	2.1.1	2.3.1	2.4.2	2.5.1	2.7.1	3.1.1	3.1.3	3.2.1	3.4.21	4.1.1	4.2.1
	8	10	4	6	10	7	10	4	12	13	4	11
ATRIBUTOS SELECINADOS	Energia total de contatos	CPO r=8.5Å l=15 no CA	CED no CA r=5Å	CLO r=8.5 Å l=20 no CA	CPO r=5Å l=15 no CA	HB-MWWM	CLO r=3.5Å l=15 no LHA	Contatos Hidrofóbicos (WNA-Surf)	Acc. Relativa	HB-SS	CLO r=8.5Å l=20 no CA	CLO r=5Å l=20 no LHA
	Esponjidade no CA r=4Å	CPO r=8.5Å l=15 no LHA	Close-ness	Densidade hidrofóbica local	CPO r=8.5Å l=15 no CA	CLO r=3.5Å l=30 no CA	CPO r=35Å l=30 no CA	Contatos carregados repulsivos (VD)	HB-SS	CLO r=5Å l=15 no CB	CLO r=8.5Å l=15 no LHA (VD)	Esponjidade no LHA r=4Å
	CED no LHA r=3Å	CPO r=8.5Å l=30 no CB	Contatos aromáticos (GN)	Proporção de esferas apolares (pocket)	Esponjidade no LHA r=5Å	CLO r=5Å l=30 no LHA	CPO r=5Å l=20 no CA	SSBOND (VD)	CPO r=8.5Å l=30 no CA	CPO r=3.5Å l=30 no CA	CPO r=8.5Å l=30 no LHA (VD)	Distância ao C-Terminal
	ASA apolar	CLO r=8.5Å l=20 no LHA (WNA-Surf)	CPO r=8.5Å l=30 no CA (VD)	CPO r=8.5Å l=30 no LHA (WNA-Dist)	CED no CA r=6Å	CPO r=5Å l=20 no CA	Centralidade radial	CED no LHA r=5Å l=5Å (WNA-Surf)	Distância ao Centro de Massa	CPO r=3.5Å l=30 no LHA	Esponjidade LHA r=4Å (GN)	Local Close-ness
	Centralidade alfa	CLO r=8.5Å l=30 no LHA WNA-Surf		CPO r=8.5Å l=15 no CB (VD)	HB-MWWM não usadas	CPO r=8.5Å l=30 no LHA	Contatos hidrofóbicos (WNA-Surf)		CED no CA r=4Å	CPO r=5Å l=30 no CB		Contatos carregados atrativos (WNA-Surf)
	CLO r=3.5Å l=15 no CA (WNA-Surf)	CLO r=8.5Å l=15 no CB (GN)		Esponjidade no LHA r=6Å (WNA-Surf)	HB-MWS não usadas	Local Close-ness	HB-SS (VD)		Local Close-ness	Densidade no LHA r=4Å		HB-SWS (WNA-Dist)
	Esponjidade no CA r=5Å (WNA-Dist)	CPO r=8.5Å l=30 no LHA (WNA-Surf)			CLO r=3.5Å l=15 no CA (WNA-Dist)	CPO r=5Å l=20 no LHA (WNA-Surf)	CPO r=8.5Å l=30 no CA (WNA-Dist)		Contatos aromáticos não usados	Distância ao centro de massa		Energia total de contatos (GN)
	Densidade no LHA r=4Å (GN)	CPO r=3.5Å l=20 no CB (VD)			CLO r=3.5Å l=30 no CA (VD)		CPO r=5Å l=30 no LHA (GN)		CLO r=8.5Å l=20 no CB (WNA-Dist)	CED no CA r=4Å		CLO r=5Å l=20 no CA (WNA-Dist)
		CED no LHA r=7Å (WNA-Dist)			CPO r=5Å l=30 no CA (WNA-Surf)		CED no CA r=6Å (WNA-Surf)		CPO r=3.5Å l=30 no LHA (GN)	Contatos carregados atrativos não usados		CPO r=8.5Å l=20 no CB (WNA-Dist)
		HB-MS não usadas (WNA-Surf)			CED no CA r=7Å (VD)		CED no CA r=4Å (VD)		Esponjidade no CA r=6Å (WNA-Dist)	CLO r=3.5Å l=20 no CB (WNA-Surf)		CED no LHA r=5Å (GN)
								Densidade no LHA r=3Å (VD)	CLO r=3.5Å l=15 no CA (VD)		Pot. Eletrostático Médio (VD)	
								HB-MWS não usadas	CPO r=5Å l=20 no LHA (VD)			



CPO
r=3.5Å
l=30 no
CB
(GN)

CLO: Cross Link Order; CPO: Cross Presence Order; CED: Densidade de Energia de contatos; HB: Hydrogen Bonds; SS: Side chain-side chain; CA: Carbono alfa; CB: carbono beta; LHA: Last Heavy Atom; MW(W)M: Main chain-water-(water)-main chain; MWS: Main chain-water-side chain; VD: Voronoi Diagram; WNASurf: WNA Surface; WNADist: WNA Distance; GN: Graph Neighbors; r: raio; l: tamanho da janela deslizante; SSBOND: ponte dissulfeto;