SABRINA DE AZEVEDO SILVEIRA

ENZYMAP: EXPLORANDO METADADOS PROTÉICOS PARA MODELAGEM E PREVISÃO DE MUDANÇAS DE ANOTAÇÃO NO UNIPROT/SWISS-PROT

Belo Horizonte
25 de janeiro de 2013

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

ENZYMAP: EXPLORANDO METADADOS PROTÉICOS PARA MODELAGEM E PREVISÃO DE MUDANÇAS DE ANOTAÇÃO NO UNIPROT/SWISS-PROT

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

SABRINA DE AZEVEDO SILVEIRA

Belo Horizonte
25 de janeiro de 2013

UNIVERSIDADE FEDERAL DE MINAS GERAIS

ENZYMAP: Explorando metadados protéicos para modelagem e previsão de mudanças de anotação no UniProt/Swiss-Prot

SABRINA DE AZEVEDO SILVEIRA

Ph. D. Wagner Meira Jr – Orientador
Universidade Federal de Minas Gerais

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

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

Belo Horizonte, 25 de janeiro de 2013

Resumo

A geração de dados biológicos experimentou um crescimento sem precedentes nas últimas décadas. Fatores como otimização e diminuição do custo de diversos processos laboratoriais associados às novas técnicas de sequenciamento de DNA, o sequenciamento de proteínas e a determinação de sua estrutura foram responsáveis pela geração de uma grande quantidade de dados. Muitos deles estão em bases de dados biológicos disponibilizadas publicamente através da Internet. Essas bases armazenam não apenas dados brutos biológicos, mas também informações relevantes a respeito de função de proteína, dados da literatura e relação entre proteína e seu gene codificante, dentre outros metadados, também chamados de anotação.

Nesse trabalho é proposta uma estratégia baseada em aprendizagem supervisionada para caracterizar e prever mudanças de anotação em dados temporais denominada *ENZYmatic Metadata Annotation Predictor* (ENZYMAP). Mais precisamente, estamos interessados em prever anotação de função de enzima com base em metadados das entradas do repositório UniProt/Swiss-Prot. Essa proposta permite sugerir possíveis correções para anotações e pode ser utilizada de modo complementar a outros métodos de anotação ajudando a aprimorar a qualidade e confiabilidade do repositório usando dados já disponíveis, o que não demanda experimento de bancada. Além disso, há um enorme volume de dados que não pode ser analisado manualmente, daí a importância de métodos de anotação automática confiávies.

Foi realizada uma exploração inicial dos dados na qual as mudanças de anotação EC foram modeladas considerando a natureza numérica e hierárquica desse sistema classificação de enzimas. Essa etapa deu origem à uma ferramenta de visualização interativa chamada Advise e a um artigo publicado no *IEEE Symposium on Biological Data Visualization (BioVis), 2012*. Na sequência foram selecionados metadados do Swiss-Prot (OC, RP e KW) para descrever entradas que sofreram um tipo específico de mudança de EC das entradas cuja anotação se manteve constante. Matrizes de ocorrência foram propostas para modelar as mudanças de EC *number* em termos dos metadados do Swiss-Prot e serviram como insumo para a estratégia de aprendizagem supervisionada.

Para caracterizar e prever as mudanças de anotação EC, três experimentos foram realizados: *Descritivo Multiclasse*, no qual conclui-se que os metadados selecionados foram capazes de discriminar entradas que experimentaram uma mudança específica no EC *number* daquelas entradas em que a anotação permaneceu constante; *Previsivo Multiclasse*

nos indicou que prever a última ocorrência de um determinado tipo de mudança de EC utilizando um único classificador multiclasse com número escasso de exemplos não foi possível; *Previsivo Origem Comum*, no qual conclui-se que é possível fazer previsão de um determinado tipo de mudança de EC utilizando classificadores mais especializados mesmo com a restrição do número de exemplos.

As previsões realizadas pelo ENZYMAP foram comparadas às previsões feitas pelo *software* DETECT, que associa um EC *number* à sequência de resíduos de uma proteína, e ambas foram confrontadas com as anotações do Swiss-Prot. O percentual de previsões feitas pelo ENZYMAP que está de acordo com o Swiss-Prot é maior que o mesmo percentual para o DETECT para todos os quatro níveis da anotação EC.

Abstract

In recent decades there has been a surge in the amount of available biological data. New DNA sequencing technologies have made economically possible an increasing number of large data projects, which led to an exponential increase in DNA sequence data. Also, vast amounts of data such as protein sequences and structures, gene-expression measurements, protein and genetic interactions and phenotype studies have been produced. Much of these data are organized and publicly available to the scientific community in biological repositories via the Internet. These repositories store not only biological raw data but also relevant information such as protein function, literature information and the relationship between a protein and its encoding gene, among other metadata, also called annotation.

In this work we propose a supervised learning approach to characterize and predict annotation changes in temporal data, which we term ENZYmatic Metadata Annotation Predictor (ENZYMAP). More precisely, we are interested in predict enzyme function annotation based on UniProt/Swiss-Prot entry metadata. This proposal allows us to suggest possible corrections to annotations from biological repositories and can be used in a complementary manner to other annotation methods improving the quality and realiability of these data. Our approach uses data already available to enhance the repository, which does not demand new expensive bench experiments. Furthermore, there is a huge volume of data that can not be analyzed manually, hence the importance of reliable automatic annotation methods.

We performed an initial exploration of the data in which EC number changes were modeled considering the numeric and hierarchical nature of EC enzyme classification system. This step led to the creation of an interactive visualization tool called Advise and also to the publication of an article in IEEE Symposium on Biological Data Visualization (BioVis), 2012. Then some metadata from Swiss-Prot (OC, RP e KW) were selected to discriminate entries that experienced a specific EC change type from those which annotation remained constant. Ocurrence matrices were proposed to model EC number changes in terms of Swiss-Prot metadata and such matrices served as input for the supervised learning approach.

We performed three experiments to characterize and predict EC number changes: *Descriptive Multiclass*, in which we concluded that selected metadata were able to discriminate entries that undergone a specific EC number change from those which annotation

remained constant; *Predictive Multiclass* indicated that predicting the last ocurrence of a EC change type using a single multiclass classifier with a scarce number of examples was not possible; *Predictive Common Source*, in which we concluded that predicting an EC change type using more specialized classifiers is possible even with a scarce number of examples.

We compared predictions made by ENZYMAP to predictions made by DETECT, a technique able to associate an EC number to the residues' sequence of a protein, and both were checked against Swiss-Prot annotations. The percentage of predictions made by our approach that is in accordance with Swiss-Prot is greater than the same percentage for DETECT for all four levels of EC annotation.

"De tudo ficaram três coisas: a certeza de que ele estava sempre começando, a certeza de que era preciso continuar e a certeza de que seria interrompido antes de terminar. Fazer da interrupção um caminho novo. Fazer da queda um passo de dança, do medo uma escada, do sono uma ponte, da procura um encontro."

Fernando Sabino

Agradecimentos

Agradeço a Deus pela vida e pela fé que me sustenta nos momentos mais difíceis.

Aos meus pais, Jaildo e Bárbara, agradeço pelo amor incondicional, pela humildade e simplicidade, pelo exemplo de caráter e pelas orações. Em especial à minha mãe pela dedicação e por sempre acreditar em mim.

Ao meu marido Ronan pelo amor, apoio, compreensão, paciência e tolerância que foram idispensáveis para que eu pudesse me dedicar ao doutorado. Ronan, você vai direto pro céu, e sem escala!

Ao meu orientador, professor Wagner Meira Jr, pela oportunidade de trabalhar com pesquisa ainda na graduação e por me acompanhar nos momentos mais críticos desse doutorado. Ao meu co-orientador, professor Carlos Silveira, obrigada por me acompanhar mesmo à distância, quando eu estava no México, e por ter me apresentado à Bioinformática. Ao professor Marcelo Matos Santoro pelo apoio e por me ajudar com as questões semânticas da Bioquímica.

Deixo um agradecimento especial à professora Raquel Minardi que com seu talento e inteligência me ajudou imensamente. Obrigada pela sua generosidade em acompanhar de perto esse trabalho. Sua participação foi fundamental e decisiva.

Agradeço também a todos os colegas do Laboratório de Bioinformática e Sistemas (LBS) pelo apoio, pela troca de experiências e pelos momentos de descontração que tornaram a caminhada mais suave. Agradeço ao Douglas que por diversas vezes interrompeu o próprio trabalho para me ajudar. À Valdete e à Nilma pelas palavras de conforto. Ao Sandro pelo feijão da sorte. À Elisa pelo inglês impecável que me ajudou nessa etapa final. Ao Coutinho que fez a Hydra funcionar para eu executar meus experimentos.

Por último, porém mais importante, agradeço à minha filha Laís pelo grande amor, incentivo, motivação, compreensão, apoio, paciência, tolerância e pela alegria inocente. Laís, a você, que desde a minha graduação passou intermináveis horas montando Lego para que eu pudesse fazer os trabalhos práticos (TPs) de AEDS III, meu muito obrigada ainda é pouco. Você é realmente um angelito!

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nas últimas décadas houve um enorme aumento na quantidade de dados biológicos disponíveis. De acordo com [Fritz et al. (2011)], as novas tecnologias de sequenciamento de DNA possibilitaram a diminuição dos custos do sequenciamento e tornaram viáveis um crescente número de grandes projetos de dados, o que levou a um aumento exponencial nos dados de sequência de DNA. Adicionalmente, uma enorme quantidade de dados de sequência e estrutura de proteínas, expressão gênica, interação de proteínas e estudos de fenótipo foram produzidos [Howe et al. (2008)]. Muitos desses dados estão organizados e foram disponibilizados publicamente para a comunidade científica através de repositórios de dados biológicos na Internet. Segundo [Lesk (2005)], tais repositórios armazenam não apenas dados brutos biológicos mas também informações relevantes a respeito das condições experimentais, dos seres vivos envolvidos, de função de proteína, dados da literatura e relação entre proteína e seu gene codificante, dentre outros metadados, também chamados de anotação.

Como a quantidade de dados biológicos está aumentando rapidamente, é comum que subconjuntos selecionados e relevantes de tais dados sejam manualmente revisados enquanto a maior parte dos dados é automaticamente anotada [Mewes et al. (2011)]. Na maioria dos casos, os papéis de genes foram anotados através de similaridade de sequência e propagados para diversos repositórios de dados, sem evidência experimental [Furnham et al. (2009); Brenner et al. (1999)].

A glicoproteína G de Nipah virus (entrada com identificador Q9IH62 no UniProt/Swiss-Prot) ilustra os riscos dessa abordagem. Tal proteína apresenta mais de 50% de similaridade de sequência com as hemaglutinina-neuraminidases, um grupo de enzimas associado ao processo de fusão viral na célula hospedeira. As estruturas da glicoproteína G de Hendra e Nipah virus foram resolvidas (identificadores 2VSK e 2VSM do PDB[1], respectivamente) e possuem o motivo estrutural conhecido como *six-blade β propeller* (uma espécie de hélice formada por 6 folhas beta), típico dessas hidrolases (hemaglutinina-neuraminidases) [Bowden et al. (2008)]. Um alinhamento estrutural

[1] http://www.rcsb.org/pdb/

com uma neuramidase legítima do virus Parainfluenza Type III (indentificador 1V3D no PDB), que também pertence à mesma família Paramyxoviridae de Henipavirus, resultou num RMSD menor que 2,0 Å[Lawrence et al. (2004)]. Um sistema automático de anotação poderia, com base na similaridade de sequência e estrutura, classificar a glicoproteína G de Henipavirus como neuramidase. De fato, até a versão 14 (julho de 2008) do UniProt/Swiss-Prot, a entrada com identificador Q9IH62 era considerada uma enzima. Entretanto, apesar da similaridade no nível de sequência e estrutura, hoje sabe-se que as glicoproteínas G de Henipavirus não são enzimas, e sua atividade é de hemaglutinina, realizando interações proteína-proteína com receptores do hospedeiro [Bowden et al. (2008)]. No momento em que esse texto era escrito, o PDB ainda indicava erroneamente as proteínas (2VSK e 2VSM) como hidrolases.

Desse modo, existe uma preocupação na comunidade científica com relação à qualidade e confiabilidade dos dados e anotações dos grandes repositórios disponíveis publicamente, o que é demonstrado por diversos estudos que abordam as taxas consideráveis de erros de anotação e, de maneira mais geral, o problema da anotação de bases biológicas. Alguns dessses estudos são abordados brevemente abaixo.

Em [Brenner et al. (1999)] e [Devos e Valencia (2001)], as diferenças entre anotações feitas por diferentes grupos de pesquisa para genomas específicos foram analisadas. Um erro sistemático de anotação decorrente da interpretação incorreta de EC *numbers* (um sistema de classificação de enzimas) parciais foi reportado em [Green e Karp (2005)]. Em [Schnoes et al. (2009)], os níveis de falhas de anotação nos repositórios de dados biológicos UniProt [Consortium et al. (2012)], GenBank [Benson et al. (2009)] and KEGG [Kanehisa et al. (2012)] foram investigados com base em Modelos Ocultos de Markov para 37 famílias de enzimas. Uma ferramenta para prever função de enzima com base em alinhamentos global e local de sequência foi proposta em [Hung et al. (2010)]. Em [Quester e Schomburg (2011)], anotações de função enzimática de algumas bases de dados foram comparadas e avaliadas. Finalmente, em [Furnham et al. (2012)] uma ferramenta que combina dados filogenéticos, funcionais, de estrutura e sequência foi apresentada e tais dados podem ajudar a elucidar a evolução de funções enzimáticas apoiando a previsão de função para enzimas ainda não caracterizadas.

Conforme mencionado, os repositórios biológicos armazenam metadados que caracterizam e dão contexto biológico aos dados brutos. Seriam tais metadados capazes de indicar que uma mudança de anotação irá ocorrer? Em caso afirmativo, como esses metadados podem ser processados para capturar essa informação e prever uma mudança de anotação?

Nesse trabalho é proposta uma estratégia baseada em aprendizagem supervisionada para caracterizar e prever mudanças de anotação em dados temporais denominada *ENZYmatic Metadata Annotation Predictor* (ENZYMAP). Mais precisamente, estamos interessados em prever a anotação de função de enzima com base em metadados das entradas do repositório UniProt/Swiss-Prot[Consortium et al. (2012)]. Essa proposta permite sugerir

possíveis correções para anotações e pode ser utilizada de modo complementar a outros métodos de anotação de modo a aprimorar a qualidade e confiabilidade do repositório utilizando dados já disponíveis, o que não demanda experimento de bancada. Além disso, há um enorme volume de dados que não pode ser analisado manualmente, daí a importância de métodos de anotação automática confiávies.

Neste capítulo são apresentados os conceitos básicos necessários ao entendimento do trabalho. São introduzidos os conceitos de bases dados e anotação, seguidos de uma breve apresentação do repositório de dados biológicos UniProt. Na sequência, são abordadas brevemente as enzimas e sua classificação EC e, finalmente, é introduzido o conceito de Mineração de Dados.

1.1 Bases de Dados

Uma base de dados é uma coleção de dados relacionados. Dados são fatos conhecidos que podem ser armazenados e que possuem significado implícito [Elmasri e Navathe (2008)]. Mais especificamente, ainda segundo [Elmasri e Navathe (2008)], uma base de dados deve ter algumas propriedades:

- É uma representação de alguns aspectos do mundo real, também chamado de *universe of discourse* UoD, e mudanças no UoD devem ser refletidas nessa representação.
- É um conjunto de dados que possui significado. O termo base de dados não é usado para referenciar um conjunto aleatório de dados.
- Possui um propósito e um grupo de usuários interessados nas possíveis aplicações da base de dados. Ela é modelada, construída e populada para esse propósito específico.

Nas últimas décadas houve um grande aumento na quantidade de dados biológicos gerados por técnicas experimentais [Luscombe et al. (2001)]. As novas tecnologias de sequenciamento de DNA [Ansorge (2009)], bem como o sequenciamento de proteínas e a determinação da estrutura secundária [Rost et al. (2004)] e terciária [Otwinowski e Minor (1997)] das mesmas foram responsáveis pela geração de uma enorme massa de dados. Devido a isso, tornou-se indispensável o armazenamento desses dados de um modo estruturado e confiável, que permitisse sua recuperação, análise e ainda sua integração com outros dados. Para atender a essa necessidade, foram utilizadas as tecnologias de bases de dados, o que colocou à disposição da Biologia um ferramental consolidado da área computacional cujos estudos iniciais datam da década de 1970 [Codd (1970)].

Existem diversas bases de dados que disponibilizam dados biológicos de diversos tipos publicamente através da Internet. Além desses dados, as bases podem armazenar também dados da literatura e as mais variadas anotações (como função de proteínas ou mesmo

relacionar uma proteína a seu gene codificante, dentre outros) [Stein (2003)]. São essas anotações que conferem significado e valor aos dados pois, por exemplo, uma sequência de nucleotídeos não é de grande utilidade até que sejam identificadas as suas codificações funcionais.

Dentre as grandes bases de dados disponíveis na Internet podemos citar o *Protein Data Bank* (PDB) [Berman et al. (2000)], GenBank [Benson et al. (2011)], *DNA Data Bank of Japan* (DDBJ) [Ogasawara et al. (2012)], *European Nucleotide Archive* (ENA) [Leinonen et al. (2011)], *Structural Classification of Proteins (SCOP)* [Murzin et al. (1995)] (CATH) [Orengo et al. (1997)], *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [Kanehisa e Goto (2000), Kanehisa et al. (2012)], *Universal Protein Resource* (UniProt) [Consortium et al. (2012)], MEROPS [Rawlings et al. (2012)], *BRaunschweig ENzyme Database* (BRENDA) [Scheer et al. (2011)]. Cada uma dessas bases armazena um determinado tipo de dado, como pode ser visto na Tabela 1.1.

Tabela 1.1: Exemplos de bases de dados biológicos.

Base de dados	Tipo de dado
PDB	Estruturas de proteínas, ácidos nucleicos e complexos.
GenBank	Sequências de nucleotídeos
DDBJ	Sequências de nucleotídeos
ENA	Sequências de nucleotídeos
SCOP	Classificação estrutural de proteínas (famílias)
CATH	Classificação estrutural de proteínas (domínios)
KEGG	Subdividido em informação sistêmica, genômica e química.
UniProt	Sequências e funções de proteínas.
MEROPS	Proteases, seus inibidores e substratos
BRENDA	Enzimas anotadas manualmente.

1.2 Anotação

De acordo com [Lesk (2005)], anotar uma sequência de nucleotídeos ou aminoácidos é determinar suas características biológicas nos seguintes níveis: processos moleculares e celulares, tecidos, órgãos e processos fisiológicos. Consiste, portanto, num processo de conferir semânticas, contextos, relações, história, proveniência e outras informações aos dados brutos biológicos, como sequências, estruturas, reações, vias metabólicas, dentre outros.

As anotações podem ser feitas utilizando texto livre ou um vocabulário controlado, como EC *number*, que será abordado ainda nesse capítulo, sendo que um vocabulário controlado amplamente conhecido pela comunidade científica permite melhor compartilhamento dos dados entre diferentes grupos de pesquisa. Ainda segundo [Lesk (2005)], há

coleções primárias de dados - com anotaçõs feitas pelos autores que submeteram os dados, como o PDB - e coleções secundárias - que são derivadas das primárias por outros grupos de pesquisa e possuem mais informações biológicas, como o UniProtKB/Swiss-Prot.

Tomemos como exemplo a entrada do UniProtKB/Swiss-Prot, cujo identificador é P07477. A Figura 1.1 mostra um pequeno subconjunto das anotações disponibilizadas para a enzima nessa base de dados. Vemos, dentre os vários atributos de anotação, a informação de que o nome recomendado dessa enzima é Trypsin-1, que ela possui três nomes alternativos, que seu EC *number* é 3.4.21.4 e que ela é encontrada na espécie humana (*Homo sapiens*).

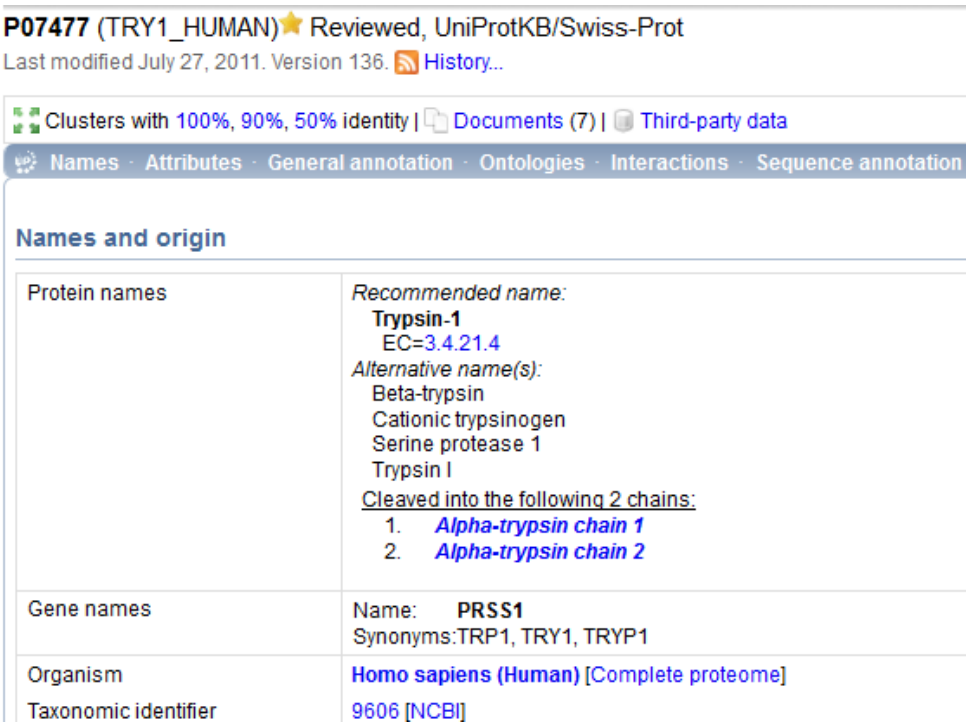


Figura 1.1: Alguns exemplos de atributos de anotação do UniProtKB/Swiss-Prot.

1.3 Uniprot

O *Universal Protein Resource* (UniProt) é o mais completo catálogo de sequências protéicas e anotação funcional para as mesmas. É uma base de dados estável, completa, classificada, rica e cuidadosamente anotada, com interface de consulta intuitiva e referências cruzadas (para um amplo conjunto de bases de dados biológicos), disponibilizada livremente para a comunidade científica [Consortium et al. (2012)]. Sua atualização acontece a cada quatro semanas, quando é lançada uma nova versão. Nesse meio tempo não acontecem atualizações na base. A versão atual, bem como um conjunto de versões históricas do Uniprot estão disponíveis para *download* em <`http://www.uniprot.org`>.

A primeira versão do UniProt foi lançada em dezembro de 2003 como resultado da criação do *UniProt Consortium*, que surgiu da união das bases de dados Swiss-Prot [Bo-

eckmann et al. (2003)], TrEMBL [Boeckmann et al. (2003)] e PIR [Wu et al. (2003)]. De acordo com [Consortium (2011)], o UniProt possui quatro principais componentes:

- *UniProt Archive* (UniParc): é a mais completa coleção de sequências não redundantes, oferecendo uma cobertura completa das sequências protéicas publicamente disponíveis nas mais diversas bases de dados. Contém apenas sequências e referências cruzadas, demais dados devem ser obtidos das bases de origem [Leinonen et al. (2004)].

- *UniProt Knowledgebase* (UniProtKB): repositório de sequências protéicas e anotações para as mesmas. Possui duas partes.

  UniProtKB/Swiss-Prot: contém dados anotados manualmente, resultado de extração de informações da literatura e análise computacional manualmente revisada por um especialista.

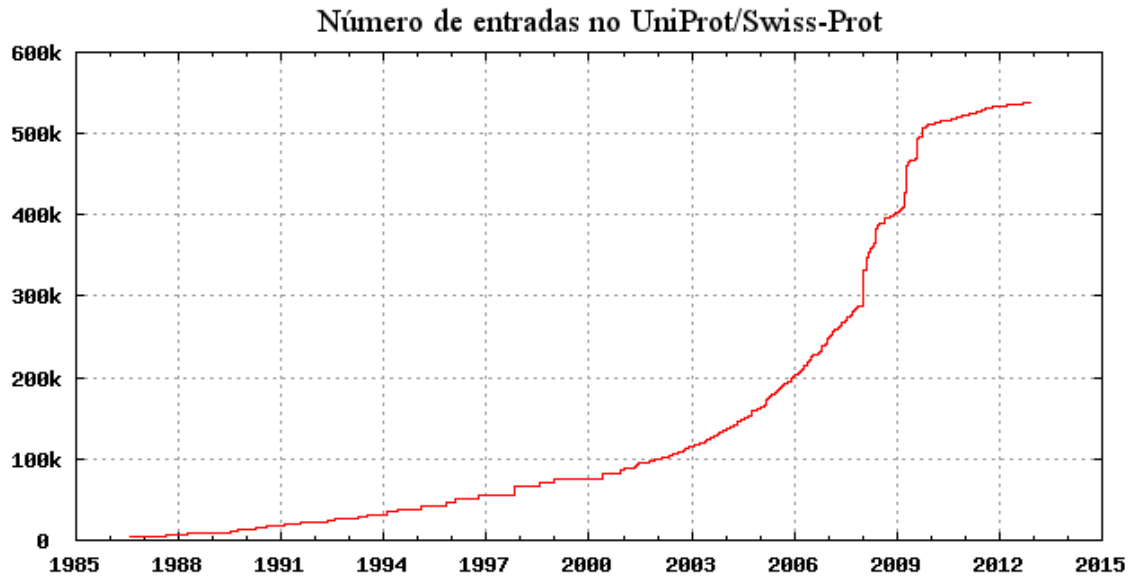
  UniProtKB/TrEMBL: dados analisados computacionalmente, que ainda carecem de revisão manual.

- *UniProt Reference Clusters* (UniRef): conjuntos de sequências similares agrupadas, facilitando as buscas por similaridade [Suzek et al. (2007)].

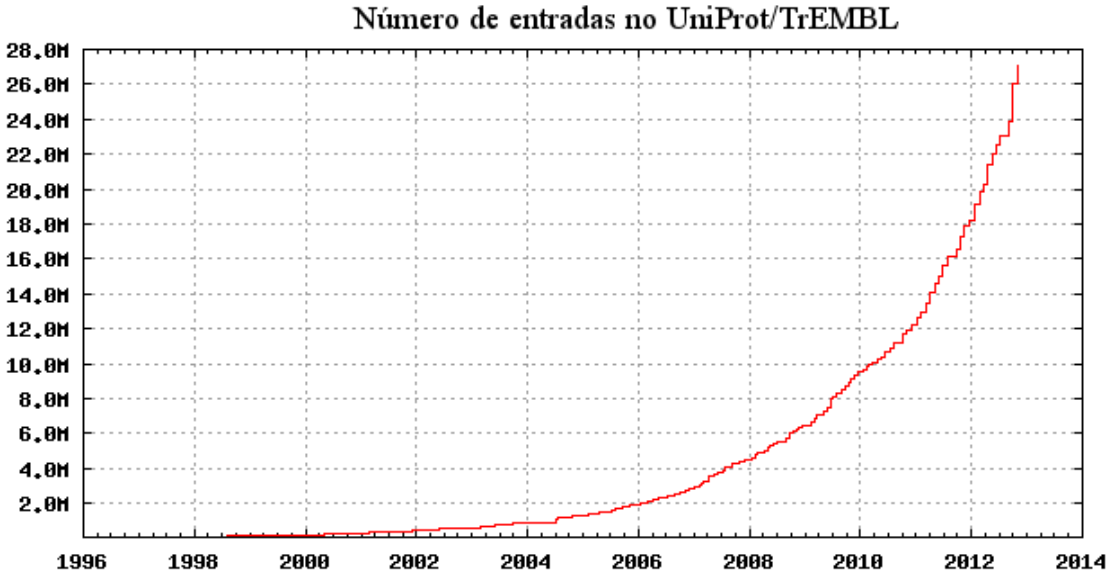
- *UniProt Metagenomic and Environmental Sequence Database* (UniMes): repositório de dados específico para dados metagenômicos e ambientais.

O UniProtKB é a peça chave do *UniProt Consortium*. A figura 1.2 ilustra o crescimento do UniProtKB/Swiss-Prot e do UniProtKB/TrEMBL até a versão 2012_11 do UniProtKB. Ele atua como ponto de acesso central para informações biomoleculares, pois está conectado, através de referências cruzadas, a mais de 140 bases de dados com informações sobre estrutura protéica, sequências de nucleotídeos, famílias e domínios de proteínas, entre outros. Para criar e manter essas referências há a colaboração com a comunidade científica e com desenvolvedores de outros repositórios para garantir que elas estejam atualizadas e confiáveis. Outra ferramenta essencial para possibilitar a interoperabilidade de bases de dados heterogêneas é o mapeamento de identificadores. Em bases de dados diferentes, uma mesma entidade biológica pode ter identificadores distintos. Para contornar essa situação, o UniProt fornece um serviço de mapeamento para mais de 100 tipos de identificadores além de disponibilizar suas tabelas de mapeamento para *download*.

É importante mencionar que o UniProtKB/Swiss-Prot é considerada uma base de dados padrão ouro para anotação de proteínas, pelo fato de ser curada e anotada manualmente. Num estudo sobre erros na anotação de enzimas, a base UniProtKB/Swiss-Prot foi considerada a mais bem anotada. Em quatro das seis superfamílias estudadas, o percentual de erro de anotação foi 0% [Schnoes et al. (2009)]. Ainda assim, isso não significa



(a)



(b)

Figura 1.2: Gráfico representativo do crescimento da base de dados UniProtKB em dezembro de 2012. (a) UniProtKB/Swiss-Prot, imagem obtida em (http://web.expasy.org/docs/relnotes/relstat.html), (b) UniProtKB/TrEMBL, imagem obtida em (http://www.ebi.ac.uk/uniprot/TrEMBLstats/)

.

que essa base seja livre de erros. De acordo com o mesmo estudo, dentre as 50 enzimas da família *Adenosine deaminase* que foram analisadas, 70% estavam incorretamente anotadas.

Já o *Translation of EMBL Nucleotide Sequence Database* (UniProtKB/TrEMBL) apresentou níveis de erros de anotação entre 25% e 60% em quatro das seis superfamílias estudadas em Schnoes et al. (2009). Essa base possui maiores níveis de erros que o UniProtKB/Swiss-Prot devido ao seu processo de anotação. Dado o rápido aumento no número de sequências disponíveis, não é possível fazer a anotação com revisão manual, seguindo os padrões de qualidade do UniProtKB/Swiss-Prot, para todas as sequências. Para tratar essa questão, existe o UniProtKB/TrEMBL, que é um conjunto de sequências computacionalmente anotadas derivadas de bases de dados de nucleotídeos.

1.4 Enzimas e Classificação EC

As enzimas, consideradas as mais notáveis e especializadas proteínas, são as catalisadoras de reações químicas dos sistemas biológicos. Com exceção de um pequeno grupo de moléculas de RNA catalítico, todas as enzimas são proteínas. Elas catalisam diversas reações que degradam moléculas de nutrientes, conservam e transformam energia química e produzem macromoléculas a partir de simples precursores. Seu estudo é de grande importância prática. Algumas doenças, especialmente as genéticas, se devem à deficiência ou ausência de uma ou mais enzimas. Sua atividade excessiva também pode ser prejudicial. As medidas da atividade de enzimas no sangue, plasma ou tecidos, podem ajudar no diagnóstico de enfermidades. Muitos medicamentos agem através de interações com enzimas. Elas são ainda importantes ferramentas na engenharia química, tecnologia de alimentos e agricultura [Lehninger et al. (2008)].

Algumas enzimas não precisam de substâncias adicionais para desempenhar sua atividade. Outras necessitam de cofatores, que são substâncias orgânicas ou inorgânicas necessárias para o funcionamento de uma enzima. Um cofator orgânico é chamado de coenzima. Sob condições normais, muitas reações químicas aconteceriam lentamente e em pequeno número num organismo vivo. Para contornar o problema, uma enzima gera um ambiente no qual uma reação pode ocorrer de modo mais rápido. Tal reação acontece num "bolso" da enzima, chamado sítio ativo. A molécula que se liga ao sítio ativo e sobre a qual a enzima atua é chamada substrato e existe certa especificidade entre uma enzima e seu substrato. Para que uma reação aconteça, é necessária uma determinada energia de ativação. O que a enzima faz é diminuir a energia de ativação de uma reação, aumentando as taxas em que dita reação acontece, porém mantendo o equilíbrio da mesma (uma mesma quantidade de reagentes gera a mesma quantidade de produtos).

No UniProtKB, mais especificamente no UniProtKB/Swiss-Prot (que possui dados detalhados e curados), há varios tipos de anotações para as enzimas. Dentre esses podemos citar a atividade catalítica, cofatores, vias metabólicas, mecanismos de regulação, doenças

associadas à deficiência enzimática, estágios do desenvolvimento nos quais a enzima está presente no organismo, conflitos na sequência de aminoácidos e variantes [Apweiler et al. (2004b), Apweiler et al. (2004a)]. Para essas anotações, há um esforço no sentido de utilizar um vocabulário controlado, que possa representar as particularidades e detalhes das entidades no mundo real, nesse caso as enzimas, possibilitando assim que um mesmo termo possa ser utilizado pela comunidade científica com uma semântica bem definida e clara. Isso permite que dados sejam compartilhados entre diferentes grupos de pesquisa e que os especialistas possam utilizar as informações já conhecidas e disponíveis para fazer novas análises e chegar a novas conclusões e resultados. Um dos sistemas de classificação que atende a esses critérios e é largamente utilizado no UniProt Swiss-Prot para anotação de função enzimática é o EC *number*.

O *Enzyme Commission* (EC) *number* [NC-IUBMB (1999)] é um sistema numérico e hierárquico de classificação de enzimas, amplamente conhecido e utilizado, estabelecido pela IUBMB (*International Union of Biochemistry and Molecular Biology*) e que baseia-se nas reações químicas catalisadas pelas enzimas. Um EC *number* possui o formato #.#.#.#, onde cada # representa um número e, da esquerda para a direita, cada número fornece progressivamente mais detalhes sobre a reação enzimática. Esse sistema define quatro níveis de profundidade para classificação das enzimas, sendo que no nível mais alto da hierarquia (número mais à esquerda) há seis categorias: (1) Oxidoredutases, (2) Transferases, (3) Hidrolases, (4) Liases, (5) Isomerases e (6)Ligases [NC-IUBMB (1999)].

Tomemos como exemplo do uso desse sistema de classificação, o EC *number* 3.4.21.4. O primeiro dígito (3) nos informa que essa enzima é uma hidrolase (responsável pela ruptura de uma ligação química envolvendo uma molécula de água); o segundo (4) agrega a informação de que ela é uma peptidase (rompe ligações peptídicas); o terceiro (21) nos diz que é uma endopeptidase (quebra ligações peptídicas em aminoácidos que não sejam os terminais) e que possui uma serina no sítio ativo; o quarto (4) nos informa que é uma tripsina (quebra as ligações peptídicas preferencialmente após os resíduos de arginina e lisina).

Um EC *number* caracteriza uma reação química, desse modo, um mesmo EC *number* pode estar associado a diferentes enzimas que catalisam uma mesma reação, como é o caso da Hexokinase-2 em *Saccharomyces cerevisiae* (identificador P04807 no UniProtKB/Swiss-Prot) e da Hexokinase-1 em *Homo sapiens* (identificador P19367 no UniProtKB/Swiss-Prot), que possuem EC *number* 2.7.1.1. Uma determinada enzima também pode estar associada a mais de um EC *number* se catalisa reações distintas, como é o caso da enzima humana com identificador P12821 do UniProtKB/Swiss-Prot, que catalisa reações com os EC *numbers* 3.2.1.- e 3.4.15.1.

1.5 Mineração de Dados

Abordaremos aqui o conceito de Mineração de Dados segundo [Tan et al. (2006)]. Mineração de Dados é o processo de extrair padrões novos e relevantes de modo automático em grandes repositórios de dados com o objetivo de extrair conhecimento a partir dos dados e apresentá-lo numa estrutura interpretável. É parte do processo de *knowledge discovery in databases* (KDD) ou descoberta de conhecimento em bases de dados, que consiste em uma série de transformações, que vão do pré processamento ao pós processamento dos resultados da mineração, conforme esquematizado em 1.3.

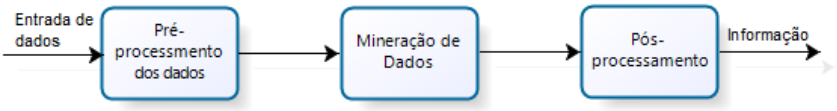


Figura 1.3: O processo de descoberta de conhecimento em bases de dados, adaptado de [Tan et al. (2006)]

.

A tarefa de pré-processamento tem o objetivo de preparar os dados brutos para a subsequente tarefa de mineração, de modo a remover ruído, selecionar registros e atributos (também chamados de características) relevantes, reduzir a dimensionalidade, dentre outros. O pós-processamento deve assegurar que resultados válidos e úteis sejam considerados. Exemplos de pós-processamento são visualização e medidas estatísticas que permitem explorar os resultados da mineração de diversos pontos de vista.

Um dos desafios que motivou o desenvolvimento da área foi a necessidade de analisar dados de modo não tradicional. A abordagem estatística tradicional é baseada no paradigma hipótese-teste, no qual uma hipótese é proposta, um experimento é definido para coletar dados e esses são analisados com relação à hipótese. As tarefas de análise de dados atuais frequentemente demandam geração e avaliação de milhares de hipóteses e, consequentemente, o desenvolvimento de algumas técnicas de Mineração de Dados foi motivado pelo desejo de automatizar o processo de geração e avaliação de hipóteses. Além disso, os conjuntos de dados analisados através de técnicas de mineração são muitas vezes amostras associadas à oportunidade em dado domínio e não amostras aleatórias. Tais conjuntos de dados comumente envolvem tipos de dados e distribuições não tradicionais.

As tarefas de Mineração de Dados podem ser divididas em duas grandes categorias:

- Tarefas de previsão, cujo objetivo é prever o valor do atributo de interesse, chamado de alvo ou variável dependente, com base nos valores de outros atributos, chamados explicativos ou variáveis independentes.

- Tarefas descritivas, cujo propósito é derivar padrões capazes de resumir os relacionamentos subjacentes presentes nos dados.

1.5.1 Redução de Dimensionalidade

A redução de dimensionalidade é uma tarefa de pré-processamento, ou seja, é realizada antes da tarefa de Mineração de Dados propriamente dita e procura reduzir a dimensionalidade do conjunto de dados original, ou seja, reduzir o número de atributos ou características através da criação de atributos novos que são uma combinação dos atributos originais [Tan et al. (2006)].

De acordo com [Han e Kamber (2006)], a redução da dimensionalidade pode trazer benefícios. Um deles é que, em geral, os algoritimos de mineração funcionam melhor quando a dimensionalidade é menor. Isso porque a redução da dimensionalidade pode eliminar atributos irrelevantes e reduzir ruído e também devido ao problema da dimensionalidade[2]. Adicionalmente, a redução de dimensionalidade diminui os requisitos de tempo e memória do algoritmo de mineração.

Para reduzir a dimensionalidade, existem várias técnicas, porém, em muitos casos, após sua aplicação existirá um número menor de atributos que serão diferentes dos atributos originais, mas igualmente válidos. Alguns exemplos de tais técnicas são *Singular Value Decomposition* (SVD), *Principal Component Analysis* (PCA) e *Locally Linear Embedding* (LLE).

1.5.2 Classificação

Classificação é uma técnica de Mineração de Dados que consiste em associar um dentre vários rótulos ou categorias pré-definidas a objetos de dados. Tais categorias são chamadas de classes. Um modelo de classificação pode ser visto como uma função f que mapeia um conjunto de atributos x para uma determinada classe.

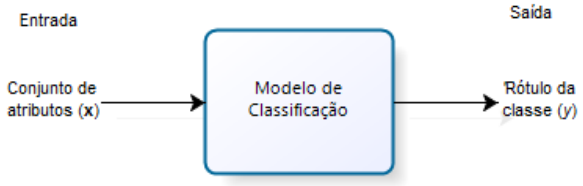


Figura 1.4: Classificação vista como a tarefa de mapear um conjunto de atributos de entrada para as classes às quais pertencem, adaptado de [Tan et al. (2006)]

.

Um classificador é uma abordagem para construção de modelos de classificação a partir de dados de treinamento fornecidos como entrada. Cada instância dos dados de treino pode ser vista como uma tupla da forma (x, y) onde x representa um conjunto de atributos ou características e y representa a classe associada à essa tupla. Uma tupla

[2]O problema da dimensionalidade é referente ao fenômeno de que vários tipos de análises de dados se tornam mais difíceis à medida que aumenta a dimensionalidade dos dados. Com o aumento da dimensionalidade, os dados vão se tornando muito dispersos no espaço, o que prejudica, por exemplo, algoritmos de agrupamento e classificação.

pode ser, por exemplo, uma linha da Tabela 1.2. Nessa tabela é mostrado um exemplo de conjunto de dados que poderia ser utilizado por um classificador. Nesse caso, entradas da base de dados UniProtKB/Swiss-Prot podem pertencer à classe mudança (para as entradas que sofreram a mudança de EC *number* 3.1.3.2 → 3.1.3.5) ou à classe controle (para as entradas que se mantiveram como 3.1.3.2), dependendo dos atributos que estejam associados a cada entrada.

Tabela 1.2: Dados referentes à entradas do UniProtKB/Swiss-Prot que experimentaram a mudança de EC *number* 3.1.3.2 → 3.1.3.5 ou se mantiveram 3.1.3.2

id	nucleotide-binding	magnesium	eukaryota	metal-binding	signal	classe
Q8TUG3	sim	sim	não	sim	não	mudança
O67004	sim	sim	não	sim	não	mudança
Q9HY05	sim	sim	não	sim	não	mudança
P58683	não	sim	não	sim	sim	controle
P34724	não	não	sim	não	sim	controle
P44009	não	sim	não	sim	sim	controle

Um conjunto de treino, contendo exemplos das várias classes, deve ser fornecido para que o classificador possa aprender com esses dados e posteriormente tentar prever as classes para um conjunto de dados de teste. Os dados de teste são instâncias que não foram utilizadas para a construção do modelo e para os quais os rótulos são conhecidos de modo a permitir uma avaliação do desempenho do classificador, ou seja, quão bem ele pode classificar instâncias novas. Dizemos que um classificador possui boa capacidade de generalização quando é capaz de prever corretamente as classes para dados que não participaram da construção do modelo.

Existem diversas técnicas de classificação e cada uma utiliza um determinado algoritmo de aprendizagem para definir um modelo que melhor se ajuste ao conjunto de atributos e classes fornecidos como treinamento. Como exemplo podemos citar Árvores de Decisão, Redes Neurais, Naïve Bayes, *K-Nearest Neighbor* (KNN) ou K vizinhos mais próximos, *Support Vector Machine* (SVM) e classificadores baseados em regras. É importante pontuar que não existe um classificador que seja o melhor para todos os problemas de classificação. A relação entre o problema a ser resolvido, ou seja, os dados a serem classificados, e o desempenho dos algoritmos de classificação é um tópico em estudo [Garg e Roth (2003)], [Tang et al. (2006)].

Um modelo de classificação pode ser utilizado com objetivo de descrição ou previsão.

- *Descrição*: nesse caso o modelo atua como uma ferramenta que ajuda a explicar como são discriminados os objetos de diferentes classes. Como exemplo podemos citar o modelo construído nesse trabalho para verificar se alguns metadados selecionados dos arquivos texto do Swiss-Prot são capazes de discriminar entradas

que sofreram determinada mudança de EC das entradas em que o EC se manteve constante. Tal modelo é detalhado na Seção 4.3.4.1;

- *Previsão*: nesse caso o classificador é utilizado para prever classes para dados desconhecidos, que não foram utilizados na construçao do modelo. Um exemplo são os modelos contruídos nesse trabalho com o propósito de utilizar o conhecimento já disponível no repositório Swiss-Prot a respeito das mudanças de EC para prever tais mudanças numa versão posterior do repositório. Maiores detalhes sobre esses modelos podem ser encontrados nas Seções 4.3.4.2 e 4.3.4.3.

1.6 Motivação do Trabalho

A geração de dados biológicos experimentou um crescimento sem precedentes nas últimas décadas. Fatores como otimização e diminuição do custo de diversos processos laboratoriais associados às novas técnicas de sequenciamente de DNA, o sequenciamento de proteínas e a determinação de sua estrutura foram responsáveis pela geração de uma grande quantidade de dados [Luscombe et al. (2001)]. Muitos deles estão em bases de dados biológicos disponibilizadas publicamente através da Internet. Essas bases armazenam não apenas os dados propriamente ditos, mas também várias outras informações relevantes relacionadas a eles, como dados da literatura, função de proteína, relação entre uma proteína e seu gene codificante, entre outros [Lesk (2005)]. Tais dados são chamados de anotação.

Em geral, os repositórios de dados biológicos são volumosos, heterogêneos, dinâmicos e mantidos de forma independente, cada um com seu próprio padrão de modelagem, armazenamento, acessibilidade e evolução. Em muitos casos, mudanças silenciosas ocorrem sem aviso prévio, e nem mesmo um histórico de versões é disponibilizado [Buneman et al. (2006)]. Manter a integridade e sincronia de dados neste contexto é certamente um grande desafio enfrentado pela Bioinformática atual.

Nesse cenário, surge uma grande preocupação da comunidade científica, que é com relação à qualidade e confiabilidade dos dados e anotações das grandes bases de dados disponibilizadas publicamente [Dall'Olio et al. (2010), Schnoes et al. (2009), Naumoff et al. (2004), Jones et al. (2007), Brenner et al. (1999), Devos e Valencia (2001), Green e Karp (2005), Gilks et al. (2005), Hung et al. (2010), Egelhofer et al. (2010), Quester e Schomburg (2011)]. Pesquisadores utilizam esses dados para realizar estudos e análises em larga escala. Além disso, muitas das bases de dados são integradas em menor ou maior grau, o que vai desde um *hiperlink* que conecta um dado em uma base ao seu correspondente em outra , até uma cópia de dados de uma ou mais bases seguida de algum tipo de processamento, originando uma nova base de dados. Assim, um dado ou anotação incorreto poderia comprometer os resultados de diversos trabalhos científicos ou, ainda pior, ser propagado entre as diversas bases de dados.

Dessa maneira, uma proposta que permita prever mudanças de anotação em repositórios de dados biológicos seria uma importante contribuição à Bioinformática. Nesse trabalho propomos uma estratégia de aprendizagem supervisionada para caracterizar e prever mudanças de anotação EC no repositório UniProt/Swiss-Prot com base em metadados das entradas de tal repositório. No decorrer desse trabalho não foram encontradas no nosso levantamento bibliográfico técnicas capazes de prever mudanças de anotação com base em metadados protéicos. Essa estratégia foi denominada *ENZYmatic Metadata Annotation Predictor* (ENZYMAP) e permite sugerir possíveis correções para as anotações EC, podendo ser utilizada de modo complementar a outros métodos de anotação, ajudando a aprimorar a qualidade e confiabilidade do repositório usando dados já disponíveis, o que não demanda experimento de bancada. Além disso, há um enorme volume de dados que não pode ser analisado manualmente, daí a importância de métodos de anotação automática confiávies.

Um fenômeno comum em repositórios biológicos é que, dado que uma correção foi feita, esse conhecimento não necessariamente é propagado para as demais entradas de uma única vez, mas sim gradual e lentamente. Nossa proposta pode apoiar na sugestão de correções da base de dados, propagando o conhecimento implícito presente na base para todas as entradas.

1.7 Contribuições do Trabalho

A seguir são descritas as principais contribuições do presente trabalho:

- Artigo com resultados dessa tese, intitulado *ENZYMAP: Exploiting protein metadata for modeling and predicting annotation changes in UniProt/Swiss-Prot*, foi submetido à revista Bioinformatics (Oxford) e pode ser visto no Apêndice C.

- Artigo intitulado *Advise: Visualizing the dynamics of enzyme annotations in UniProt/Swiss-Prot* publicado no evento *IEEE Symposium on Biological Data Visualization (BioVis), 2012* realizado em Seattle, EUA. Esse trabalho pode ser visto no Apêndice B e é resultado de uma exploração inicial das mudanças de anotação EC descrita na Seção 4.2.1.

- ENZYMAP: estratégia baseada em aprendizagem supervisionada capaz de prever mudanças de anotação EC em dados temporais do repositório biológico UniProt/Swiss-Prot com base em metadados presentes nas entradas de tal repositório. Em nosso levantamento bibliográfico não foram encontrados trabalhos que utilizam metadados de repositórios biológicos para prever mudanças na anotação de proteínas. Tal estratégia:

  Utiliza dados já disponíveis no repositório para fazer as previsões, o que não demanda novos experimentos de bancada;

Antecipa mudanças na base de dados, sugerindo alterações de anotação EC tão logo os metadados indiquem essa possibilidade;

Pode ser utilizada de modo complementar com técnicas de previsão de função de enzima baseadas em sequência e estrutura;

- ADVISe: ferramenta de visualização interativa que permite explorar as mudanças de anotação EC ao longo de diversas versões do repositório <`https://github.com/arturhoo/ADVISe`.

1.8 Organização do Texto

No Capítulo 2, Revisão da literatura, fez-se um levantamento bibliográfico de trabalhos correlatos e foram discutidos o sistema de classificação EC e o GO. No Capítulo 3, Objetivos, foram listados os objetivos geral e específicos do trabalho. O capítulo 4, Materiais e métodos, descreve a metodologia e técnicas utilizadas nas análises. Os resultados são apresentados e discutidos no Capítulo 5. O Capítulo 6 apresenta as conclusões e possíveis desdobramentos futuros para esse trabalho.

Capítulo 2

Revisão da literatura

Nesse capítulo serão revisados o sistema EC de classificação de enzimas e o *Gene Ontology*, pois ambos podem ser utilizados para anotação de função catalítica de enzimas, sendo amplamente conhecidos e adotados. No decorrer desse projeto, não foram encontrados no nosso levantamento bibliográfico técnicas capazes de prever mudanças de anotação do tipo EC *number* em bases de dados temporais, assim abordaremos alguns trabalhos que tratam dos níveis de erros de anotação em repositórios biológicos e, de modo mais amplo, do problema de anotação em tais repositórios. Acreditamos que tais temas sejam correlatos e relevantes para a nossa proposta.

2.1 Sistema de Classificação de Reações Enzimáticas

2.1.1 *Gene Ontology* (GO)

O *Enzyme Comission number* não é o único sistema de classificação para reações enzimáticas, embora seja o mais amadurecido e consolidado. Uma alternativa a esse sistema seria o *Gene Ontology* (GO) [Ashburner et al. (2000)], criado pelo *Gene Ontology Consortium*, que nasceu como um projeto conjunto de três bases de dados de organismos modelo, o FlyBase [Tweedie et al. (2009)], o *Mouse Genome Informatics* (MGI) [Blake et al. (2011)] e o *Saccharomyces Genome Database* (SGD) [Engel et al. (2010)] .Trata-se de uma iniciativa da área de Bioinformática que tem o objetivo de padronizar a representação de genes e dos atributos dos produtos de genes entre diferentes espécies e bases de dados. Para isso, provê um vocabulário estruturado, controlado e classificações que abrangem diversos domínios da biologia molecular e celular e é disponibilizado livremente para que seja utilizado pela comunidade científica para anotação de genes, seus produtos e também sequências [Harris et al. (2004)]. É importante dizer que o GO não é um sistema específico para classificação de reações enzimáticas. Ele é mais geral, porém contempla, dentre vários outros tipos de anotações, a função catalítica.

Segundo [Ashburner et al. (2001)], o GO estrutura um amplo conhecimento biológico através de ontologias. Uma ontologia represena formalmente o conhecimento como um

conjunto de conceitos dentro de um determinado domínio e relações entre esses conceitos. De acordo com [Gruber et al. (1995)], uma ontologia é uma especificação explícita de uma conceitualização.

O GO inclui três ontologias, que foram definidas porque representam conjuntos de informação comuns para as formas de vida e servem como base para a anotação de genes e seus produtos em três domínios não sobrepostos da biologia molecular. Em cada uma delas os termos possuem definições em formato texto e identificadores únicos e estáveis. Segundo [Harris et al. (2004)] as ontologias do GO são:

- *Molecular Function* (MF), que descreve atividades catalíticas ou de ligação em nível molecular. Os termos de MF representam as atividades e não as entidades responsáveis pela ação. Além disso, não especificam onde, quando ou em que contexto uma ação ocorre. Alguns exemplos de termos de MF são *kinase activity* (mais geral) e *6-phosphofructokinase activity*, que representa um subtipo do anterior.

- *Biological Process* (BP), que descreve objetivos biológicos realizados através de um ou mais conjuntos ordenados de funções moleculares. Processos de mais alto nível como *cell death* podem ter subtipos (como *apoptosis*) e subprocessos (como *apoptotic chromosome condensation*).

- *Cellular Component* (CC), que descreve localizações no nível de estruturas subcelulares e complexos macromoleculares. Exemplos de termos que fazem parte de CC são *nuclear inner membrane* e *ubiquitin ligase complex*.

Na figura 2.1 observa-se um exemplo de anotação de enzima com termos do GO das três ontologias. Trata-se do identificador Q8RXD9 do UniProt/Swiss-Prot que, de acordo com a classificação EC, é anotado da seguinte maneira: 2.4.1.25 (*4-alpha-glucanotransferase*). Vemos, para o mesmo identificador do UniProt/Swiss-Prot, um esquema para os conceitos da ontologia MF, que estão em níveis superiores ao termo *4-alpha-glucanotransferase activity* (um dos termos do GO usados para anotar Q8RXD9).

Ainda de acordo com [Harris et al. (2004)], cada anotação do GO consiste de um termo do GO associado a uma referência ao trabalho ou análise no qual se baseia a associação de tal termo com o produto de um gen. Cada anotação precisa também incluir um *evidence code* para indicar em que tipo de evidência uma anotação se baseia. Podemos dizer que esses *evidence codes* são uma espécie de proveniência de dados, pois nos dão informação a respeito do processo de derivação dos mesmos.

Apesar de ser uma ferramenta importante e largamente utilizada, o GO possui algumas limitações:

- É organizado como um grafo direcionado acíclico, estrutura hierárquica similar a uma árvore, porém permite que um termo tenha mais de um pai. Essa estrutura pode não ser a mais adequada, pois tende a poluir a visualização [Zeeberg et al. (2003)].

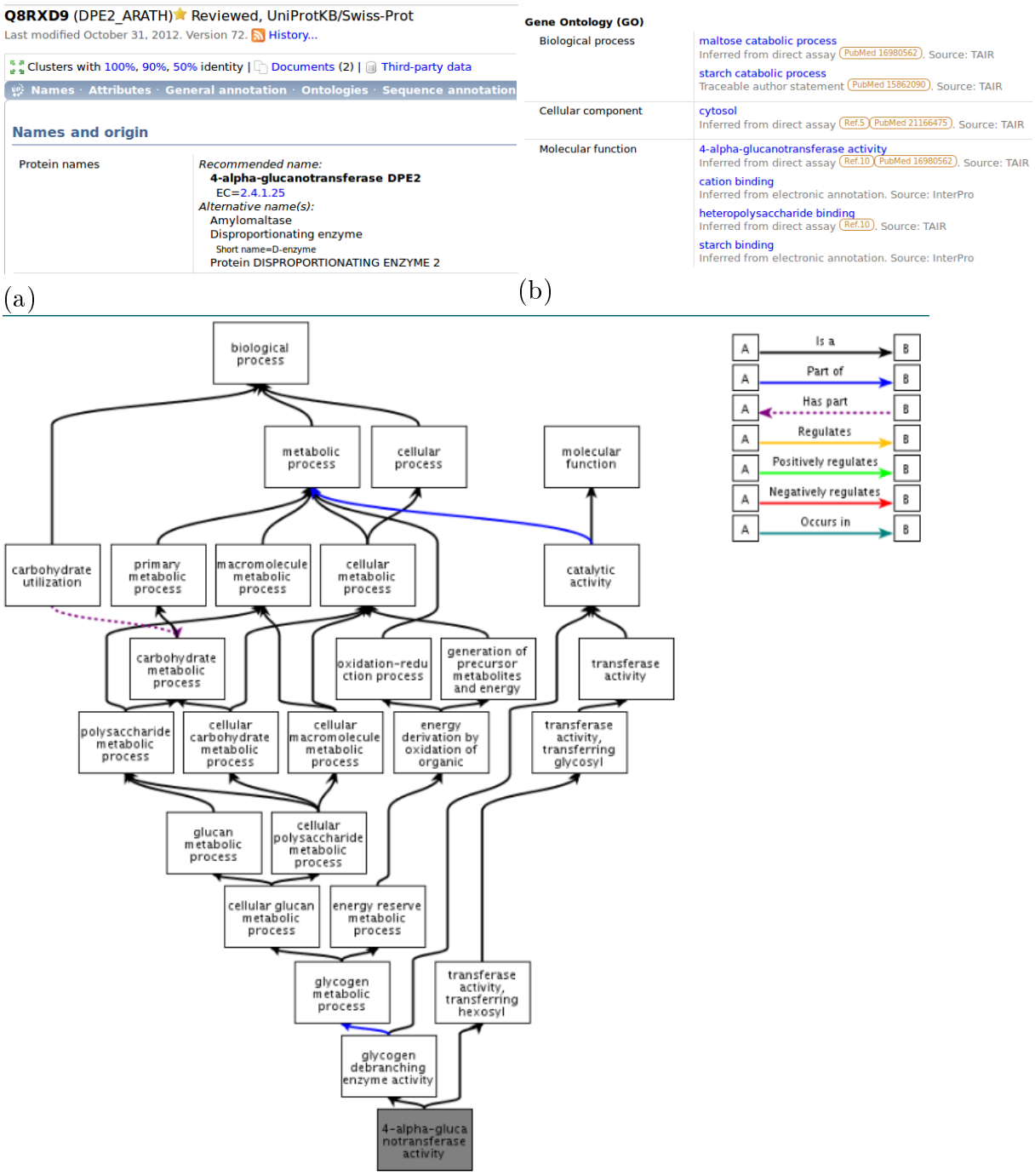


Figura 2.1: Anotação do tipo EC *number* e GO para entrada Q8RXD9 do UniProt/Swiss-Prot. (a) Anotação do tipo EC, (b) Anotação do tipo GO e (c) Conceitos da ontologia MF superiores ao termo *4-alpha-glucanotransferase activity*. A imagen (c) foi adaptada do QuickGO [Binns et al. (2009)]

- Segundo Camon et al. (2005), dentre as 19.490 anotações de BP disponíveis para *Homo sapiens*, 11.434 foram inferidas de anotações totalmente automáticas (a maioria estava correta). Apesar disso, muitas eram termos de muito alto nível na hierarquia do GO, o que limita sua utilidade.

2.1.2 *Enzyme Comission Number* (EC)

O EC number, como já descrito anteriormente em nossa introdução, é um sistema numérico e hierárquico de classificação de reações químicas da forma #.#.#.#, onde cada # representa um número e fornece progressivamente, da esquerda para direita, maiores detalhes sobre a reação. Na Tabela 2.1 estão as seis categorias básicas (de mais alto nível) da classificação EC.

Tabela 2.1: Classificação de enzimas, adaptado de [Lehninger et al. (2008)]

# classe	Classes	Tipo de reação catalisada
1	Oxidoredutase	Transferencia de elétrons
2	Transferase	Reações de transferência de grupos
3	Hidrolase	Reações de hidrólise
4	Liase	Adição de grupos a ligações duplas ou formação de ligações duplas por remoção de grupos
5	Isomerase	Transferência de grupos dentro de moléculas dando formas isoméricas
6	Ligase	Formação de C-C, C-S, C-O e C-N mediante reações de condensação acopladas à quebra de ATP.

Fizemos, de maneira simplificada, uma busca por cada uma dessas classes do EC nas bases de dados Google Scholar [Google (2012)], PDB [Berman et al. (2000)] e PubMed [PubMed (2012)]. O Google Acadêmico ou Google Scholar é uma máquina de busca *web* que permite pesquisar por artigos científicos, livros e trabalhos acadêmicos de modo geral. O PDB é uma base de estruturas protéicas e o PubMed é uma base de citações voltada para a literatura Biomédica. Buscamos pelos nomes das classes de mais alto nível da hierarquia EC em inglês nas bases Google Scholar e PubMed, e, no PDB, buscamos pelo dígito referente à cada classe. Em seguida, verificou-se quantos resultados eram retornados em cada consulta e calculamos o percentual de resultados relacionados à cada classe para cada uma das três bases. Aparentemente, algumas classes têm sido mais estudadas que outras ao longo dos anos. Observamos que as transferases possuem percentual de resultados significativo para as três bases (superior a 26% em todas elas). Os resultados relacionados a hidrolases também são significativos, com percentual que varia de aproximadamente 11% a 42%. Já as classes liase e isomerase apresentam percentual inferior a 10% para todas as bases. Os resultados podem ser vistos na Tabela 2.2. O fato de determinadas classes EC

serem mais estudadas que outras pode ter reflexo no nosso trabalho, pois possivelmente haverá mais exemplares de enzimas e suas mudanças de anotação EC associadas a tais classes.

Tabela 2.2: Resultado da busca pelas classes EC nas bases Google Scholar, PDB e PubMed (número absoluto e percentual).

Classe EC	Scholar		PDB		PubMed	
	absoluto	(%)	absoluto	(%)	absoluto	(%)
oxidoreductase	122.000	6,5	7.731	1,8	499.969	20,2
transferase	942.000	50,0	10.897	26,5	712.758	28,8
hydrolase	215.000	11,4	16.054	39,1	1.040.771	42,1
lyase	154.000	8,2	3.202	7,8	118.865	4,8
isomerase	177.000	9,4	1.655	4,0	47.984	1,9
ligase	273.000	14,5	1.517	3,7	52.562	2,1

No final da década de 50, como o número de enzimas conhecidas crescia rapidamente, os enzimologistas começaram a lidar com problemas decorrentes da falta de um vocabulário controlado. Muitos nomes distintos eram usados para descrever uma mesma enzima, além disso, era comum várias enzimas que catalisavam reações diferentes receberem o mesmo nome. Para contornar esse problema, os especialistas da área desenvolveram o EC *number* como um sistema padronizado e hierárquico de classificação de enzimas, que teve sua primeira versão em 1961 [Commission (1961)]. Desde então, o EC *number* tem sido largamente utilizado pela comunidade científica passando por diversas revisões ao longo dos anos, tendo na publicação de 1999 sua mais recente versão [NC-IUBMB (1999)].

Por outro lado, existem alguns problemas relacionados ao uso do EC *number*, como:

- O uso de EC *numbers* incompletos, como, por exemplo, a entrada do Uniprot $AK1C3\_HUMAN$, que possui identificador P42330 e foi anotada com o EC *number* 1.-.-.- [Egelhofer et al. (2010)]. Basicamente, isso acontece ou porque o especialista não se sente seguro para inferir a função exata da enzima ou porque, embora saiba exatamente a função da mesma, somente o IUBMB pode atribuir um EC *number* a uma enzima descoberta. Como esse é um processo rigoroso que pode levar meses, o especialista pode optar por depositar a mesma com o EC *number* incompleto. Não é possível distinguir entre um caso ou outro [Green e Karp (2005)].

- Um gene que codifica uma enzima anotada com um EC *number* parcial pode ser associado a muitas ou todas as reações bioquímicas anotadas com o mesmo EC *number* parcial, o que seria uma inferência incorreta dada a natureza ambígua desse tipo de EC *number* [Green e Karp (2005)]. Tomemos como exemplo o gene b3787 [identificador P27829 no Swiss-Prot, *UDP-N-acetyl-d-mannosamine dehydrogenase* (EC 1.1.1.-)]. A função do produto desse gene no KEGG[1] é *UDP-N-acetyl-*

[1] `http://www.genome.jp/dbget-bin/www_bget?eco:b3787`

d-mannosaminuronic acid dehydrogenase. Apesar disso, o KEGG associa esse gene a 15 reações diferentes e nenhuma delas corresponde a tal atividade enzimática.

- Outro problema do EC *number* e de todos os sistemas de classificação, segundo [Ma et al. (2007)], é que a diferente especificidade de substrato das enzimas nos diversos organismos não é capturada pelos vários modelos.

- Segundo [Egelhofer et al. (2010)] a reação catalisada pela enzima *sterol 14-demethylase* (1.14.13.70) foi corretamente anotada com a sub-subclasse 1.14.13 que, de acordo com a classificação EC *number* compreende as enzimas "acting on paired donors, with incorporation or reduction of molecular oxygen, with NADH or NADPH as one donor, and incorporation of one atom of oxygen". Porém, essa enzima também poderia ter sido anotada com a sub-subclasse 1.14.21, que contém enzimas "acting on paired donors, with incorporation or reduction of molecular oxygen, with NADH or NADPH as one donor, and the other dehydrogenated". Tais sub-subclasses são muito semelhantes e poderiam ser agrupadas em uma só sem perda de informação.

- De acordo com [Schmidt et al. (2003)], a reação $ATP + H2O = ADP + phosphate$ é catalisada pelas enzimas *adenosinetriphosphatase* (3.6.1.3) e *myosin* ATPase (3.6.4.1). Aqui, o princípio geral de que a classe de enzima é definida por sua reação química é violado. O mesmo é acontece para as peptidases, subclasse 3.4, que são classificados de acordo com o mecanismo e não com a reação.

Em nossas análises de mudanças de anotação, utilizamos a classificação EC *number* e não o GO por se tratar de uma análise histórica, na qual estudamos versões mais antigas do UniProt/Swiss-Prot. Como o GO é uma classificação mais recente, não está presente em diversas versões históricas do repositório de dados tratado.

2.2 Análise de Anotações

Existem diversos trabalhos que abordam a questão de qualidade da anotação de sequências protéicas e de nucleotídeos. Aqui faremos uma breve discussão abordando alguns deles.

Parte desses estudos foi desenvolvida no começo do que é chamado na literatura de *genomic era*, como é o caso de [Brenner et al. (1999)] e [Devos e Valencia (2001)]. O primeiro examina as anotações para o genoma de *Mycoplasma genitalium* realizadas por três grupos diferentes e encontra um percentual de erro de anotação entre 7% e 15% (dependendo dos genes analisados e do grupo ou grupos responsáveis pela análise). O segundo calcula percentuais de erros contando o número de diferenças de anotação em conjuntos de proteínas similares para os genomas *Mycoplasma genitalium*, *Haemophilus influenzae*

e *Methanococcus jannaschii* e conclui que, para o primeiro genoma, esse percentual fica entre 4% e 40%, enquanto nos dois últimos fica entre 4% e 34% (dependendo to tipo de anotação considerado). Nota-se que esses trabalhos se basearam em discrepâncias de anotações feitas por grupos distintos para genomas bem específicos, o que, segundo [Schnoes et al. (2009)], estabelece um limite inferior dos prováveis níveis de falhas de anotação.

Há também estudos mais recentes, como [Green e Karp (2005)], no qual reportam um novo tipo de falha sistemática de anotação, que resulta da interpretação equivocada dos *Enzyme Commission numbers* parciais. Esse tipo de interpretação leva à associação de genes anotados com EC *number* parcial a várias reações bioquímicas anotadas com o mesmo EC *number* parcial, o que pode ser uma inferência incorreta. Ainda de acordo com Green e Karp (2005), dentre os 135 genes de *E.coli* do KEGG anotados com EC *number* parcial, 43,7% estão incorretamente anotados.

Em Jones et al. (2007) foi desenvolvida uma metodologia para estimar os níveis de erros de anotação de sequências. Erros são adicionados (de maneira artificial e em taxas previamente determinadas) a anotações de sequências e usam regressão para modelar o impacto que isso provocaria nas anotações que se baseiam no BLAST. A metodologia foi aplicada à base de dados GOSeqLite, mais precisamente às anotações de sequência com termos do GO, e concluíram que a taxa de erro de anotação varia de 28% a 30%, sendo que para as anotações não baseadas em similaridade de sequência a taxa é de 13% a 18% e para as anotações que se baseiam em similaridade a taxa é de 49%.

Já em [Gilks et al. (2005)] o formalismo desenvolvido em [Gilks et al. (2002)] foi aplicado a um modelo de base de dados protéica e hierarquicamente estruturada. Concluíram que o poder discriminatório é perdido mais rapidamente dentro de uma determinada superclasse do que entre superclasses. Sugerem o uso de um *copy number* h, onde $h = 0$ quando a anotação provém de dados experimentais e $h > 0$ quando a anotação é copiada de uma sequência com *copy number* $h - 1$. Afirmam que isso diminuiria a propagação de erros de anotação.

Schnoes et al. (2009) investigou os níveis de falha de anotação de função molecular nos repositórios de dados biológicos UniProtKB/Swiss-Prot, GenBank Non-redundant (NR), UniProtKB/TrEMBL e KEGG para 37 famílias de enzimas com evidência experimental no *Structure-Function Linkage Database* (SFLD) (Pegg et al., 2006). O Swiss-Prot apresentou percentual de erro próximo de 0 para a maioria das famílias enquanto GenBank NR, TrEMBL e KEGG apresentaram percentual de erro entre 5% e 63%. Ainda em Schnoes et al. (2009), uma análise das sequências do GenBank NR revelou que, em 1999, o nível de falha de anotação era próximo de 0% e em 2005 era próximo de 40%, indicando que as falhas de anotação aumentaram significativamente nesse período.

Em Hung et al. (2010) é proposto um método para previsão de enzima baseado em alinhamento global e local de sequência chamado Density Estimation Tool for Enzyme ClassificaTion (DETECT). Essa técnica utiliza o teorema de Bayes para integrar informação de perfis de estimação de densidade ou *density estimation profiles* para cada EC

number, de modo que uma probabilidade é calculada a partir da similaridade de todas as proteínas relevantes para um determinado EC *number* e não com base em apenas uma sequência. Quando comparado ao BLAST, o DETECT revelou melhora na acurácia de anotação de enzimas e, quando aplicado ao *Plasmodium falciparum*, erros de anotação foram identificados.

Egelhofer et al. (2010) estudou inconsistências no esquema de classificação de enzimas EC *number* pois podem levar a problemas na anotação das mesmas. Dados de 3788 reações enzimáticas foram validados e mais de 80% das associações de um EC *number* a tais reações estava de acordo o esquema EC. Os resultados podem ser utilizados para fazer correções e aprimorar o sistema de classificação EC.

Em Quester e Schomburg (2011), EnzymeDetector foi proposto para comparar e avaliar de modo automático as funções enzimáticas associadas a entradas dos repositórios NCBI RefSeq (Pruitt et al., 2009), KEGG, PEDANT (Walter et al., 2009), Pseudomonas Genome Database (Winsor et al., 2011) e UniProt/Swiss-Prot. A ferramenta ainda complementa essas informações com sua própria previsão de função, que é baseada em análise de similaridade de sequência, em informações do BRENDA (Braunschweig Enzyme Database) e em busca de padrões em sequências. Nesse mesmo trabalho, nove genomas de procariotos foram analisados e encontraram aproximadamente 70% de inconsistências nas previsões de enzimas dos repositórios considerados.

A ferramenta FunTree foi apresentada em Furnham et al. (2012) e reúne dados filogenéticos, sequência, estrutura bem como informações químicas e funcionais para um conjunto de superfamílias de enzimas definidas estruturalmente. Os autores afirmam que a combinação desse conjunto de dados permite investigar a evolução de novas funções enzimáticas dentro de cada superfamília, o que pode apoiar na previsão de função para enzimas ainda não caracterizadas.

Esses trabalhos evidenciam o interesse da comunidade científica em aferir os níveis de falhas de anotação dos repositóris biológicos e, de modo mais geral, no problema de anotação. Observamos que esses níveis de falhas de anotação são significativos e que as diversas estratégias que são utilizadas para anotar sequências protéicas e de nucleotídios de modo automático possuem limitações.

Capítulo 3

Objetivos

3.1 Objetivo Geral

Projetar, implementar e avaliar uma estratégia de aprendizagem supervisionada que permita prever mudanças de anotação de enzimas em dados temporais de repositórios biológicos com base metadados das entradas de tal repositório. Esse objetivo se baseia na hipótese de que metadados de anotação de repositórios biológicos podem indicar que uma mudança de anotação ocorrerá.

3.2 Objetivos Específicos

- Coletar todas as versões disponíveis das entradas do UniProt/Swiss-Prot.
- Modelar o problema da dinâmica das anotações, definindo categorias para as mudanças de EC *number* observadas no repositório de dados de acordo com a natureza hierárquica da classificação EC, considerando especializações e generalizações.
- Construir um banco de dados contendo as informações das mudanças de EC *number* das entradas do UniProtKB/Swiss-Prot conforme item anterior.
- Analisar as mudanças de EC *number*, sua frequência e impacto ao longo das versões e das diferentes classes de enzimas.
- Modelar as mudanças de EC *number* em termos dos metadados do Swiss-Prot selecionados para discriminar entradas com anotação estável das que sofreram um tipo específico de mudança de EC .
- Elaborar, implementar e avaliar uma estratégia que permita verificar se os metadados selecionados são capazes de discriminar entradas estáveis das que sofreram um tipo específico de alteração.

- Projetar, implementar e avaliar um modelo de aprendizagem supervisionada baseado nesses atributos discriminantes para prever alterações de EC *number* nas entradas do UniProt/Swiss-Prot.
- Comparar os resultados da estratégia proposta com outra técnica capaz de fazer previsões de anotação EC.

Capítulo 4

Materiais e Métodos

Neste capítulo são detalhadas as etapas de construção do ENZYMAP, nossa estratégia baseada em aprendizagem supervisionada para previsão de mudanças de EC *number*.

Para caracterizar e prever as mudanças de EC *number*, três experimentos foram realizados: *Descritivo Multiclasse*, cujo objetivo é verificar se é possível separar entradas do UniProt/Swiss-Prot que experimentaram uma mudança de EC específica daquelas em que o EC permaneceu o mesmo com base em metadados das entradas do repositório; *Previsivo Multiclasse*, no qual todos os dados disponíveis no repositório a respeito de um tipo de mudança de EC são utilizados para prever uma mudança do mesmo tipo; *Previsivo Origem Comum*, que segmenta as mudanças de EC pelo prefixo comum (EC *number* antes da mudança) para aprimorar o experimento anterior.

Na Seção 4.1 são descritos os dados utilizados nesse trabalho, mais especificamente as versões do Swiss-Prot analisadas e também os metadados desse repositório selecionados para caracterizar as mudanças de anotação. Na Seção 4.2 é abordada a modelagem dos dados, num primeiro momento é descrita uma exploração inicial das mudanças de EC e em seguida descreve-se como os metadados do Swiss-Prot foram modelados para alimentar a estratégia de aprendizagem supervisionada proposta. Na Seção 4.3 são detalhadas as técnicas empregadas no processamento dos metadados, os experimentos realizados e as técnicas de redução de dimensionalidade e classificação utilizadas.

4.1 Dados

Foram obtidas, através do ftp do Uniprot[1], as versões completas da base disponíveis em maio de 2012, chamadas *major releases*, de 1 a 44. Trabalhamos com a parte manualmente revisada, referente ao Swiss-Prot. Para analisar uma mudança de EC *number*, é preciso observar a anotação de uma mesma entrada da base em duas versões distintas. Desse modo, as versões citadas foram estudadas par a par e foi tomado o conjunto interseção

[1] ftp.uniprot.org

dos identificadores de cada par de versões. Esses conjuntos tiveram suas alterações de EC *number* estudadas.

O Swiss-Prot é uma base de sequências protéicas e anotação funcional para as mesmas, que conta com revisão manual de especialistas e é considerada padrão ouro na anotação de proteínas. Como neste trabalho a proposta é prever as mudanças de anotação EC com base em metadados presentes nas entradas do repositório, optou-se pelo Swiss-Prot devido à riqueza e qualidade de suas anotações.

Na Tabela 4.1 são apresentadas algumas informaçãoes das versões utilizadas, como data em que foram disponibilizadas publicamente, número e percentual de entradas que possuem EC *number* e total de entradas para cada versão. A Figura 4.1 sintetiza esses dados. Na Tabela 4.2 e na Figura 4.2 estão os dados dos pares de versões.

Alguns metadados foram selecionados no conjunto de 44 versões do UniProt/Swiss-Prot para caracterizar e prever mudanças de anotação de enzimas. Esses metadados são descritos a seguir.

4.1.1 Metadados Selecionados

Nos experimentos Descritivo e Previsivo da aborgagem de aprendizagem supervisionada, estamos interessados em metadados (atributos de anotação) presentes nas entradas do repositório Swiss-Prot que sejam capazes de discriminar e caracterizar as mudanças de anotação EC. As *line types* ou linhas *Organism Classification* (OC), *Reference Position* (RP) e *KeyWord* (KW) dos *flat files* ou arquivos texto das entradas do Swiss-Prot foram selecionadas como metadados candidatos. Maiores detalhes a respeito do formato dos arquivos texto podem ser obtidos no manual do usuário do UniProt[2].

De acordo com o manual do usuário:

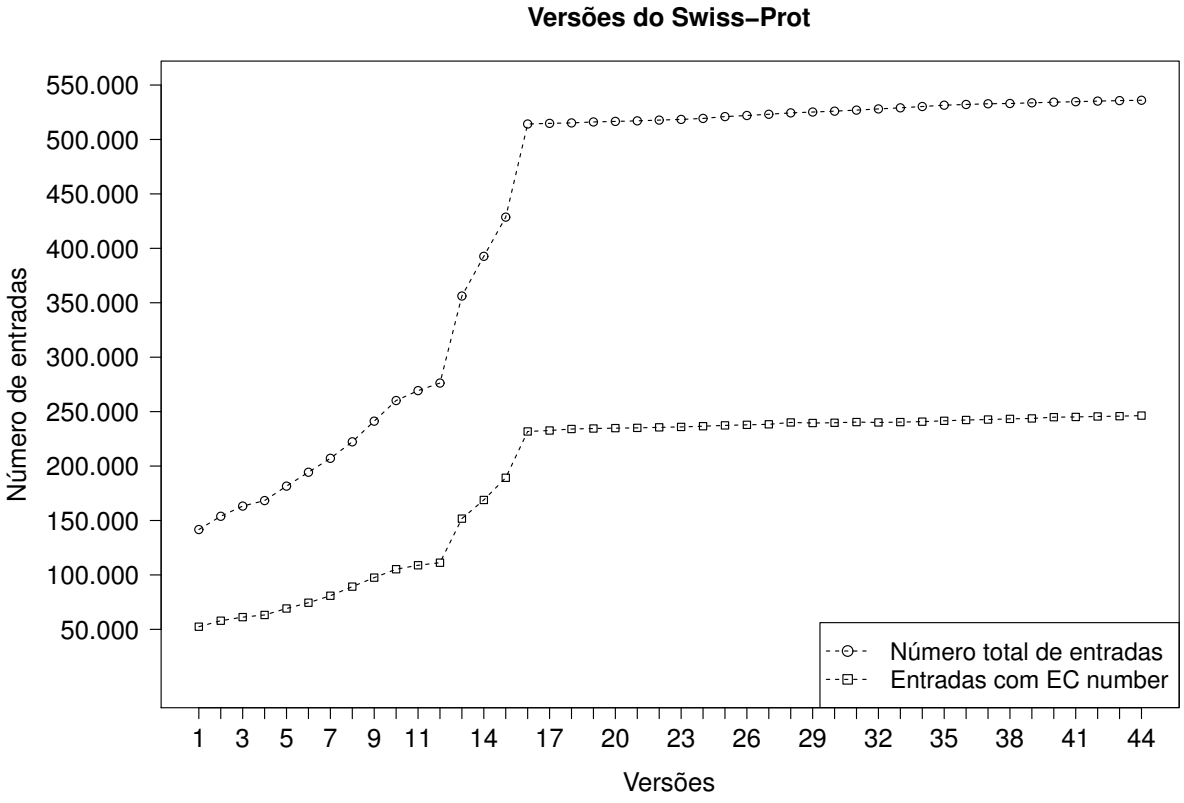
> *The RP (Reference Position) lines describe the extent of the work relevant to the entry carried out by the authors. It should contain a description of the information that has been propagated in the Swiss-Prot entry.*
>
> *The OC (Organism Classification) lines contain the taxonomic classification of the source organism. The taxonomic classification used is that maintained at the NCBI (see http://www.ncbi.nlm.nih.gov/Taxonomy/) and used by the nucleotide sequence databases (EMBL/GenBank/DDBJ)*
>
> *The KW (KeyWord) lines provide information that can be used to generate indexes of the sequence entries based on functional, structural, or other categories. The keywords chosen for each entry serve as a subject reference for the sequence.*

Assim, a linha OC é referente à taxonomia do organismo ao qual a enzima pertence, RP nos informa a porção de uma referência bibliográfica relevante para anotar determinada

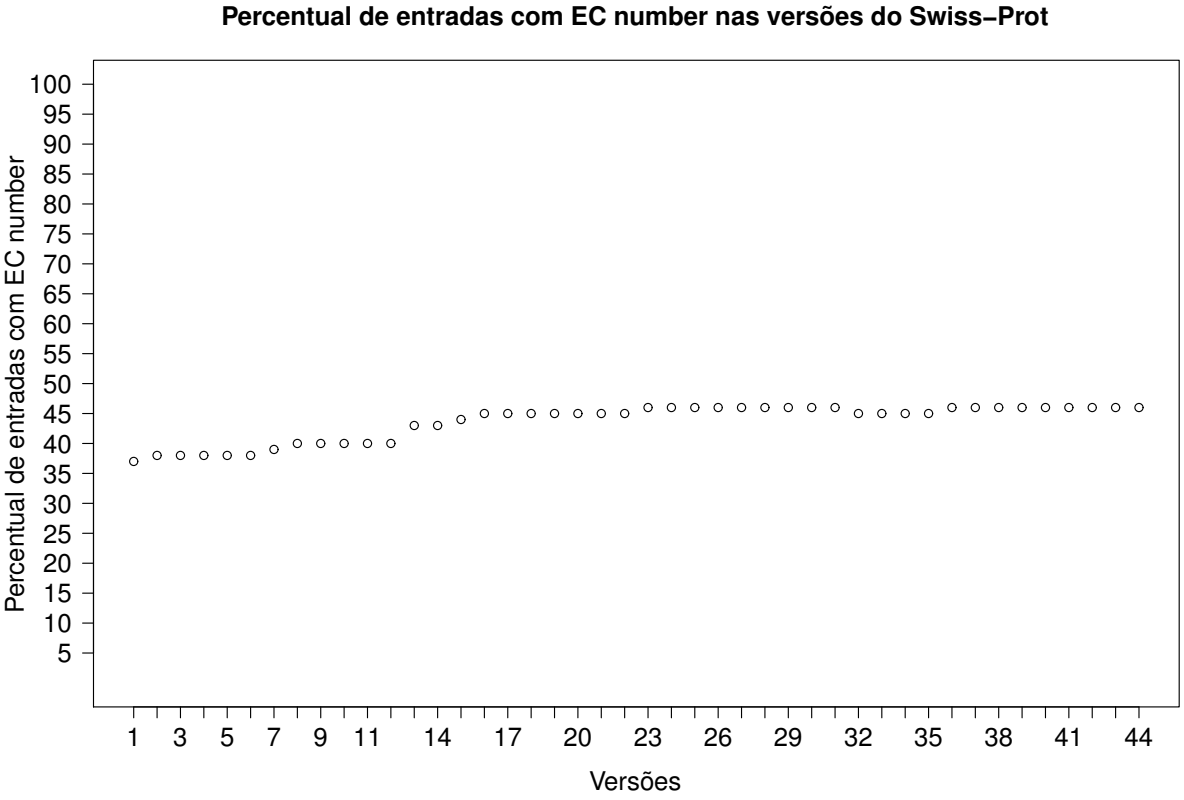
[2]http://web.expasy.org/docs/userman.html

Tabela 4.1: Versões 1 a 44 do Swiss-Prot: índice e nome da versão, data de lançamento, percentual e número absoluto de entradas com EC *number* e total de entradas.

Índice da versão	Nome da versão	Lançamento (MM/DD/AAAA)	% entradas com EC	# de entradas com EC	Total de entradas
1	1	12/15/2003	37	52.434	141.681
2	2	07/05/2004	38	57.931	153.871
3	3	10/25/2004	38	61.229	163.235
4	4	02/01/2005	38	63.221	168.297
5	5	05/10/2005	38	69.164	181.571
6	6	09/13/2005	38	74.468	194.317
7	7	02/07/2006	39	80.874	207.132
8	8	05/30/2006	40	89.245	222.289
9	9	10/31/2006	40	97.508	241.242
10	10	03/06/2007	40	105.225	260.175
11	11	05/29/2007	40	108.876	269.293
12	12	07/24/2007	40	111.230	276.256
13	13	02/26/2008	43	151.694	356.194
14	14	07/22/2008	43	168.849	392.667
15	15	03/24/2009	44	189.234	428.650
16	2010_01	01/19/2010	45	231.776	514.212
17	2010_02	02/09/2010	45	232.662	514.789
18	2010_03	03/02/2010	45	234.040	515.203
19	2010_04	03/23/2010	45	234.494	516.081
20	2010_05	04/20/2010	45	234.843	516.603
21	2010_06	05/18/2010	45	235.081	517.100
22	2010_07	06/15/2010	45	235.561	517.802
23	2010_08	07/13/2010	46	235.952	518.415
24	2010_09	08/10/2010	46	236.597	519.348
25	2010_10	10/05/2010	46	237.361	521.016
26	2010_11	11/02/2010	46	237.872	522.019
27	2010_12	11/30/2010	46	238.344	523.151
28	2011_01	01/11/2011	46	240.052	524.420
29	2011_02	02/08/2011	46	239.545	525.207
30	2011_03	03/08/2011	46	239.775	525.997
31	2011_04	04/05/2011	46	240.406	526.969
32	2011_05	05/03/2011	45	240.055	528.048
33	2011_06	05/31/2011	45	240.374	529.056
34	2011_07	06/28/2011	45	240.787	530.264
35	2011_08	07/27/2011	45	241.578	531.473
36	2011_09	09/21/2011	46	242.309	532.146
37	2011_10	10/19/2011	46	242.742	532.792
38	2011_11	11/16/2011	46	243.333	533.049
39	2011_12	12/14/2011	46	243.749	533.657
40	2012_01	01/25/2012	46	244.898	534.242
41	2012_02	02/22/2012	46	245.113	534.695
42	2012_03	03/21/2012	46	245.566	535.248
43	2012_04	04/18/2012	46	245.826	535.698
44	2012_05	05/16/2012	46	246.347	536.029



(a)

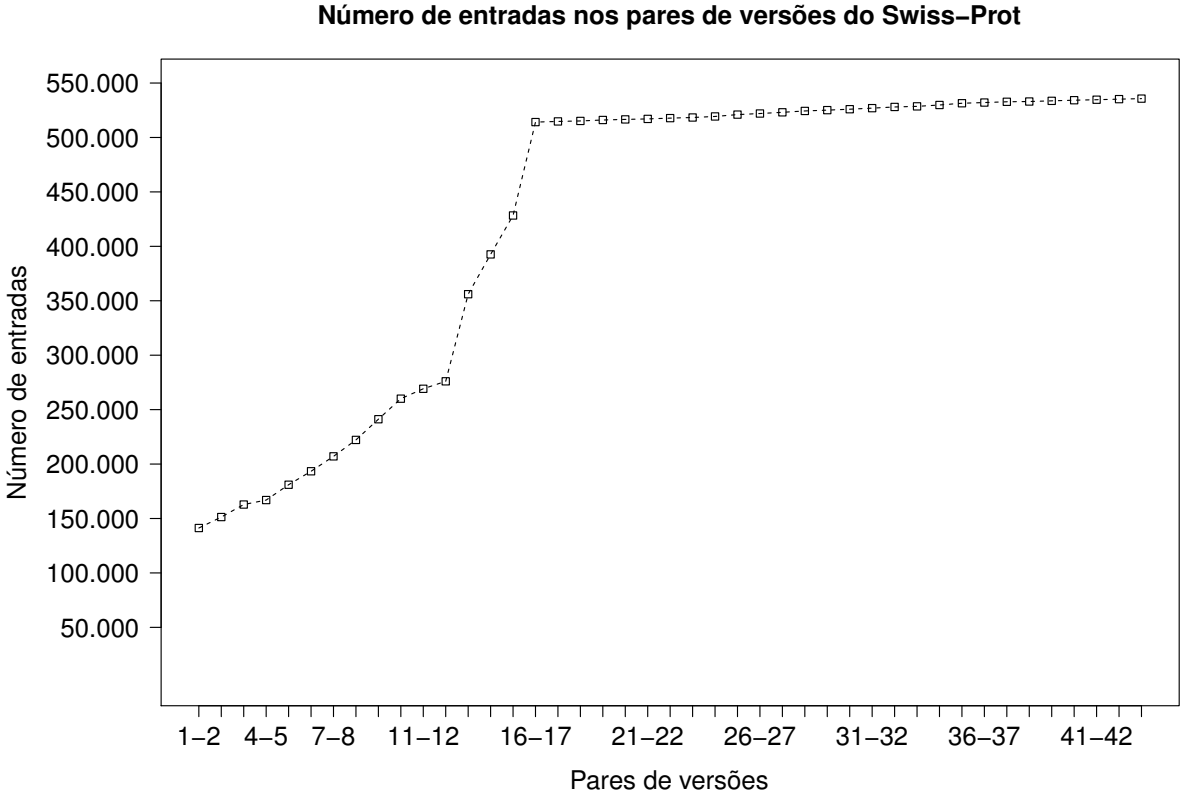


(b)

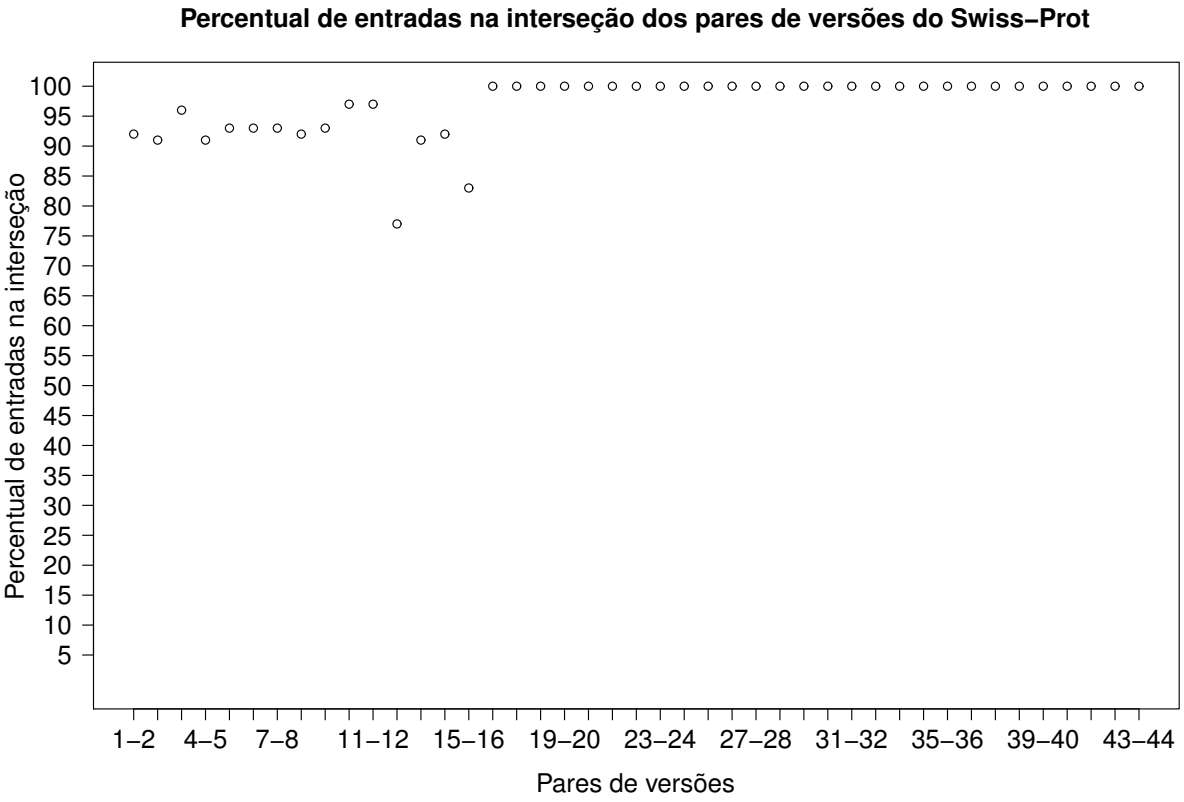
Figura 4.1: Dados das versões do UniProt/Swiss-Prot referentes à Tabela 4.1. (a) Número total de entradas da base e número de entradas anotadas com EC *number*. (b) Percentual de entradas anotadas com EC *number*.

Tabela 4.2: Pares de versões analisadas e número de entradas estudadas em cada par.

Pares de versões	Número de entradas na ∩
1-2	141.249
2-3	151.318
3-4	162.812
4-5	166.933
5-6	181.005
6-7	193.382
7-8	207.069
8-9	222.181
9-10	241.189
10-11	260.065
11-12	269.152
12-13	276.011
13-14	356.036
14-15	392.597
15-16	428.331
16-17	514.121
17-18	514.740
18-19	515.180
19-20	516.049
20-21	516.593
21-22	517.045
22-23	517.769
23-24	518.350
24-25	519.302
25-26	521.007
26-27	522.001
27-28	523.101
28-29	524.367
29-30	525.107
30-31	525.960
31-32	526.934
32-33	528.024
33-34	528.573
34-35	529.826
35-36	531.443
36-37	532.076
37-38	532.780
38-39	533.028
39-40	533.643
40-41	534.227
41-42	534.678
42-43	535.207
43-44	535.682



(a)



(b)

Figura 4.2: Dados dos pares de versões do UniProt/Swiss-Prot referentes à Tabela 4.2. (a) Número de entradas no conjunto interseção dos identificadores de cada par de versões. (b) Percentual de entradas do par de versões que está no conjunto interseção.

entrada e KW contém termos relacionados a uma entrada e que podem ser utilizados para indexá-la com base base em função e estrutura, dentre outros.

Esses atributos foram escolhidos porque, no caso do OC, há organismos a respeito dos quais já existem maiores estudos, o que possivelmente levaria a anotações de melhor qualidade. Como exemplo podemos citar os organismos *Saccharomyces cerevisiae*, *Drosophila melanogaster* e *Caenorhabditis elegans*, considerados modelo e a respeito dos quais existem numerosos estudos. De maneira semelhante, como o RP informa porque determinada referência foi utilizada para anotar uma entrada, acreditamos que entradas com referências mais específicas, como *function*, seriam mais bem anotadas do que entradas com referências mais gerais, como *nucleotide sequence large scale genomic dna*. Já a tag KW representa uma espécie de sumário de cada entrada da base, contendo palavras relevantes relacionadas à ela. Um exemplo dessas tags é mostrado abaixo para o identificador P66880 do Swiss-Prot cujo o EC *number* é atualmente 3.1.3.5.

```
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA]
OC   Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;
KW   Complete proteome; Cytoplasm; Hydrolase; Metal-binding;
KW   Nucleotide-binding.
```

4.2 Modelagem

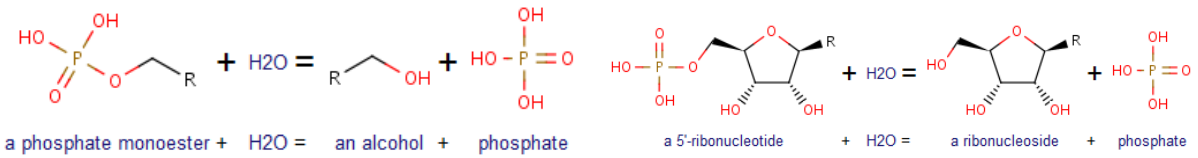
4.2.1 Exploração Inicial

Na etapa de exploração inicial das mudanças de EC *number* foram utilizadas as versões 1 a 15 do Swiss-Prot, que eram as versões disponíveis em Março de 2009, no início desse estudo. Tal etapa resultou na publicação do artigo [Silveira et al. (2012)] intitulado *Advise: Visualizing the dynamics of enzyme annotations in UniProt/Swiss-Prot* no *IEEE Symposium on Biological Data Visualization (BioVis), 2012* realizado em Seattle, EUA. O artigo está anexado ao final desse texto. Abaixo são descritas a modelagem proposta no artigo e suas principais conclusões. Posteriormente as versões desse estudo foram atualizadas e agora ele contempla as 44 versões do Swiss-Prot.

Com base na natureza hierárquica da classificação EC, foram definidas algumas categorias para classificar as mudanças observadas ao longo das versões 1.0 a 15.0. É importante saber o nível da hierarquia EC em que as mudanças ocorrem, pois mudanças nos níveis mais altos (mais à esquerda) são mais graves que nos níveis mais baixos. Desse modo, definimos os parâmetros prefixo comum, generalizações e especializações, que representam respectivamente o tamanho do prefixo comum de dois EC *numbers* envolvidos numa mudança, número de níveis que foram apagados e número de níveis adicionados.

Tomemos como exemplo a seguinte mudança de EC *number*:

$$3.1.3.2 \Rightarrow 3.1.3.5$$



(a) (b)

Figura 4.3: Esquema da reação catalisada por enzimas com EC *number* 3.1.3.2 (a) e com EC 3.1.3.5 (b). Adaptado do BRENDA <http://www.brenda-enzymes.org/>.

Nessa mudança uma enzima *acid phosphatase* (Figura 4.3 (a)) passa a ser classificada como *5'-nucleotidase* (Figura 4.3 (b)). Nela, o prefixo comum é de tamanho 3, houve 1 generalização (o último nível foi apagado) e 1 especialização (o último nível foi escrito). Na Tabela 4.3 há mais exemplos de mudanças de EC *number* experimentadas por entradas do Swiss-Prot com suas generalizações, especializações e tamanho do prefixo comum.

Tabela 4.3: Exemplos de mudanças de EC *number* com identificadores das entradas do Swiss-Prot que sofreram tais mudanças, versões em que ocorreram, tamanho do prefixo comum, generalizações e especializações.

EC anterior	EC novo	Id	Versões	Prefixo	Generalização	Especialização
-.-.-.-	-.-.-.-	Q9K5T1	1-2	0	0	0
3.1.4.14	1.7.-.-	P41407	7-8	0	4	2
1.1.1.-	1.-.-.-	P52895	5-6	1	2	0
5.3.-.-	5.3.1.27	P42404	14-15	2	0	2
2.5.1.64	2.5.1.-	P17109	13-14	3	1	0
4.1.1.22	4.1.1.22	P95477	1-2	4	0	0

4.2.1.1 Visualização das Mudanças de EC *Number*

Devido à numerosidade dos pares de EC *number*, 3.281.206, observados da versão 1.0 a 15.0 do Swiss-Prot, foi proposto um mapa que permite visualizar essas anotações segmentadas pelos parâmetros propostos em 4.2.1, que são tamanho do prefixo comum, generalizações e especializações e ainda segmentar pelas versões do repositório.

A unidade básica dessa visualização será chamada de *frame*, e pode ser vista na figura 4.4. Nela, o eixo x representa especializações, o eixo y representa as generalizações e ambos podem variar de 0 a 4, dado que o EC *number* possui 4 níveis que podem ser removidos ou adicionados. É interessante destacar algumas posições importantes:

(a) (b)

Figura 4.4: Unidades básicas da visualização proposta. (a) Heatmap: quanto mais escura a cor, maior o valor representado. (b) Quadmap: quanto maior a área do retângulo maior o valor. Vermelho representa entradas acima da diagonal, azul representa entradas abaixo da diagonal e bege representa entradas na diagonal. Em (a) e (b), cinza escuro representa mudanças que não podem acontecer devido ao tamanho do prefixo comum representado pelo *frame*. O cinza claro representa posições vazias.

- *Posição (0,0)*: corresponde às entradas que não sofreram mudanças num dado par de versões da base;
- *Diagonal*: representa entradas que sofreram o mesmo número de especializações e generalizações (exibida em vermelho na Figura 4.4 (a)) e são potenciais correções de anotação. Está representada em bege no Quadmap;
- *Matriz triangular inferior*: corresponde às entradas que sofreram mais especializações que generalizações (posições abaixo da diagonal na Figura 4.4 (a)) e são representadas em azul no Quadmap;
- *Matriz triangular superior*: compreende as entradas com mais generalizações que especializações, ou seja, entradas que perderam anotação (posições acima da diagonal na Figura 4.4 (a)). Está representada em vermelho no Quadmap.
- *Posições inválidas*: tomemos como exemplo uma mudança com prefixo comum de tamanho 3. Nesse caso não é possível que tal mudança tenha generalizações ou especializações em 2 ou mais níveis do EC. Esse tipo de evento é representado em cinza escuro.

Diversos *frames* como esses foram organizados de acordo com a técnica de Pequenos Múltiplos [Tufte (1990)] como mostrado na Figura 4.5. Na visualização como um todo, o eixo x representa pares de versões do Swiss-Prot e o eixo y representa o parâmetro prefixo comum, que aqui varia de 0 a 3, pois o prefixo comum de tamanho 4 é referente a uma entrada que não sofreu mudança de anotação e estamos particularmente interessados nas mudanças.

Dois tipos de visualizações foram propostos:

- *Heatmap*: a cor é utilizada como atributo pré-atentivo que representa a frequência de mudanças numa determinada posição do *frame*. Quanto mais escuro o verde, maior a frequência de mudanças de uma determinada posição. Essa representação fornece um panorama geral dos dados, permitindo que sejam facilmente identificadas a diagonal e matrizes triangulares inferior e superior. Tais matrizes representam tendências de especialização e generalização nas anotações.

- *Quadmap*: a área é utilizada para representar a frequência de mudanças numa determinada posição do *frame* dado que é um atributo visual mais preciso que a cor para demonstrar quantidade. Na Figura 4.5, é mais fácil estimar as frequências de mudança no Quadmap que no Heatmap. No Quadmap, o tamanho das posições (retângulos) é diferente de um *frame* para o outro. Para contornar isso, usamos as cores bege, vermelha e azul para representar, respectivamente, pontos na diagonal, acima e abaixo da mesma.

Além das representações citadas, foram adicionados alguns filtros e a possibilidade de visualizar um *frame* em particular e ainda posições específicas dentro de um *frame*, exibidas como histogramas onde as mudanças são separadas pelas grandes classes do EC *number* (nível mais à esquerda). São mostrados ainda os metadados referentes a cada uma das mudanças de EC (OC, RP e KW). Dessa maneira, essa visualização, chamada de Advise, se tornou interativa. Um vídeo[3] da ferramenta destacando suas principais funcionalidades, bem como o código e passos para a instalação[4] estão disponíveis na Internet.

Os filtros que podem ser aplicados à visualização são:

- *Escala linear ou logarítimica*: a cor do Heatmap ou a área dos retângulos no Quadmap são computados de acordo com o número absoluto das frequência ou com o logarítmo das mesmas.

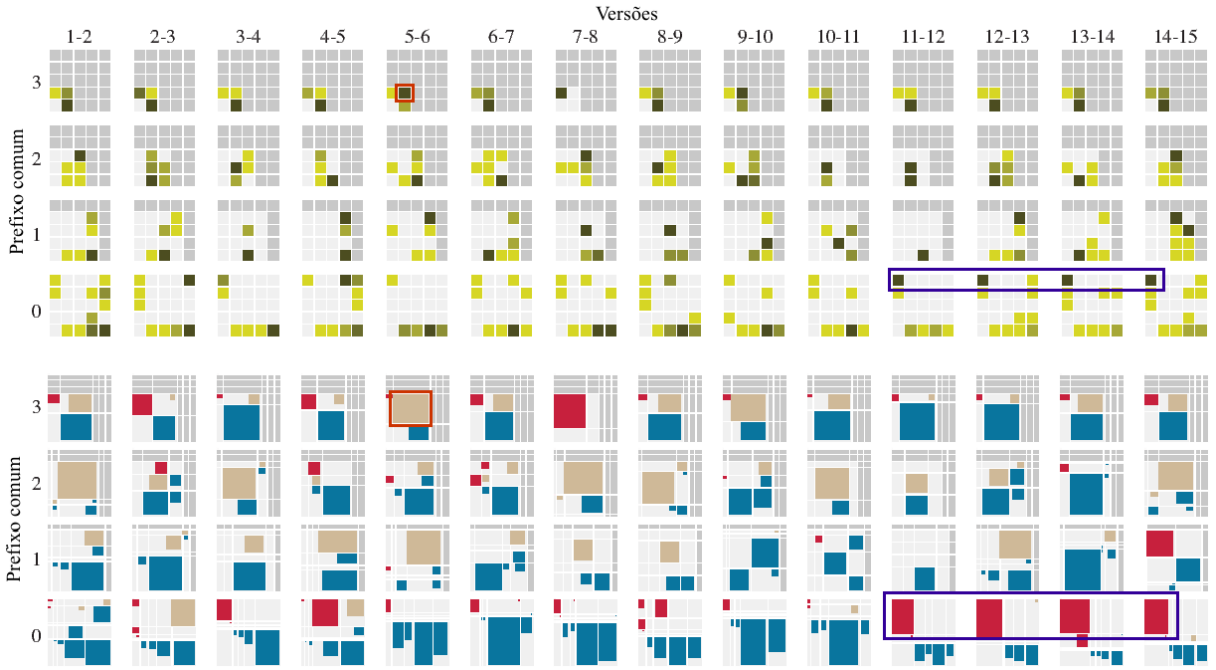
- *Normalização global ou local*: a normalização global destaca posições de alta frequência considerando o conjunto de dados como um todo, enquanto a normalização local destaca posições de alta frequência dentro de *frames* específicos.

- *Somente mudanças ou dados completos*: exibe somente entradas que sofreram mudança de EC ou o conjunto de dados completo.

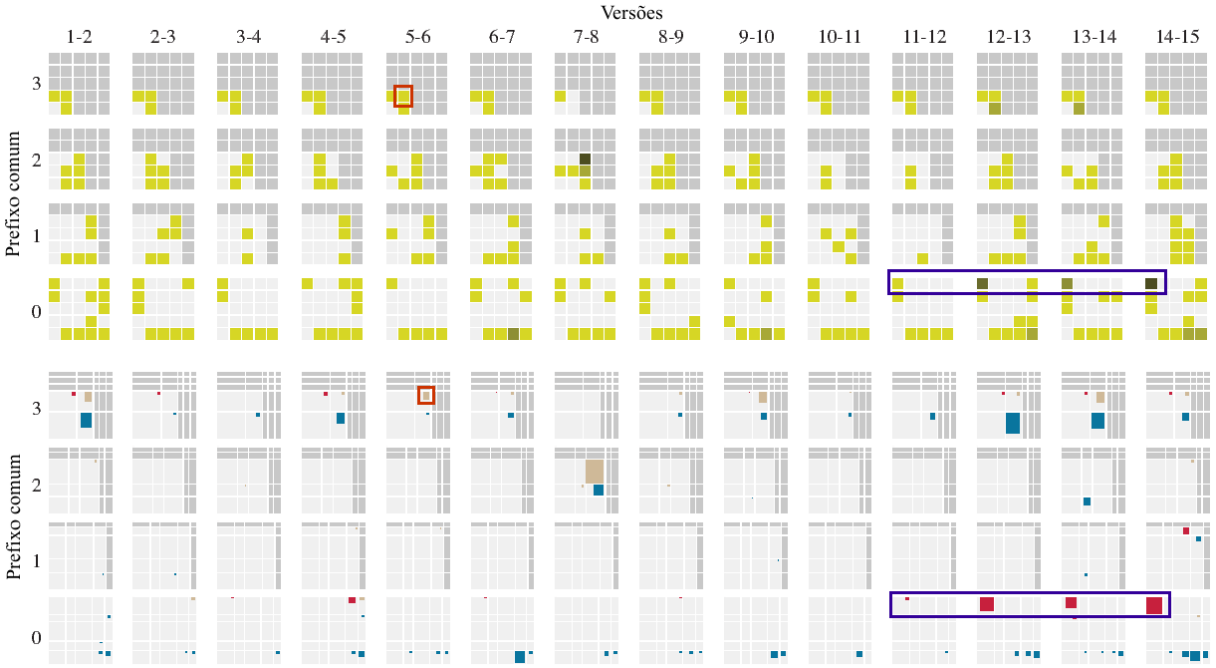
A seguir discutiremos posições destacadas na Figura 4.5 (a) e (b), tanto no Heatmap quanto no Quadmap, que permitem elucidar a representatividade da visualização proposta e também alguns eventos interessantes detectados. Outros vários casos relevantes são abordados e discutidos em [Silveira et al. (2012)].

[3] `http://vimeo.com/41296155`

[4] `https://github.com/arturhoo/ADVISe`



(a)



(b)

Figura 4.5: (a) Heatmap e Quadmap com escala linear, somente mudanças exibidas e normalização local. (b) Heatmap e Quadmap com escala linear, somente mudanças exibidas e normalização global. Em (a) a normalização local destaca mudanças numerosas dentro de cada *frame* e em (b) a normalização global destaca mudanças numerosas em relação a todo o conjunto de dados considerado.

Nas posições destacadas por quadrados alaranjados do par de versões 5 e 6 e na linha cujo tamanho do prefixo comum é 3, são representadas as 115 mudanças que ocorreram da versão 5 para a versão 6 do Swiss-Prot, cujos EC *numbers* envolvidos possuem prefixo comum de tamanho 3 e experimentaram uma generalização (um nível foi apagado) e uma especialização (um nível foi escrito). As mudanças referentes a esse ponto estão na Tabela 4.4. Na Figura 4.5 (a) esses pontos estão mais destacados devido à normalização local, enquanto em (b) o destaque é menor em relação ao conjunto de dados completo (normalização global).

Tabela 4.4: Mudanças referentes aos quadrados de cor laranja nas versões 5-6 da figura 4.5.

EC anterior	EC novo	Frequência
2.4.1.21	2.4.1.42	12
2.7.7.19	2.7.7.21	1
3.1.3.2	3.1.3.5	77
3.1.3.2	3.1.3.6	6
3.1.4.17	3.1.4.35	18
4.1.1.17	4.1.1.19	1

Nas 4 posições destacadas por retângulos roxos no Heatmap e no Quadmap da Figura 4.5 (a) e (b), mais especificamente os pontos com prefixo comum de tamanho 0, 4 generalizações e nenhuma especialização, estão representadas entradas da base cujos EC *numbers* tiveram os 4 níveis apagados, uma mudança drástica, já que tais entradas perderam esses EC *numbers*. Isso ocorreu em 146 entradas nos pares de versões 11-12, em 1357 entradas nas versões 12-13, em 1006 entradas nas versões 13-14 e em 1976 nas versões 14-15. De acordo com o UniProtKB/Swiss-Prot, eles procuram associar EC *numbers* apenas a subunidades catalíticas, de modo que, muitas vezes, em grandes complexos protéicos apenas uma ou poucas subunidades receberão anotação EC. Quando descobrem que esse procedimento foi violado, o EC *number* é completamente removido das subunidades que não possuem atividade enzimática. Para ilustrar, tomemos as seguintes entradas nas versões 12-13:

- Q6FSJ2, possuía anotação EC 1.10.2.2 e essa foi removida porque a subunidade 7 do *cytochrome b-c1* não é a subunidade com atividade de redutase.
- Q8LX28, é a subunidade 8 de *ATP synthase* e faz parte de *membrane proton channel*. Teve o EC *number* 3.6.3.14 removido.
- Q6AY96, teve o EC *number* 2.7.11.1 removido porque é uma subunidade de *transcription factorm*, mas não possui atividade de *serine/threonine kinase*.

Após essa etapa de exploração inicial, na qual foram identificadas algumas tendências e exceções nas mudanças de EC *number* ao longo de várias versões do repositório, foi

realizada a modelagem dos dados para a etapa de caracterização e previsão de mudanças de anotação EC.

4.2.2 Experimentos Descritivo e Previsivo

Três experimentos foram realizados para caracterizar e prever as mudanças de EC *number*: *Descritivo Multiclasse*, *Previsivo Multiclasse* e *Previsivo Origem Comum*. A modelagem das mudanças de anotação EC em termos dos metadados selecionados (OC, RP e KW) é a mesma para os três experimentos. Neles foram utilizadas as 44 versões do Swiss-Prot.

Dados de treinamento contendo entradas que sofreram mudanças de EC *number* e dados de entradas em que o EC *number* se manteve constante são necessários para caracterizar e prever mudanças de anotação EC utilizando a estratégia de aprendizagem supervisionada proposta. Nela, o algoritmo deve aprender com esses dados numa etapa de treinamento para que num passo posterior possa separar um conjunto de entradas que sofreu mudanças na anotação EC (conjunto mudança) de um conjunto em que a anotação não mudou (conjunto controle). Como exemplo de um tipo de mudança de EC podemos citar a entrada com indentificador Q9PKH4 do Swiss-Prot, cujo EC mudou de 3.1.3.2 para 3.1.3.5 da versão 5 para 6. Como exemplo de controle, podemos citar o identificador P20611, cujo EC 3.1.3.2 se manteve o mesmo da versão 5 para 6.

Para modelar mudanças e não mudanças de EC *number* foi proposta uma matriz de ocorrência. Nela, as colunas representam as características ou atributos (termos obtidos a partir das tags OC, RP e KW e processados conforme Seção 4.3.1) e as linhas representam instâncias do conjunto mudança ou controle. Uma posição i, j dessa matriz é 1, se a instância de índice i (uma dada entrada) possui a característica correspondente à coluna de índice j, e 0 caso contrário. A última coluna representa as classes para cada instância. As classes foram modeladas considerando o EC *number* de origem (antes da mudança) e o EC *number* de destino (depois da mudança), desse modo a classe 3.1.3.2 $\rightarrow$ 3.1.3.5 representa que uma dada entrada era anotada com EC 3.1.3.2 e essa anotação foi substituída por 3.1.3.5. Um fragmento de uma matriz de ocorrência que mostra algumas instâncias da mudança 3.1.3.2 $\rightarrow$ 3.1.3.5, que aconteceu da versão 5 para 6, e seu controle é mostrado na Tabela 4.5.

4.2.3 Criação do Banco de Dados

Um banco de dados foi criado utilizando o Sistema Gerenciador de Banco de Dados (SGBD) MySql versão 5.1.41. Nele estão armazenados os dados referentes às 18.727.155 mudanças de EC *number* observadas ao longo das versões 1 a 44 do Swiss-Prot tomadas par a par.

Um modelo entidade relacionamento (ER) é uma representação conceitual e abstrata de dados que captura as características do mundo real que são relevantes para uma de-

Tabela 4.5: Fragmento de matriz de ocorrência para a mudança 3.1.3.2 → 3.1.3.5 e seu controle.

id	nucleotide-binding	magnesium	eukaryota	metal-binding	signal	classe
Q8TUG3	1	1	0	1	0	3.1.3.2 → 3.1.3.5
O67004	1	1	0	1	0	3.1.3.2 → 3.1.3.5
Q9HY05	1	1	0	1	0	3.1.3.2 → 3.1.3.5
P58683	0	1	0	1	1	3.1.3.2 → 3.1.3.2
P34724	0	0	1	0	1	3.1.3.2 → 3.1.3.2
P44009	0	1	0	1	1	3.1.3.2 → 3.1.3.2

terminada aplicação. Os blocos básicos para a construção desse modelo são as entidades, relacionamentos e atributos. Uma entidade pode ser vista como algo capaz de existir de modo independente e que possa ser univocamente identificado. Os relacionamentos descrevem como as entidades se relacionam e são, de um modo geral, verbos. Um atributo é uma propriedade das entidades e seu significado depende das mesmas [Elmasri e Navathe (2008)].

O banco de dados aqui criado é extremamente simples, contendo apenas uma entidade (*mudança*) e seus atributos. Foi desenvolvido devido à necessidade de se poder visualizar e quantificar de modo prático e rápido as mudanças de anotação EC e suas características (como tamanho do prefixo comum, níveis escritos e apagados, bem como as linhas OC, RP e KW) nas diferentes versões do Swiss-Prot. Na figura 4.6 há um diagrama Entidade Relacinamento do banco de dados das mudanças de EC *number*. Na Tabela 4.6 há uma breve descrição dos atributos.

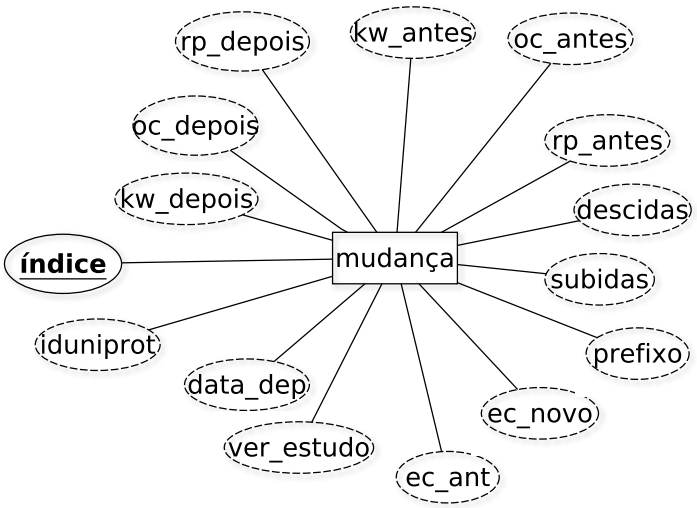


Figura 4.6: Diagrama ER do banco criado.

Tabela 4.6: Atributos da entidade *mudança*

Atributos	Significado
índice	chave primária
iduniprot	identificador da entrada
data_dep	data de depósito da entrada
ver_estudo	versão superior do par estudado
ec_ant	EC anterior
ec_novo	EC novo
prefixo	tamanho prefixo comum
subidas	níveis apagados
descidas	níveis escritos
rp_antes	RP antes da mudança
oc_antes	OC antes da mudança
kw_antes	KW antes da mudança
rp_depois	RP depois da mudança
oc_depois	OC depois da mudança
kw_depois	KW depois da mudança

4.3 Técnica

Nessa Seção serão descritas as técnicas utilizadas no processamento dos metadados selecionados para caracterizar as mudanças de anotação EC, bem como as técnicas adotadas na contrução de nossa estratégia baseada em aprendizagem supervisionada para previsão de tais mudanças.

4.3.1 Geração das Matrizes de Ocorrência

Nessa seção é descrito o processo para gerar as matrizes de ocorrência utilizadas pela estratégia de aprendizagem supervisionada proposta. Para cada tipo de mudança de EC *number* e para cada versão do Swiss-Prot na qual tal mudança aconteceu, os arquivos texto das entradas da base que experimentaram essa mudança e das entradas que formam o controle foi processado para extrair os metadados presentes nas linhas OC, RP e KW. Esses metadados passaram por um pré-processamento textual, que é um conjunto de técnicas aplicadas ao texto para reduzir as variações e aumentar as frequências observadas dos termos. As técnicas de pré-processamento aplicadas aos metadados foram:

- *Normalização*: tem objetivo de remover sinais de pontuação e acentos do texto e converter os caracteres para minúsculo.
- *Remoção de stop words*: trata-se da remoção de palavras extremamente comuns como, por exemplo, pronomes e artigos e que devido à grande frequência, não acres-

centam informação. Tais palavras são conhecidas como *stop words*.

- *N-grams*: um *n-gram* é uma sequência de n itens obtidos a partir de uma sequência de texto. Foi utilizado para capturar contexto presente nos metadados processados e para que pudessem ser considerados não apenas termos exatos, mas também aproximados. Por exemplo, dada a expressão abc, após o uso da técnica *n-grams* teríamos a, b, c, ab, bc. Aqui foram obtidos *n-grams* de tamanho até 2, pois para valores maiores que 2 a matriz de ocorrência gerada é muito grande (aproximadamente 5GB) e não pôde ser processada pelo passo seguinte (redução da dimensionalidade via SVD no software R)

- *Stemming*: reduz as palavras à sua raiz. O algoritmo de stemmer para a língua inglesa utilizado foi uma implementação em Java do *Porter stemming* [Porter et al. (1980)] obtido do *website*[5] do autor.

Abaixo há um exemplo dos metadados antes e depois do pré-processamento.
Antes:

```
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
OC   Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.

RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RP   GENOME REANNOTATION.
RP   LEVEL OF PROTEIN EXPRESSION [LARGE SCALE ANALYSIS].

KW   Complete proteome; Glycoprotein; Hydrolase; Reference proteome;
KW   Signal.
```

Depois:
Tag OC

```
eukaryota; fungi; dikarya; ascomycota; saccharomycotina;
saccharomycet; saccharomycetal; saccharomycetacea; saccharomyc
```

Tag RP

```
nucleotid; sequenc; larg; scale; genom; dna;
genom; reannot;
level; protein; express; analysi;
nucleotid sequenc; sequenc larg; larg scale; scale genom; genom dna
genom reannot;
level protein; protein express; express larg; larg scale; scale analysi
```

[5] http://tartarus.org/~martin/PorterStemmer/java.txt

Tag KW

```
complet; proteom; glycoprotein; hydrolas; refer; proteom;
signal;
complet proteom; refer proteom
```

Os termos das linhas OC e KW são originalmente separados por ponto e vírgula (;), enquanto na *tag* RP, se houver mais de um termo, esses são separados por vírgula (,). No exemplo fornecido, os metadados depois do pré-processamento estão separados por (;). Após o pré-processamento textual, cada um dos termos resultantes foi utilizado como um atributo da matriz de ocorrência. Dado um tipo de mudança de EC *number*, a versão da base em que ela ocorreu e uma entrada que sofreu tal mudança, metadados dessa entrada foram extraídos para todas as versões da base antes da mudança (até a versão imediatamente anterior à mudança).

4.3.2 Seleção de Mudanças de EC

Para nossa estratégia de aprendizagem supervisionada foram selecionados tipos de mudanças de EC que possuem pelo menos 10 exemplos ao longo das 44 versões estudadas do Swiss-Prot. Uma lista de tais mudanças está disponível no Apêndice A.3. Essa escolha se deve ao fato de que, na etapa Descritiva, foi realizada uma validação cruzada na qual o conjunto de dados foi dividido em 10 partes, chamada de *ten fold cross-validation*. Mais detalhes sobre os experimentos podem ser encontrados na Seção 4.3.4. Aqui, considera-se 3.1.3.2 $\rightarrow$ 3.1.3.5 como tipo de mudança de EC. Os identificadores do Swiss-Prot Q8TUG3 e O67004 são dois exemplos ou instâncias do tipo de mudança de EC 3.1.3.2 $\rightarrow$ 3.1.3.5. A Tabela 4.7 mostra alguns dados sobre mudanças de EC obtidos a partir das versões analisadas do repositório.

Para gerar as matrizes de ocorrência, todos os tipos de mudanças de EC com pelo menos 10 exemplos ao longo de todas as versões foram considerados (tipos de mudanças de EC utilizadas e descartadas estão representados na Figura 4.7). Porém, para mudanças como, por exemplo, $-.-.-.- \rightarrow 5.2.1.8$ que ocorre da versão 39 para 40, há uma enorme quantidade de exemplos de controle (288.932) representado por $-.-.-.- \rightarrow -.-.-.-$. Esse conjunto controle representa entradas que não possuíam anotação EC na versão 39 e permaneceram sem anotação EC na versão 40.

Assim, foi definido um limite superior para o número de instâncias de controle, caso contrário, o controle seria super representado nas matrizes de ocorrência e também aumentaria o custo computacional das tarefas de redução de dimensionalidade (Seção 4.3.3) e classificação (Seção 4.3.4). O limite superior escolhido para o número de instâncias de controle foi a mediana do número de exemplos para os tipos de mudança de EC, que é 27, dado que esse valor é mais representativo para o número de exemplos dos diferentes tipos de mudança de EC do que a média, que é 102,2 com desvio padrão 224,6. Mais detalhes são fornecidos na Figura 4.8.

Tabela 4.7: Mudanças de EC *number* nas 44 versões do Swiss-Prot

Total de pares de EC	Pares com ECs diferentes	Tipos de mudança de EC	Tipo de mudança de EC pelo menos 10 exemplos
18.727.155	55.908	1.968	508

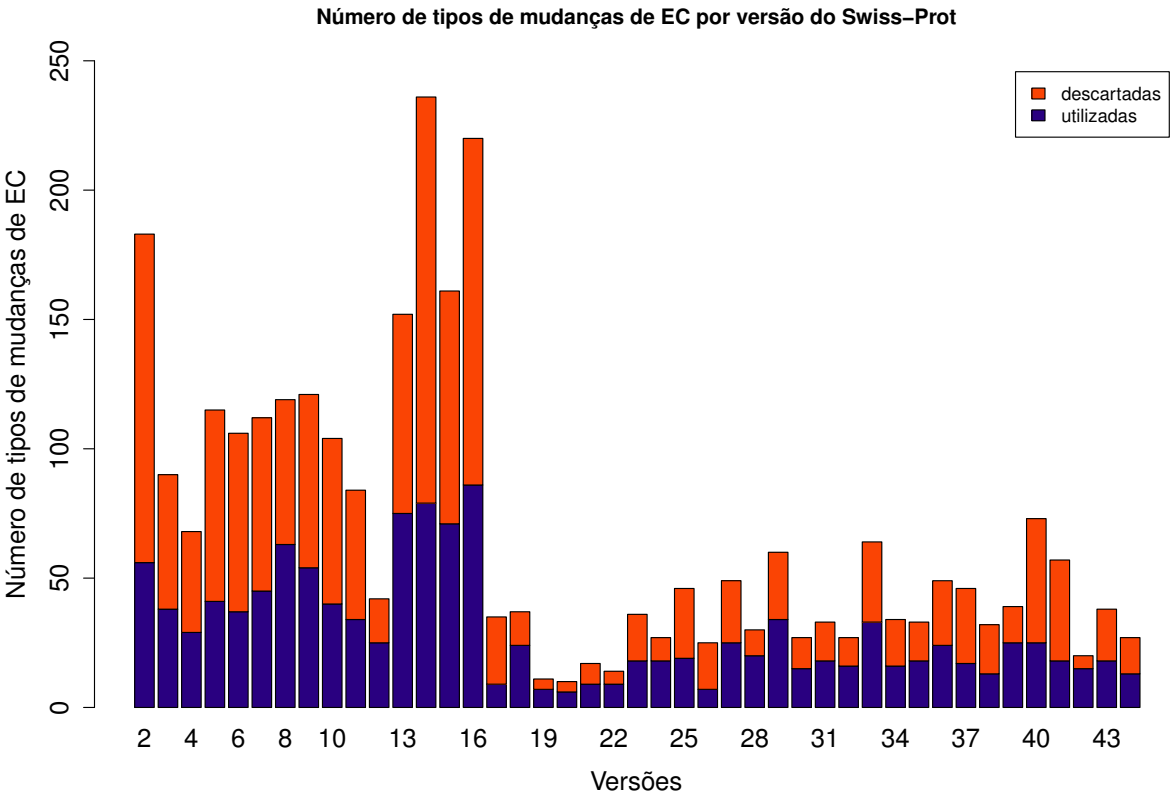


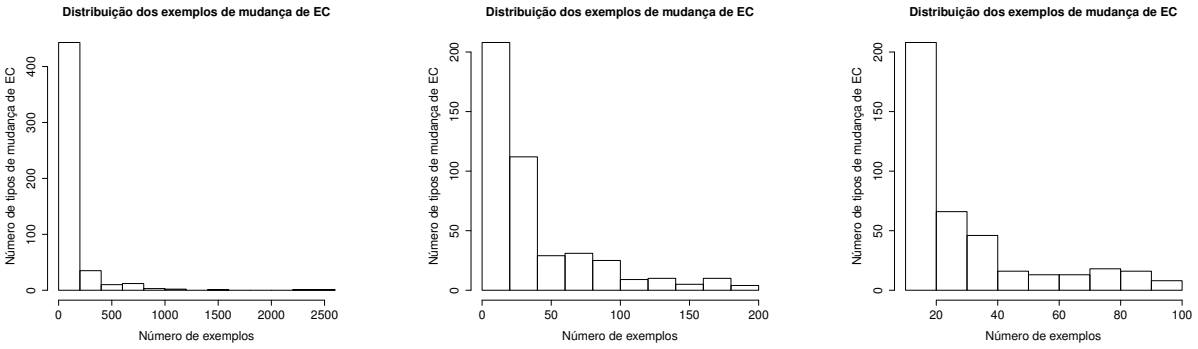
Figura 4.7: Número de tipos de mudanças de EC utilizadas e descartadas. Tipos de mudanças de EC com pelo menos 10 exemplos ao longo das 44 versões do Swiss-Prot foram usadas neste trabalho.

4.3.3 Redução de Dimensionalidade

Neste trabalho as matrizes de ocorrência passaram por um processo de redução de dimensionalidade realizado através da *Singular value decomposition* (SVD) ou decomposição por valor singular. A SVD é uma técnica da álgebra linear que se baseia no fato de que uma matriz A, de dimensões m por n, pode ser representada pelo produto $U\Sigma V^T$:

$$A = U\Sigma V^T \tag{4.1}$$

onde U é uma matriz m por m e suas colunas são os vetores singulares à esquerda de A; Σ é uma matriz diagonal m por n com os valores singulares de A em ordem decrescente; V é uma matriz n por n e suas colunas representam os vetores singulares à direita de A. Para fazer a compressão dos dados utilizados na tarefa de classificação, reduzindo o número de características ou atributos, ruído e ainda mantendo as relações relevantes entre os



(a) (b) (c)

Figura 4.8: O número de exemplos de mudanças de EC é apresentado no eixo x e o número de tipos de mudanças de EC é apresentado no eixo y. Em (a) o histograma mostra o número de exemplos de mudanças de EC para todos os 508 tipos de mudanças de EC com pelo menos 10 exemplos; em (b) somente tipos de mudanças com menos de 200 exemplos são apresentadas; em (c) tipos de mudanças com menos que 100 exemplos são exibidos. O limite superior definido para o número de exemplos do conjunto controle foi a mediana do número de exemplos de mudança de EC, que é 27. Tal valor é mais representativo que a média, que é 102,2 com desvio padrão 224,6.

termos, a matriz A pode ser aproximada pela matriz A_k (de posto k onde k é menor que o posto de A), ou seja:

$$A_k = U_k \Sigma_k V_k^T \tag{4.2}$$

Para obter A_k, são utilizados os primeiros k valores singulares de A, de modo que a matriz resultante tenha k características ou atributos. De acordo com [Eldén (2006)], a matriz A_k pode ser aproximada através da matriz D_k:

$$A_k = U_k \Sigma_k V_k^T = U_k(\Sigma_k V_k^T) = U_k(D_k) \tag{4.3}$$

Ou seja:

$$A_k = \Sigma_k V_k^T = D_k \tag{4.4}$$

A mesma estratégia para aproximar a matriz A_k utilizando D_k foi adotada em [Pires et al. (2011)], o que é razoável dado que, segundo [Tan et al. (2006)], padrões entre os atributos são capturados pelos vetores singulares à direita, ou seja, as colunas de V. Como afirma [Deerwester et al. (1989)], a escolha do k é empírica, assim aproximaçãoes para a matriz A com k variando de 1 a 100 foram geradas e a matriz que levou ao melhor modelo de classificação foi selecionada. É importante destacar que a redução de dimensionalidade através da técnica SVD permite reduzir o custo computacional e os requisitos de memória dos algoritmos aplicados na tarefa de classificação. A SVD foi utilizada e discutida de modo similar em diversos estudos [Berry et al. (1995); del Castillo-Negrete et al. (2007); Bécavin et al. (2011) e Deerwester et al. (1989)].

4.3.4 Classificação

A tarefa de classificação deste trabalho está representada no esquema da Figura 4.9 e foi realizada em duas etapas: *Descritiva*, com o objetivo de verificar se os metadados selecionados dos arquivos texto do Swiss-Prot são capazes de discriminar entradas que sofreram determinada mudança de EC das entradas em que o EC se manteve constante; *Previsiva*, com o propósito de utilizar o conhecimento já disponível a respeito das mudanças de EC para prever tais mudanças numa versão posterior do repositório.

Para caracterizar e prever mudanças de anotação EC, três experimentos foram realizados: Descritivo Multiclasse, Previsivo Multiclasse e Previsivo Origem Comum. Esses experimentos compõem a tarefa de classificação e são descritos a seguir.

4.3.4.1 Experimento Descritivo Multiclasse

Tem objetivo de verificar se os metadados selecionados nos arquivos texto do Swiss-Prot, OC, RP e KW, são capazes de discriminar entradas do repositório que sofreram um tipo de mudança específica na anotação EC de entradas nas quais o EC *number* se manteve constante.

Modelos de classificação foram gerados usando as matrizes de ocorrência (construídas a partir do conjunto de dados completo, ou seja, as 44 versões do Swiss-Prot) reduzidas através da SVD com k variando de 1 a 100 e foi selecionado o melhor modelo de classificação (ver Seção 4.3.6). Além disso, as matrizes de ocorrência foram geradas com e sem o uso das técnicas de pré-processamento textual *n-gram* e *stemming* e a melhor configuração foi mantida nos experimentos Previsivos posteriores.

O desempenho do modelo de classificação foi avaliado através da técnica de validação cruzada estratificada com 10 partições ou *ten fold cross-validation*. Segundo [Han e Kamber (2006)], tal técnica consiste em segmentar aleatoriamente o conjunto de dados em dez partições mutualmente exclusivas, chamadas *fold*, de tamanho aproximadamente igual. A cada execução, uma das partições é usada para testar o classificador e o restante das partições é usado para treino. Esse procedimento é repetido dez vezes de modo que cada partição é utilizada como teste apenas uma vez. No caso da validação cruzada estratificada a distribuição de classes das intâncias de cada partição é aproximadamente a mesma do conjunto de dados original.

4.3.4.2 Experimento Previsivo Multiclasse

O propósito do experimento Previsivo Multiclasse é fazer previsões de mudança de EC *number* utilizando um único classificador multiclasse. Aqui, os tipos de mudança de EC *number* previamente modelados no experimento Descritivo Multiclasse foram usados para construir um modelo de classificação e prever mudanças de EC. São ditos tipos de mudança de EC molelados aqueles que possuem $F_1 > 0,5$ (a métrica F_1 é detalhada na Seção 4.3.6). Somente esses tipos de mudança foram utilizados porque não é esperado

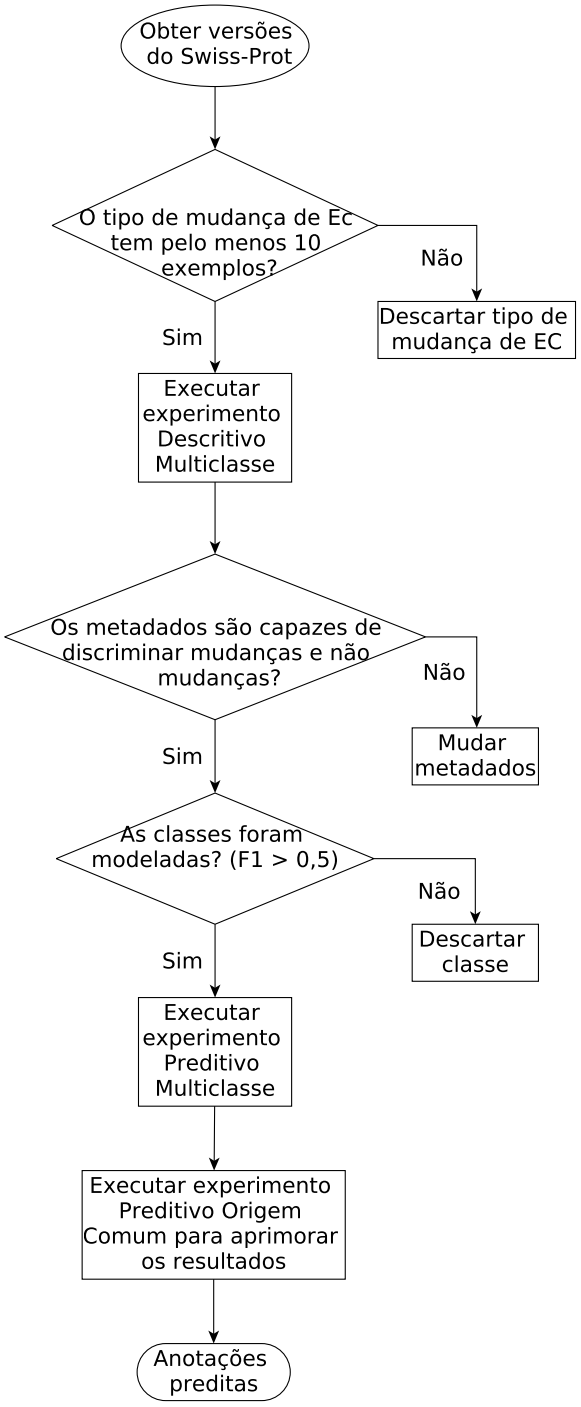


Figura 4.9: Fluxo da tarefa de classificação: Experimentos Descritivo Multiclasse, Previsivo Multiclasse e Previsivo Origem Comum.

que os tipos de mudanças que não foram nem caracterizados no experimento Descritivo possam ser previstos.

As matrizes de ocorrência para este experimento são geradas da mesma forma que no anterior e também passam pela redução de dimensionalidade através da SVD. O conjunto de dados de teste é composto pela última ocorrência de um determinado tipo de mudança de EC e o conjunto de dados de treino é formado pelas demais ocorrências desse tipo de mudança. Tomemos como exemplo a mudança $-. - . - .- \rightarrow$ 2.3.1.48, que ocorreu nas versões $2, 6, 8, 9, 12, 14, 15, 43, 44$. As entradas do Swiss-Prot que sofreram essa mudança nas versões $2, 6, 8, 9, 12, 14, 15, 43$ fazem parte dos dados de treino e as entradas que sofreram a mesma mudança na versão 44 fazem parte dos dados de teste.

Aqui é simulado um cenário no qual toda a informação disponível no repositório a respeito de um dado tipo de mudança de EC é utilizado para prever uma próxima mudança de EC desse mesmo tipo.

4.3.4.3 Experimento Previsivo Origem Comum

Este experimento foi realizado com o objetivo de aprimorar os resultados do experimento Previsivo Multiclasse. Os mesmos dados de tal experimento (tipos de mudanças de EC modeladas no experimento Descritivo Multiclasse) foram segmentados por origem comum e cada origem comum corresponde a um classificador. Origem comum é referente ao EC *number* associado a uma entrada antes da mudança de anotação. Tomemos com exemplo os tipos de mudanças de EC 2.1.1.$- \rightarrow$ 2.1.1.189, 2.1.1.$- \rightarrow$ 2.1.1.190 e seu controle 2.1.1.$- \rightarrow$ 2.1.1.$-$, que possuem a origem comum 2.1.1.$-$. Nesse caso, há um classificador específico para essa origem comum no qual as possíveis classes são 2.1.1.$- \rightarrow$ 2.1.1.189, 2.1.1.$- \rightarrow$ 2.1.1.190 e 2.1.1.$- \rightarrow$ 2.1.1.$-$.

Dessa maneira, há 24 possíveis origens comuns e, consequentemente, 24 classificadores que são mais especializados do que os dos classificadores multiclasse anteriores. Esse experimento foi realizado na expectativa de que seria mais fácil fazer previsões corretas com classificadores mais específicos, nos quais há menos classes a serem previstas.

As matrizes de ocorrência para este experimento são geradas da mesma forma que nos anteriores e também passam pela redução de dimensionalidade através da SVD. O conjunto de dados de teste é composto pela última ocorrência de um determinado tipo de mudança de EC e o conjunto de dados de treino é formado pelas demais ocorrências desse tipo de mudança, tal como realizado no experimento Previsivo Multiclasse.

4.3.5 Algoritmos de Classificação

Em cada um dos experimentos realizados nesse trabalho foram utilizados três algoritmos de classificação, Naïve Bayes (John e Langley, 1995), K-Nearest-Neighbor (KNN) ou K vizinhos mais próximos (Aha et al., 1991) e J48, uma implementação em Java do algo-

ritmo C4.5 (Quinlan, 1993). Tais algoritmos são descritos brevemente nas três próximas Seções.

4.3.5.1 K-Nearest-Neighbor

De acordo com [Han e Kamber (2006)], K-Nearest-Neighbor (KNN) ou K vizinhos mais próximos é uma técnica de classificação que baseia-se no aprendizado por analogia, comparando uma dada instância de teste com as instâncias de treino similares a ela. As instâncias de treino possuem n características ou atributos e assim representam um ponto no espaço n-dimensional. Quando é fornecida uma instância desconhecida (teste), o algoritmo obtém as K instâncias de treino mais próximas de tal instância no espaço n-dimensional. Essas K instâncias são os K vizinhos da tupla de teste. A proximidade dos vizinhos foi determinada utilizando a distância Euclidiana. Tal distância entre dois pontos ou instâncias $X_1 = (x_{11}, x_{12}, ..., x_{1n})$ e $X_2 = (x_{21}, x_{22}, ..., x_{2n})$ é calculada:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \tag{4.5}$$

A classe mais comum dentre os K vizinhos considerados é associada à instância de teste. A Figura 4.10 ilustra os três vizinhos mais próximos da instância X. Imagine que X é uma instância de teste e que há duas opções de classe: positivo, que representa a mudança 3.1.3.2 → 3.1.3.5 e negativo, que representa o controle 3.1.3.2 → 3.1.3.2. Nesse caso, a classe mais comum dentre os três vizinhos mais próximos de X, que é 3.1.3.2 → 3.1.3.5 será associada a esse ponto. Dessa maneira, X será classificado como 3.1.3.2 → 3.1.3.5. A escolha do K é empírica, em geral esse parâmetro é variado até obter a melhor classificação. Nesse método de classificação, as decisões se baseiam em informação local enquanto os classificadores baseados em árvore de decisão buscam um modelo global que melhor se adeque ao conjunto de dados.

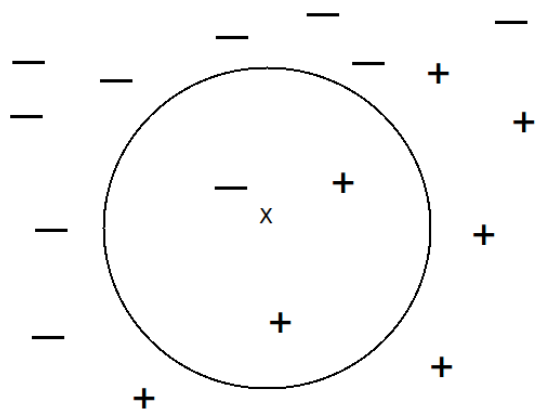


Figura 4.10: Exemplo de KNN para K=3.

4.3.5.2 Árvore de Decisão

Segundo [Han e Kamber (2006)] na classificação baseada em árvore de decisão, num primeiro momento é necessário construir a árvore com base em dados de treinamento, um processo conhecido como *decision tree induction* e num segundo momento essa árvore é utilizada para classificar instâncias novas (teste) que não participaram do processo de construção.

A árvore de decisão é uma estrutura na qual cada nó interno denota um teste num atributo, cada ramo é a saída de um teste e cada folha está associada a uma classe. O nó mais alto é chamado raiz. A proposta da árvore de decisão é escolher o atributo que melhor divide os dados (gerando partições mais puras possível, ou seja, na qual as classes não estejam misturadas) a cada etapa da construção da árvore. O algoritmo usa uma estratégia gulosa que faz decisões ótimas locais com relação ao atributo utilizado para particionar os dados. Na Tabela 4.8 há um exemplo de matriz de ocorrência utilizada na construção de uma árvore de decisão. As instâncias podem ser vistas como entradas do Swiss-Prot referentes à mudança 3.1.3.2 → 3.1.3.5 e seu controle 3.1.3.2 → 3.1.3.2.

Tabela 4.8: Matriz de ocorrência geradora da árvore de decisão da Figura 4.11.

Instância	magnesium binding	metal-binding	classe
A	1	1	3.1.3.2 → 3.1.3.5
B	1	1	3.1.3.2 → 3.1.3.5
C	1	1	3.1.3.2 → 3.1.3.5
D	1	0	3.1.3.2 → 3.1.3.2
E	0	1	3.1.3.2 → 3.1.3.2
F	1	0	3.1.3.2 → 3.1.3.2

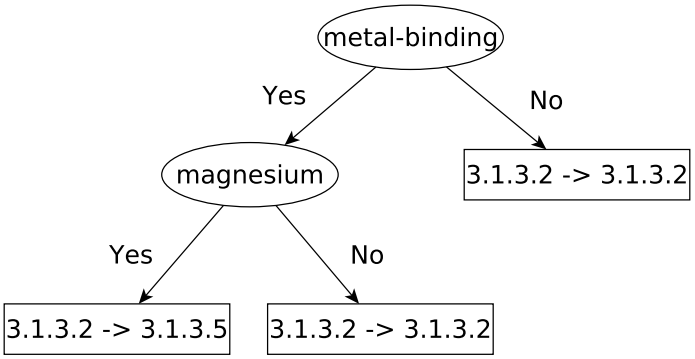


Figura 4.11: Árvore de decisão gerada com base nos dados da Tabela 4.8.

O algoritmo de Hunt [Tan et al. (2006)] é a base para diversos algorimos de árvore de decisão, dentre eles o J48, uma implementação em Java do algoritmo C4.5, utilizado

nesse trabalho. Segue uma definição de tal algoritmo: considere D_t como o conjunto de instâncias de treino associadas ao nó t e $y = \{y_1, y_2, ..., y_c\}$ os rótulos ou classes.

- *Passo 1*: Se todos os registros em D_t pertencem à mesma classe y_t, então t é um nó folha rotulado como yt.
- *Passo 2*: If D_t possui instâncias de classes diferentes, um atributo é selecionado como *condição de teste* para particionar os dados em conjuntos menores. Um nó filho é criado para cada resultado da *condição de teste* e os registros em D_t são distribuídos entre os nós filho com base no valor que possuem para o atributo selecionado como *condição de teste*. O algorimo é aplicado recursivamente a cada nó filho.

Para classificar uma instância X de classe desconhecida, os valores dos atributos dessa instância são testados a partir da raiz até uma folha da árvore de decisão. Essa folha contém a classe que deve ser associada à instância.

4.3.5.3 Naïve Bayes

A seguir descreveremos brevemente a técnica Naïve Bayes segundo Han e Kamber (2006).

Seja D um conjunto de tuplas com suas respectivas classes. Novamente, uma tupla pode ser uma linha da Tabela 4.5 por exemplo. Cada tupla é representada por um vetor de atributos n-dimensional $X = (x_1, x_2, ..., x_n)$, representando n medições realizadas para os n atributos $A_1, A_2, ..., A_n$.

Suponha que existam m classes, $C_1, C_2, ..., C_m$. Dada uma tupla X, o classificador irá prever que X pertence à classe com maior probabilidade a *posteriori* condicional a X. Ou seja, o classificador Naïve Bayes associa uma tupla X a uma classe C_i se e somente se

$$P(C_i|X) > P(C_j|X) \quad \text{for} \quad 1 \leq j \leq m, j \neq i.$$

Assim, $P(C_i|X)$ deve ser maximizada. Pelo teorema de Bayes,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Como $P(X)$ é constante para todas as classes, apenas $P(X|C_i)P(C_i)$ deve ser maximizado. Computar $P(X|C_i)$ é extremamente caro computacionalmente, assim assume-se que o valor dos atributos são condicionalmente independentes dada a classe da tupla. Desse modo:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

$$P(X|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times ... \times P(x_n|C_i).$$

Aqui x_k é referente ao valor do atributo A_k para a tupla X. O classificador associa à tupla X a classe C_i se e somente se:

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for} \quad 1 \leq j \leq m, j \neq i.$$

Ou seja, se $P(X|C_i)P(C_i)$ é máxima.

4.3.6 Estratégia de Avaliação dos Classificadores

Diversos experimentos foram realizados para selecionar o melhor modelo de classificação. Para cada uma das matrizes resultantes da SVD foram aplicados três algoritmos de classificação: Naïve Bayes, KNN com $K = 1, 3, 5, 7, 10$ (lembrando que no contexto do KNN o parâmetro K representa o número de vizinhos mais próximos considerados para fazer a classificação) e J48. Para avaliar o desempenho dos classificadores, as métricas F_1 (também chamada *F measure*) e *Area Under ROC Curve* (AUC) ou área sob a curva ROC [Fawcett (2006)] foram consideradas. Descreveremos brevemente tais métricas segundo [Tan et al. (2006)] e na sequência o procedimento adotado para selecionar o melhor resultado para cada um dos experimentos realizados.

A seguinte terminologia será adotada: *Verdadeiros positivos* (TP) corresponde às instâncias positivas previstas corretamente pelo modelo de classificação; *falsos negativos* (FN) são os exemplos positivos previstos incorretamente como negativos pelo modelo; *falsos positivos* (FP) corrresponde às instâncias negativas previstas incorretamente como positivas pelo modelo; *verdadeiros negativos (TN)* são instâncias negativas previstas corretamente como negativas pelo modelo.

Precisão (p) é a fração das instâncias que são efetivamente positivas dentre as que foram previstas como positivas pelo classificador ($p = \frac{TP}{TP+FP}$) e a revocação (r) é referente à fração de intstâncias realmente positivas dentre as que foram recuperadas pelo classificador ($r = \frac{TP}{TP+FN}$). Precisão e revocação podem podem ser resumidas pela métrica F_1, que é a média harmônica das mesmas e tende ao menor dos dois valores. Assim, um alto valor para F_1 garante que a precisão e revocação possuem valores altos.

$$F_1 = 2 \times \frac{p \times r}{p + r} \tag{4.6}$$

Receiver Operating Characteristic (ROC) *Curve* é uma estratégia de avaliação de classificadores que representa o compromisso entre a taxa de verdadeiros positivos[6] ($TPR = \frac{TP}{TP+FN}$) e a taxa de falsos positivos ($FPR = \frac{FP}{FP+TN}$). A taxa de verdadeiros positivos TPR é plotada no eixo y e a taxa de falsos positivos FPR é plotada no eixo x. Alguns pontos das curvas ROC possuem interpretação bem definida: ($TPR = 0$, $FPR = 1$) significa que todas as previsões estão incorretas; ($TPR = 1$, $FPR = 0$) re-

[6]Nota-se que a revocação é equivalente à taxa de verdadeiros positivos (TPR). Tais métricas são ainda conhecidas como sensibilidade.

presenta que as instâncias positivas e negativas foram previstas corretamente. O caso em que $TPR = 1$ e $FPR = 0$ representa o modelo de classificação ideal e nele a *Area Under ROC Curve* (AUC) ou área sob a curva ROC é 1. Assim, quanto mais próximo de 1 é AUC, melhor o modelo.

Nos experimentos Descritivo Multiclasse e Previsivo Multiclasse, para selecionar o melhor resultado para um determinado algoritmo de classificação, ou seja, a matriz resultante da SVD que levou a esse resultado, um esquema de votação foi aplicado. Um voto foi associado a cada resultado com maior valor para F_1 e um voto foi associado para cada resultado com maior valor para AUC. Note que mais de um resultado pode apresentar o maior valor para F_1 ou AUC. Nos casos em que houve empate, o resultado obtido através da matriz com menor número de atributos foi selecionado.

De maneira semelhante, após escolher o melhor resultado dentro de um dado algoritmo de classificação, o melhor resultado dentre todos os algoritmos foi selecionado através do mesmo esquema de votação. Nesse caso, quando houve empate, o resultado com maior valor para F_1 foi selecionado. Quando comparamos os resultados obtidos a partir de algoritmos diferentes, aqueles com valores similares de AUC podem ter valores de F_1 bem diferentes (consequentemente de precisão e revocação). Assim, priorizou-se os melhores valores para F_1 quando houve empate no esquema de votação.

No experimento Previsivo Origem Comum, o melhor resultado para os classificadores referentes à cada origem comum foi escolhido de acordo com o maior valor para F_1. Nesse experimento há casos em que mesmo classificadores com valores altos para AUC exibiam valores baixos para F_1 (e consequentemente para precisão e revocação). Assim priorizou-se a métrica F_1.

Capítulo 5

Resultados e Discussões

Neste capítulo são apresentados os resultados e discussões dos três experimentos realizados (Descritivo Multiclasse, Previsivo Multiclasse e Previsivo Origem Comum), bem como a comparação do ENZYMAP com o DETECT (técnica capaz de associar um EC *number* para uma dada sequência protéica). Apresentamos ainda alguns estudos de caso com previsões interessantes realizadas pelo ENZYMAP.

5.1 Experimento Descritivo Multiclasse

Nesta seção são apresentados os resultados do experimento da etapa descritiva, cujo objetivo é verificar se os metadados selecionados OC, RP e KW são capazes de discriminar entradas que experimentaram uma mudança de EC *number* de entradas em que a anotação EC permaneceu constante. Modelos de classificação foram gerados usando as 100 matrizes de ocorrência resultantes da SVD como entrada para cada um dos três algoritmos de classificação (KNN; J48, que é uma implementação em Java do algoritmo C4.5, e Naïve Bayes). O desempenho dos modelos foram avaliados através de uma validação cruzada com 10 partições.

A Tabela 5.1 mostra os melhores resultados para cada algoritmo de classificação. Os resultados completos são fornecidos no Apêndice A.1. Com exceção do algoritmo de Naïve Bayes, os classificadores são capazes de prever mudanças de EC *number* pois precisão, revocação e F_1 apresentaram valores em torno de 70% e AUC está acima de 90% para os resultados dos demais algoritmos. O KNN com 1 vizinho foi selecionado como o melhor resultado devido ao alto valor para F_1, que foi considerado pelo esquema de votação.

A Tabela 5.2 mostra o número de exemplos das classes modeladas ($F_1 > 0{,}5$) e não modeladas separadas por instâncias de controle e mudança. As classes são os tipos de controle e mudança de EC (por exemplo 3.1.3.2 $\rightarrow$ 3.1.3.5 e 3.1.3.2 $\rightarrow$ 3.1.3.2). É importante destacar que, em geral, as classes modeladas possuem mais exemplos do que as não modeladas, o que é evidenciado pela média e mediana que possuem valores mais altos para as classes modeladas. Isso indica que classes com maior número de exemplos

Tabela 5.1: Melhor desempenho de previsão de mudança de EC para cada técnica utilizando validação cruzada de 10 partições.

Número de votos	Algoritmo	Número de atributos	TPR	FPR	Prec.	Revoc.	F_1	AUC
0	Naïve Bayes	97	0,507	0,005	0,672	0,507	0,534	0,929
1	**KNN_K1**	**38**	**0,741**	**0,005**	**0,739**	**0,741**	**0,738**	**0,953**
0	KNN_K3	100	0,718	0,009	0,712	0,718	0,709	0,963
1	KNN_K5	100	0,711	0,013	0,697	0,711	0,696	0,966
1	KNN_K7	96	0,702	0,016	0,683	0,702	0,682	0,966
1	KNN_K10	81	0,691	0,022	0,664	0,691	0,664	0,966
0	J48	88	0,738	0,006	0,728	0,738	0,727	0,934

apresentam melhores resultados.

Tabela 5.2: Classes modeladas e não modeladas para o melhor resultado (KNN_K1 com 38 características ou atributos): média, desvio padrão, mediana e total de instâncias para classes modeladas ($F_1 > 0,5$) e não modeladas ($F_1 < 0,5$) separadas por controle e mudança. A última coluna representa o número de classes.

	Classe	Média	Desvio padrão	Mediana	Total de instâncias	Número de classes
Modeladas	Todas	183,1	1155,8	37	63.540	347
	Controle	292,6	2119,7	34	28.972	99
	Mudança	139,4	286,3	37	34.568	248
Não modeladas	Todas	61,2	123,6	23	19.414	317
	Controle	36,1	48,4	27	2.059	57
	Mudança	66,8	134,0	21	17.355	260

Na Tabela 5.3, as médias aritmética e ponderada das métricas utilizadas para avaliar o desempenho do classificador foram calculadas separadamente para classes que representam mudanças de EC e para as que representam controle. Os valores são melhores para o conjunto controle do que para as mudanças, o que é esperado pois é mais difícil prever uma mudança de anotação do que uma anotação que se manteve constante dado que há classes de controle que são muito numerosas.

Esse experimento nos dá evidências de que os metadados OC, RP e KW são capazes de discriminar e caracterizar as entradas que sofreram um tipo específico de mudança de EC das entradas em que a anotação não mudou porque, mesmo num classificador multiclasse com 664 classes (um problema de classificação complexo já que a probabilidade de acertar uma classe ao acaso é 1/664 ou 0,15% enquanto num classificador binário essa probabilidade é 1/2 ou 50%) os valores de 0.74 para F_1 e 0.95 para AUC indicam que nosso classificador não é aleatório (quando F_1 e AUC ficam em torno de 0,5)

Tabela 5.3: Médias aritmética e ponderada para as classes de controle e mudança do melhor resultado (KNN_K1 com 38 características ou atributos)

Métricas	Controle		Mudança	
	Média aritmética	Média ponderada	Média aritmética	Média ponderada
TPR	0,549	0,879	0,511	0,659
FPR	0,000	0,010	0,000	0,002
Precisão	0,592	0,864	0,529	0,664
Revocação	0,549	0,879	0,511	0,659
F_1	0,564	0,870	0,515	0,659
AUC	0,892	0,969	0,893	0,942

5.2 Experimentos Previsivos

Nesses experimentos toda a informação disponível no repositório Swiss-Prot a respeito de mudanças de EC *number* é utilizada para prever mudanças em versões futuras. Os tipos de mudança de EC *number* previamente modelados no experimento Descritivo Multiclasse foram usados para construir um modelo de classificação e prever mudanças de EC. São chamados tipos de mudança de EC modelados aqueles que possuem $F_1 > 0,5$ (a métrica F_1 é detalhada na Seção 4.3.6). Somente esses tipos de mudança foram utilizados porque não é esperado que os tipos de mudanças que não foram nem caracterizados no experimento Descritivo possam ser previstos.

O conjunto de dados de teste é formado pela última ocorrência de um dado tipo de mudança de EC e o conjunto de dados de treino é composto pelas demais ocorrências de tal tipo de mudança. Tomemos como exemplo a mudança $-.-.-.- \rightarrow$ 2.3.1.48, que ocorreu nas versões $2, 6, 8, 9, 12, 14, 15, 43, 44$. As entradas que sofreram essa mudança de EC nas versões $2, 6, 8, 9, 12, 14, 15, 43$ fazem parte dos dados de treinamento e as entradas que sofreram essa mesma mudança na versão 44 formam os dados de teste.

5.2.1 Multiclasse

O objetivo do experimento Previsivo Multiclasse é prever mudanças de EC para a última ocorrência de cada tipo de mudança de EC utilizando um único classificador multiclasse que compreende todas as possíveis classes. Esse experimento é similar ao Descritivo, porém aqui são analisadas somente as mudanças modeladas, totalizando 361 classes.

Os resultados são fornecidos na Tabela 5.4 e os resultados completos estão disponíveis no Apêndice A.2. As médias aritmética e ponderada foram calculadas separadamente para a o conjunto mudança e controle e são mostradas na Tabela 5.5. Os valores de precisão, revocação, F_1 e AUC são menores que os do experimento Descritivo. Quando a última versão na qual uma mudança aconteceu é reservada para teste, exemplos de treinamento

são perdidos, o que impacta na qualidade dos resultados.

Dessa maneira, para aprimorar os resultados são necessários mais exemplos de treino. Outra alternativa seria uma tarefa de classificação mais especializada. Como não temos controle sobre a ocorrência e a quantidade de mudanças de EC, as mudanças foram segmentadas pela origem comum e uma tarefa de classificação mais especializada foi realizada.

Tabela 5.4: Experimento Previsivo Multiclasse com dados de treino e teste: melhor desempenho para cada técnica.

Número de votos	Algoritmo	Número de atributos	TPR	FPR	Prec.	Revoc.	F_1	AUC
1	Naive Bayes	65	0,200	0,039	0,344	0,200	0,236	0,692
1	**KNN_K1**	**13**	**0,316**	**0,075**	**0,406**	**0,316**	**0,247**	**0,652**
0	KNN_K3	12	0,283	0,066	0,399	0,283	0,232	0,657
0	KNN_K5	57	0,282	0,086	0,502	0,282	0,231	0,635
0	KNN_K7	13	0,260	0,049	0,238	0,260	0,225	0,671
0	KNN_K10	100	0,270	0,085	0,497	0,270	0,225	0,666
1	J48	16	0,296	0,084	0,249	0,296	0,221	0,692

Tabela 5.5: Médias aritmética e ponderada para as classes de controle e mudança do melhor resultado (KNN_K1 com 38 características ou atributos)

	Controle		Mudança	
Métricas	Média aritmética	Média ponderada	Média aritmética	Média ponderada
TPR	0,515	0,828	0,092	0,255
FPR	0,016	0,229	0,001	0,002
Precisão	0,585	0,524	0,114	0,269
Revocação	0,515	0,828	0,092	0,255
F_1	0,512	0,605	0,078	0,188
AUC	0,804	0,826	0,641	0,721

5.2.2 Origem Comum

Esse experimento foi realizado para aprimorar os resultados da classificação do experimento Previsivo Multiclasse. O conjunto de dados foi segmentado por origem comum e cada origem comum corresponde a um classificador específico. Origem comum é referente ao EC *number* associado a uma entrada antes da mudança de anotação. Tomemos com exemplo os tipos de mudanças de EC 2.1.1.– $\rightarrow$ 2.1.1.189, 2.1.1.– $\rightarrow$ 2.1.1.190 e seu controle 2.1.1.– $\rightarrow$ 2.1.1.–, que possuem a origem comum 2.1.1.-. Nesse caso,

há um classificador específico para essa origem comum no qual as possíveis classes são 2.1.1.− → 2.1.1.189, 2.1.1.− → 2.1.1.190 e 2.1.1.− → 2.1.1.−.

Há 24 origens comuns e consequentemente 24 classificadores que são mais especializados que o classificador do experimento Previsivo Multiclasse anterior, o que aumenta as chances de fazer previsões corretas (há menos opções de classes para cada classificador). Como detalhado na Seção 4.3.6, 100 matrizes resultantes da SVD foram processadas por três algoritmos de classificação: KNN com $K = 1, 3, 5, 7, 10$, Naïve Bayes e J48. Esse processo foi repetido para cada uma das 24 origens comuns e o melhor resultado foi selecionado de acordo com os maiores valores para AUC.

A Tabela 5.6 mostra um resumo dos 24 melhores resultados (um para cada origem comum). A média desses 24 resultados foi calculada para resumí-los e pode ser vista na Tabela 5.7. A média tem valores de precisão, revocação e F_1 maiores que 0,86. Porém, há uma origem comum, $-.-.-.-$, que teve um resultado significativamente pior (TPR=0,341, FPR=0,102, precisão=0,662, revocação=0,341, F_1=0,305 and AUC=0,664) quando comparado à média da Tabela 5.7. Essa origem tem um peso alto pois contém 36 tipos de mudanças de EC (ou classes) e 2.631 instâncias de mudanças de EC.

Na Tabela 5.8, as médias aritmética e ponderada foram calculadas separadamente para as classes mudança e controle. Na coluna que contém a média ponderada para o conjunto mudança, a precisão (0,756) é maior que a revocação (0,274), também chamada de taxa de verdadeiros positivos (TPR) ou sensibilidade. Isso indica que é difícil prever uma mudança, mas se o classificador prevê uma instância como mudança, a chance de que a previsão esteja correta é alta.

É importante destacar que apesar de algumas métricas exibirem valor baixo para as mudanças se comparado ao controle, os dados considerados como resposta correta (anotações do Swiss-Prot) podem apresentar insconsistências ou até mesmo erros, dado que mudanças na anotação EC acontecem ao longo do tempo no repositório. Além disso, essas métricas calculadas a partir dos resultados do Weka não consideram resultados parciais (quando os níveis mais altos do EC estão corretos). Assim, para fornecer uma comparação mais justa entre as anotações do Swiss-Prot e nossos resultados, as anotações previstas foram comparadas com as do Swiss-Prot considerando de 1 a 4 níveis do EC *number*.

Para estender essa comparação, a ferramenta DETECT (Hung et al., 2010) foi utilizada para fazer previsões de EC para as mesmas entradas do Swiss-Prot fornecidas ao ENZYMAP e assim previsões feitas pela nossa metodologia, pelo DETECT e as anotações do Swiss-Prot foram comparadas. DETECT foi escolhido para essa comparação porque é uma técnica relativamente recente (de 2010) capaz de retornar um EC *number* com base em alinhamentos local e global das sequências. Essa ferramenta recebe como entrada as sequências de resíduos das proteínas no formato FASTA separadas por organismo e retorna previsões de EC *number*. Apesar de nossa estratégia e o Detect serem essencialmente diferentes, as previsões de EC podem ser utilizadas de modo complementar para

Tabela 5.6: Resultado do experimento Origem Comum. Cada linha corresponde ao melhor resultado para cada classificador (origem comum).

Origem	TPR	FPR	Prec.	Revoc.	F-1	AUC	Algoritmo	Número de atributos	Número de classes
-.-.-.-	0,341	0,102	0,662	0,341	0,305	0,664	KNN_K1	1	36
1.1.1.-	1,000	0,000	1,000	1,000	1,000	1,000	KNN_K1	11	2
1.10.2.2	1,000	0,000	1,000	1,000	1,000	1,000	KNN_K5	2	2
1.9.3.1	0,699	0,330	0,704	0,699	0,701	0,683	KNN_K10	2	2
2.-.-.-	0,418	0,314	0,773	0,418	0,321	0,624	Naïve Bayes	1	3
2.1.1.-	0,897	0,236	0,913	0,897	0,905	0,933	KNN_K7	74	3
2.3.1.-	0,964	0,964	0,930	0,964	0,947	0,907	KNN_K10	100	2
2.4.-.-	0,967	0,004	0,975	0,967	0,969	0,981	J48	13	2
2.7.1.-	0,882	0,031	0,925	0,882	0,891	0,894	KNN_K3	89	2
2.7.3.-	1,000	0,000	1,000	1,000	1,000	1,000	J48	30	2
2.7.7.48	0,659	0,302	0,700	0,659	0,663	0,545	KNN_K3	40	2
2.7.7.6	0,933	0,007	0,960	0,933	0,940	0,963	Naïve Bayes	32	2
3.-.-.-	0,903	0,014	0,945	0,903	0,914	0,944	KNN_K1	5	2
3.1.-.-	0,964	0,964	0,930	0,964	0,947	0,611	KNN_K1	100	2
3.1.13.-	0,946	0,058	0,951	0,946	0,946	0,905	KNN_K10	65	2
3.1.2.15	0,959	0,000	1,000	0,959	0,979	0,000	KNN_K10	100	2
3.2.1.18	0,931	0,931	0,867	0,931	0,898	0,500	J48	10	2
3.4.22.-	1,000	0,000	1,000	1,000	1,000	1,000	KNN_K10	100	2
3.4.25.-	0,995	0,331	0,995	0,995	0,995	0,970	KNN_K10	41	2
3.6.3.14	0,935	0,046	0,944	0,935	0,935	0,949	Naïve Bayes	12	2
4.2.2.-	0,718	0,635	0,800	0,718	0,622	0,796	KNN_K1	2	2
5.-.-.-	1,000	0,000	1,000	1,000	1,000	1,000	KNN_K1	4	2
6.-.-.-	0,900	0,900	0,810	0,900	0,853	0,500	Naïve Bayes	100	2
6.4.1.2	1,000	0,000	1,000	1,000	1,000	1,000	KNN_K1	10	2

Tabela 5.7: Média dos melhores resultados do experimento Origem Comum da Tabela 5.6

TPR	FPR	Precisão	Revocação	F_1	AUC
0,876	0,257	0,908	0,876	0,864	0,807

Tabela 5.8: Médias aritmética e ponderada para as classes de controle e mudança do melhor resultado para o experimento Origem Comum.

Métricas	Controle		Mudança	
	Média aritmética	Média ponderada	Média aritmética	Média ponderada
TPR	0,881	0,908	0,269	0,274
FPR	0,287	0,301	0,038	0,070
Precisão	0,855	0,741	0,287	0,756
Revocação	0,881	0,908	0,269	0,274
F-1	0,859	0,806	0,249	0,293
AUC	0,812	0,825	0,687	0,643

aprimorar as anotações.

5.3 Comparação entre ENZYMAP, DETECT e Swiss-Prot

O mesmo conjunto de dados fornecido como entrada aos experimentos da etapa Previsiva da Seção 5.2, com 3.582 mudanças de EC *number*, foi também fornecido ao DETECT 1.0[1]. O ENZYMAP fez 3.582 previsões de EC enquanto o DETECT fez 1.876. Ambos os métodos foram comparados às anotações do Swiss-Prot. A Figura 5.1 apresenta a comparação entre as previsões realizadas pelo ENZYMAP e pelo DETECT. Aqui as previsões foram comparadas por níveis do EC, de 1 a 4.

Para o primeiro nível (mais à esquerda), Figura 5.1 (a), 56% das previsões feitas pelo ENZYMAP concordam com Swiss-Prot enquanto para o DETECT esse percentual é de 49%. Considerando-se os dois métodos juntos, a interseção das mesmas com o Swiss-Prot representa 72% das anotações desse repositório, o que mostra que utilizá-los em conjunto aumenta a cobertura das anotações.

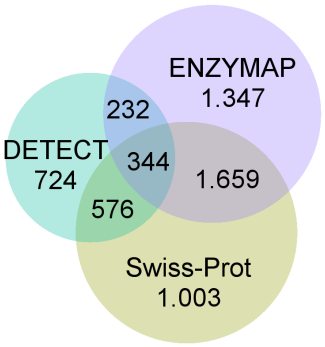
Para os níveis 2 a 4, o percentual de previsões feitas pela nossa proposta que está de acordo com as anotações do Swiss-Prot é maior que o mesmo percentual para o DETECT e os dois métodos juntos cobrem mais que 64% das anotações do repositório, como mostrado na Tabela 5.9. Entretanto, para o nível 4 o percentual de previsões feitas pelo DETECT que estão de acordo com as anotações do Swiss-Prot diminui significativamente chegando a 32%, enquanto para o ENZYMAP esse percentual é de 49%. Quanto mais específica a anotação, mais difícil é a previsão, o que pode levar a um tipo comum de erro de anotação chamado *overprediction* (quando um método de anotação associa mais níveis do que deveria) [Schnoes et al. (2009)]. Assim, nesse aspecto nossa proposta supera o

[1]http://www.compsysbio.org/projects/DETECT/

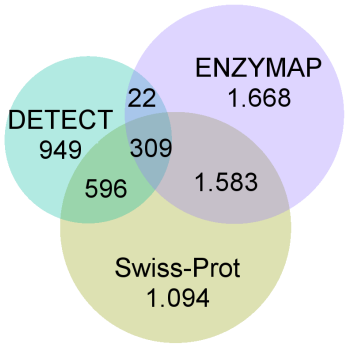
DETECT, pois consegue associar níveis mais específicos e ainda assim acertar em mais casos.

Tabela 5.9: Previsões feitas por ambos os métodos para os 4 níveis do EC *number*. As duas primeiras linhas correspondem ao percentual das previsões feitas pelo ENZYMAP e pelo DETECT que estão de acordo com as anotações do Swiss-Prot. Cobertura representa o percentual de anotações do repositório coberto quando os dois métodos são utilizados de modo complementar.

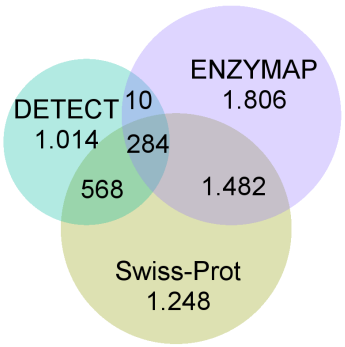
	Nível 1	Nível 2	Nível 3	Nível 4
ENZYMAP (%)	56	53	49	49
DETECT (%)	49	48	45	32
Cobertura (%)	72	70	65	64



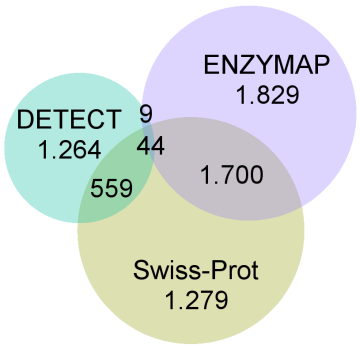
(a)



(b)



(c)



(d)

Figura 5.1: Comparação entre previsões de EC *number* realizadas pelo DETECT e pelo ENZYMAP com as anotações do Swiss-Prot (valores absolutos). Em (a) o primeiro nível da anotação EC é comparado; De modo semelhante, em (b), (c) e (d) 2, 3 e 4 níveis da anotação EC são considerados.

5.3.1 Estudos de Caso

Algumas previsões interessantes são discutidas abaixo.

A entrada com identificador Q5FWH2 foi prevista como 6.3.2.− para dados de teste da versão 44 e essa entrada realmente sofreu a mudança $-.-.-.- \rightarrow$ 6.3.2.− da versão 43 para 44. Nesse caso, o ENZYMAP fez a previsão correta de um EC de três níveis para uma entrada não anotada. O DETECT não retornou um resultado para a entrada Q5FWH2.

Na origem comum -.-.-.-, o ENZYMAP fez a previsão de que a entrada com identificador Q2GGA6 deveria ser anotada como 3.6.5.- (hidrolase que atua em guanosina trifosfato ou GTP) e tal entrada realmente sofreu a mudança $-.-.-.- \rightarrow$ 3.6.5.− da versão 27 para 28. O Q2GGA6 é denominado fator de alongamento 4 e atua como fator de correção na tradução [Qin et al. (2006)]. A previsão foi realizada utilizando dados de treino anteriores à versão 15, de março de 2009 (inclusive), e dados de teste da versão 27, de novembro de 2010. Nossa estratégia foi capaz de prever três níves de EC para uma entrada não anotada. O DETECT retornou o EC 2.7.7.4 para essa entrada.

O DETECT e o ENZYMAP fizeram a previsão de que as entradas O61694 e O94581, ambos subunidades da citocromo C oxidase de um inseto e de levedura respectivamente, deveriam receber o EC *number* 1.9.3.1, referente a oxidoredutases que agem em grupos heme como doadores de elétrons e oxigênio como receptores. No Swiss-Prot, um EC *number* não é associado a essas entradas, indicando que não são enzimas. A questão é que a citocromo C oxidase é um grande complexo protéico transmembrana, com várias subunidades, o que pode introduzir ambiguidade. A previsão do ENZYMAP e do DETECT é correta se as proteínas em questão (O61694 e O94581) são consideradas como parte do complexo enzimático da citocromo C oxidase. Entretanto, essas subunidades específicas podem não ter função catalítica direta. Esse caso ilustra a dificuldade de realizar a anotação quando as entradas pertencem a complexos protéicos de vários domínios ou cadeias, com unidades funcionais diferentes. De fato, até a versão 15 (março de 2009) do Swiss-Prot, o EC *number* 1.9.3.1 estava associado a essas entradas.

De acordo com o ENZYMAP a entrada Q9IH62 deveria permanecer anotada com o EC 3.2.1.18, entretanto, da versão 14 para 15, tal entrada perdeu a anotação EC no Swiss-Prot. Conforme citado na Seção 1, essa proteína apresenta mais de 50% de similaridade de sequência com as hemaglutinina-neuraminidases, um grupo de enzimas associado à ligação viral e ao processo de fusão na célula hospedeira. As estruturas da glicoproteína G de Hendra e Nipah virus foram resolvidas (identificadores 2VSK e 2VSM, respectivamente) e possuem o motivo estrutural conhecido como *six-blade β propeller* (uma espécie de hélice formada por 6 folhas beta), típico dessas hidrolases (hemaglutinina-neuraminidases) [Bowden et al. (2008)]. Um alinhamento estrutural com uma neuramidase legítima do virus Parainfluenza Type III (indentificador 1V3D no PDB), que também pertence à mesma família Paramyxoviridae de Henipavirus, resultou num RMSD menor que 2,0

Å[Lawrence et al. (2004)]. Entretanto, apesar da similaridade no nível de sequência e estrutura, hoje sabe-se que as glicoproteínas G de Henipavirus não são enzimas, e sua atividade é de hemaglutinina, realizando interações proteína-proteína com receptores do hospedeiro [Bowden et al. (2008)]. No momento em que esse texto era escrito, o PDB ainda indicava as proteínas (2VSK e 2VSM) como hidrolases. O DETECT também retornou o EC *number* 3.2.1.18 para a entrada Q9IH62.

Na origem comum 2.4.-.-, o ENZYMAP fez a previsão de que a entrada com identificador Q5NDL2 deveria ser anotada como 2.4.1.-. Essa entrada é uma transferase de N-acetilglucosamina ligada a oxigênio do organismo *Homo Sapiens* (humano). Tal previsão foi considerada como erro pelo Weka, pois, no conjunto de dados de teste, a anotação segundo o Swiss-Prot era 2.4.-.-. Contudo, na versão 2012_07, de julho de 2012 (lançada depois das versões utilizadas em nosso estudo), essa entrada recebeu o EC 2.4.1.255 no Swiss-Prot. A previsão do ENZYMAP foi realizada utilizando dados de treino anteriores à versão 2011_02 de fevereiro de 2011 (inclusive) e dados de teste da versão 2011_03 de março de 2011, de modo que nossa estratégia foi capaz de antecipar o terceiro nível do EC para a entrada Q5NDL2 16 meses antes de tal anotação ser disponibilizada no Swiss-Prot. O DETECT não retornou um resultado para essa entrada.

Capítulo 6

Conclusões

Nesse trabalho avaliamos se os metadados protéicos do repositório biológico UniProt/Swiss-Prot podem ser utilizados para prever mudança de anotação EC. Uma estratégia baseada em aprendizagem supervisionada foi proposta para caracterizar e prever mudanças de EC *number* nos dados temporais desse repositório. Tal estratégia foi denominada *ENZYmatic Metadata Annotation Predictor* (ENZYMAP). Nossa proposta pode ser utilizada como método complementar de anotação automática que ajuda a aprimorar a qualidade e confiabilidade de anotações de enzimas através do uso de metadados já disponíveis no repositório, sugerindo possíveis correções e antecipando mudanças na anotação. O artigo resultante dessa tese, intitulado *ENZYMAP: Exploiting protein metadata for modeling and predicting annotation changes in UniProt/Swiss-Prot*, foi submetido à revista Bioinformatics (Oxford).

Num primeiro momento, para realizar uma exploração inicial dos dados coletamos as versões disponíveis da base e modelamos as mudanças de EC em termos dos parâmetros prefixo comum, generalizações e especializações considerando a natureza numérica e hierárquica do EC. Uma ferramenta de visualização que segmenta as mudanças de anotação EC do Swiss-Prot com relação aos parâmetros citados foi proposta e permitiu ter um panorama geral das mudanças de anotação, identificando tendências de especialização e generalização. Essa etapa deu origem à uma ferramenta de visualização interativa chamada Advise e a um artigo [Silveira et al. (2012)] publicado no *IEEE Symposium on Biological Data Visualization (BioVis), 2012*.

Na sequência foram selecionados metadados do Swiss-Prot (OC, RP e KW) capazes de descrever entradas que sofreram um tipo específico de mudança de EC das entradas cuja a anotação se manteve constante. As matrizes de ocorrência foram propostas para modelar as mudanças de EC *number* em termos dos metadados do Swiss-Prot e serviram como insumo para a estratégia de aprendizagem supervisionada.

Três experimentos foram realizados para caracterizar e prever as mudanças de anotação EC: *Descritivo Multiclasse*, no qual conclui-se que os metadados selecionados (as linhas OC, RP e KW dos arquivos texto do Swiss-Prot) foram capazes de discriminar entradas

que experimentaram uma mudança específica no EC *number* daquelas entradas em que a anotação permaneceu constante; *Previsivo Multiclasse* nos indicou que prever a última ocorrência de um determinado tipo de mudança de EC utilizando um único classificador multiclasse com número escasso de exemplos não foi possível; *Previsivo Origem Comum*, no qual conclui-se que é possível fazer previsão de um determinado tipo de mudança de EC utilizando classificadores mais especializados (um para cada origem comum) mesmo com a restrição do número de exemplos.

As previsões feitas pelo ENZYMAP (experimento Previsivo Origem Comum) foram comparadas às previsões feitas pelo DETECT e ambas foram confrontadas com as anotações do Swiss-Prot pois os resultados obtidos a partir do Weka não consideram previsões parciais (quando acerta alguns níveis). Assim, para fornecer uma comparação mais justa entre os métodos, as anotações foram comparadas considerando de 1 a 4 níveis do EC *number*. O percentual de previsões feitas pelo ENZYMAP que está de acordo com o Swiss-Prot é maior que o mesmo percentual para o DETECT para todos os quatro níveis do EC. Desse modo, o ENZYMAP supera o DETECT.

6.1 Perspectivas

Nesta seção levantamos alguns pontos referentes aos desdobramentos futuros desse trabalho.

Queremos investigar se é possível associar um índice de confiabilidade às nossas previsões para ajudar o especialista do domínio a decidir se deve aceitar uma dada previsão. Tomemos como exemplo a previsão correta feita pelo ENZYMAP de que o identificador A0KGY4 sofreria a mudança de EC $2.-.-.- \rightarrow 2.4.2.-$. É desejável que um alto valor de confiabilidade esteja associado a essa previsão, por exemplo, 90%. Em contrapartida, é desejável que valores baixos estejam associados a previsões incorretas. Um possível caminho seria utilizar a probabilidade que os algoritmos de classificação liberam para cada previsão para apoiar na construção de tal índice de confiabilidade.

Gostaríamos também de explorar se há outros metadados capazes de descrever e prever mudanças de anotação EC, além de caracterizar e aferir quantitativamente a dificuldade dos dados através de medidas como, por exemplo, entropia (referente ao grau de incerteza dos dados) e informação mútua (referente a atributos que carregam a mesma informação). Dados que apresentam valores altos para entropia e informação mútua são mais difíceis de classificar.

Para elucidar quais foram os metadados mais relevantes para fazer as previsões, uma informação que não é preservada devido à utilização da SVD, estamos considerando a utilização de Análise de Conceitos Formais (FCA) de modo complementar à estratégia proposta nesse trabalho. Tal técnica, que pode ser considerada como Mineração de Dados simbólica, permitirá a extração de um conjunto de conceitos formais de bases de dados, conferindo semântica aos resultados e aprimorando a interpretabilidade por parte dos

especialistas do domínio. Essa proposta foi enviada ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e aprovada para um bolsa de pós-doutorado.

No contexto de FCA, a relação entre objetos e atributos pode ser formalizada por uma correspondência binária entre duas classes: uma que representa os objetos e outra, os atributos. Essa correspondência pode ser modelada por meio de uma tabela binária, também chamada contexto binário, onde, para cada par objeto-atributo, deve ser indicado se o atributo aplica-se ou não ao objeto. Tal representação é particularmente interessante pois trata-se da mesma representação utilizada neste trabalho para as matrizes de ocorrência. Assim, é possível gerar uma hierarquia de conceitos, que pode ser representada e visualizada através de estruturas interpretáveis chamadas de *lattice* [Cimiano et al. (2005)]. Essas estruturas podem ser visualizadas através de ferramentas como ToscanaJ [Becker e Correia (2005)], ConExp [Yevtushenko (2003)], Galicia [Valtchev et al. (2003)] e Colibri [Lindig e Götzmann (2007)], o que apoiará o especialista do domínio a identificar padrões interessantes e não óbvios nos dados. Em tarefas subsequentes de Mineração de Dados, essa informação pode ser aplicada, viabilizando a descoberta de conhecimento direcionada pelo domínio prévio ou *Knowledge Discovery Guided by Domain Knowledge* (KDDK) [Lieber et al. (2006)].

Finalmente, queremos investigar como prever mudanças de anotação considerando os níveis do EC gradualmente do 1 ao 4 (da esquerda para direita) para avaliar se acertamos mais ao tentar prever níveis mais altos da hierarquia. Se ao aprofundar na hierarquia EC, tentando prever níveis mais específicos, as previsões piorarem, pode-se optar por prever níveis mais altos, mantendo uma previsão mais geral, porém mais confiável, evitando assim *overprediction*.

Apêndice A

Informações adicionais

A.1 Experimento Descritivo Multiclasse

Esse experimento foi realizado com três configurações diferentes no que diz respeito às tarefas de preprocessamento *n-grams* e *stemmer*. (1) sem utilizar *n-grams* e *stemmer*; (2) sem *n-grams* e com *stemmer*; (3) com *n-grams* e com *stemmer*. O objetivo de usar essas três configurações foi de verificar qual delas era capaz de gerar o melhor modelo de classificação para utilizar tal configuração nos experimentos previsivos subsequentes.

A configuração com *n-grams* e sem *stemmer* não foi executada devido a restrições de *hardware*. Como a matriz de ocorrência (detalhada na Seção 4.3.1) para essa configuração era a maior delas (3.8 GB), a máquina utilizada para executar a SVD (no software R) excedeu a memória RAM. Essa matriz é maior que as demais porque a técnica *stemmer*, que reduziria o número de atributos mapeando as palavras derivadas para sua raiz, não foi aplicada.

Os resultados são apresentados nas Tabelas A.1 (sem *n-grams* e sem *stemmer*), A.2 (sem *n-grams* e com *stemmer*) e A.3 (com *n-grams* e com *stemmer*). A Tabela A.4 resume os resultados. A configuração (3), na qual foram utilizados *n-grams* e *stemmer*, é ligeiramente superior às demais, assim, para os experimentos previsivos, essa foi a configuração adotada.

A.2 Experimento Previsivo Multiclasse

Na Tabela A.5 é apresentado o resultado completo do experimento Previsivo Multiclasse. São exibidos, para cada algoritmo de classificação, os melhores resultados, escolhidos de acordo com diferentes métricas (TPR, FPR, precisão, revocação, F_1 e AUC).

Tabela A.1: Resultados da configuração 1: matriz de ocorrência gerada sem utilizar n-grams e stemmer.

Técnica	Votos	Máximo	Atributos	TPR	FPR	Prec.	Revoc.	F_1	AUC
Naïve Bayes	1	TPR	93	0,494	0,004	0,672	0,494	0,526	0,927
	0	FPR	1	0,255	0,255	0,065	0,255	0,104	0,715
	1	Prec.	93	0,494	0,004	0,672	0,494	0,526	0,927
	1	Rec.	93	0,494	0,004	0,672	0,494	0,526	0,927
	1	F_1	93	0,494	0,004	0,672	0,494	0,526	0,927
	1	**AUC**	**82**	**0,481**	**0,004**	**0,662**	**0,481**	**0,511**	**0,928**
KNN_K1	**2**	**TPR**	**99**	**0,741**	**0,005**	**0,74**	**0,741**	**0,738**	**0,952**
	0	FPR	1	0,559	0,008	0,545	0,559	0,55	0,901
	2	**Prec.**	**99**	**0,741**	**0,005**	**0,74**	**0,741**	**0,738**	**0,952**
	2	**Rec.**	**99**	**0,741**	**0,005**	**0,74**	**0,741**	**0,738**	**0,952**
	2	F_1	**99**	**0,741**	**0,005**	**0,74**	**0,741**	**0,738**	**0,952**
	1	AUC	94	0,74	0,005	0,739	0,74	0,737	0,952
KNN_K3	**2**	**TPR**	**90**	**0,713**	**0,009**	**0,705**	**0,713**	**0,703**	**0,963**
	0	FPR	1	0,487	0,017	0,458	0,487	0,466	0,887
	1	Prec.	97	0,712	0,009	0,705	0,712	0,702	0,963
	2	**Rec.**	**90**	**0,713**	**0,009**	**0,705**	**0,713**	**0,703**	**0,963**
	2	F_1	**90**	**0,713**	**0,009**	**0,705**	**0,713**	**0,703**	**0,963**
	1	AUC	97	0,712	0,009	0,705	0,712	0,702	0,963
KNN_K5	0	TPR	100	0,701	0,013	0,684	0,701	0,683	0,965
	0	FPR	1	0,46	0,024	0,41	0,46	0,428	0,879
	0	Prec.	100	0,701	0,013	0,684	0,701	0,683	0,965
	0	Rec.	100	0,701	0,013	0,684	0,701	0,683	0,965
	2	F_1	**95**	**0,701**	**0,013**	**0,683**	**0,701**	**0,684**	**0,966**
	2	**AUC**	**95**	**0,701**	**0,013**	**0,683**	**0,701**	**0,684**	**0,966**
KNN_K7	1	TPR	48	0,691	0,015	0,667	0,691	0,669	0,966
		FPR	1	0,44	0,031	0,376	0,44	0,4	0,873
	1	Prec.	64	0,691	0,016	0,669	0,691	0,669	0,966
	1	Rec.	48	0,691	0,015	0,667	0,691	0,669	0,966
	2	F_1	**55**	**0,691**	**0,016**	**0,667**	**0,691**	**0,67**	**0,966**
	1	AUC	79	0,689	0,016	0,666	0,689	0,667	0,966
KNN_K10	1	TPR	54	0,676	0,02	0,644	0,676	0,648	0,967
		FPR	1	0,419	0,04	0,341	0,419	0,369	0,866
		Prec.	86	0,676	0,022	0,647	0,676	0,647	0,966
	1	Rec.	54	0,676	0,02	0,644	0,676	0,648	0,967
	2	F_1	**21**	**0,676**	**0,018**	**0,64**	**0,676**	**0,649**	**0,967**
	1	AUC	46	0,675	0,02	0,642	0,675	0,647	0,967
J48	**2**	**TPR**	**88**	**0,744**	**0,006**	**0,732**	**0,744**	**0,733**	**0,937**
		FPR	1	0,498	0,014	0,468	0,498	0,479	0,831
	1	Prec.	90	0,743	0,006	0,732	0,743	0,732	0,937
	2	**Rec.**	**88**	**0,744**	**0,006**	**0,732**	**0,744**	**0,733**	**0,937**
	2	F_1	**88**	**0,744**	**0,006**	**0,732**	**0,744**	**0,733**	**0,937**
	1	AUC	85	0,743	0,006	0,731	0,743	0,732	0,937

Tabela A.2: Resultados da configuração 2: matriz de ocorrência gerada sem utilizar n-grams e com stemmer.

Técnica	Votos	Máximo	Atributos	TPR	FPR	Prec.	Revoc.	F_1	AUC
Naïve Bayes	0	TPR	99	0,492	0,004	0,67	0,492	0,523	0,927
	0	FPR	1	0,255	0,255	0,065	0,255	0,104	0,715
	0	Prec.	94	0,491	0,004	0,671	0,491	0,523	0,927
	1	Rec.	99	0,492	0,004	0,67	0,492	0,523	0,927
	1	F_1	100	0,492	0,004	0,671	0,492	0,524	0,926
	0	**AUC**	**89**	**0,491**	**0,004**	**0,671**	**0,491**	**0,523**	**0,928**
KNN_K1	1	TPR	97	0,741	0,005	0,739	0,741	0,737	0,952
	0	FPR	1	0,559	0,008	0,546	0,559	0,551	0,901
	1	Prec.	92	0,74	0,005	0,739	0,74	0,737	0,952
	1	Rec.	97	0,741	0,005	0,739	0,741	0,737	0,952
	2	F_1	**98**	**0,741**	**0,005**	**0,739**	**0,741**	**0,738**	**0,952**
	1	AUC	82	0,739	0,005	0,738	0,739	0,736	0,952
KNN_K3	**2**	**TPR**	**90**	**0,713**	**0,009**	**0,706**	**0,713**	**0,703**	**0,963**
	0	FPR	1	0,486	0,016	0,457	0,486	0,465	0,887
	2	**Prec.**	**90**	**0,713**	**0,009**	**0,706**	**0,713**	**0,703**	**0,963**
	2	**Rec.**	**90**	**0,713**	**0,009**	**0,706**	**0,713**	**0,703**	**0,963**
	2	F_1	**90**	**0,713**	**0,009**	**0,706**	**0,713**	**0,703**	**0,963**
	1	AUC	84	0,712	0,009	0,705	0,712	0,702	0,963
KNN_K5	1	TPR	91	0,701	0,013	0,683	0,701	0,683	0,966
	0	FPR	1	0,46	0,023	0,411	0,46	0,429	0,879
	1	Prec.	95	0,701	0,013	0,684	0,701	0,684	0,966
	1	Rec.	91	0,701	0,013	0,683	0,701	0,683	0,966
	1	F_1	**48**	**0,701**	**0,012**	**0,683**	**0,701**	**0,685**	**0,965**
	1	AUC	91	0,701	0,013	0,683	0,701	0,683	0,966
KNN_K7	1	TPR	53	0,691	0,016	0,666	0,691	0,669	0,966
	0	FPR	1	0,441	0,03	0,378	0,441	0,401	0,874
	1	Prec.	100	0,691	0,017	0,668	0,691	0,668	0,966
	1	Rec.	53	0,691	0,016	0,666	0,691	0,669	0,966
	2	F_1	**55**	**0,691**	**0,016**	**0,667**	**0,691**	**0,67**	**0,966**
	1	AUC	53	0,691	0,016	0,666	0,691	0,669	0,966
KNN_K10	0	TPR	86	0,677	0,022	0,646	0,677	0,647	0,966
	0	FPR	1	0,419	0,04	0,341	0,419	0,369	0,866
	0	Prec.	85	0,676	0,022	0,647	0,676	0,647	0,966
	0	Rec.	86	0,677	0,022	0,646	0,677	0,647	0,966
	1	F_1	**22**	**0,676**	**0,018**	**0,64**	**0,676**	**0,649**	**0,967**
	1	AUC	48	0,676	0,02	0,643	0,676	0,648	0,968
J48	1	TPR	90	0,742	0,006	0,731	0,742	0,731	0,936
	0	FPR	1	0,498	0,013	0,469	0,498	0,479	0,831
	1	Prec.	90	0,742	0,006	0,731	0,742	0,731	0,936
	1	Rec.	90	0,742	0,006	0,731	0,742	0,731	0,936
	1	F_1	90	0,742	0,006	0,731	0,742	0,731	0,936
	1	**AUC**	**61**	**0,741**	**0,006**	**0,729**	**0,741**	**0,729**	**0,937**

Tabela A.3: Resultados da configuração 3: matriz de ocorrência gerada utilizando n-grams e stemmer.

Técnica	Votos	Máximo	Atributos	TPR	FPR	Prec.	Revoc.	F_1	AUC
Naïve Bayes	**2**	**TPR**	**97**	**0,507**	**0,005**	**0,672**	**0,507**	**0,534**	**0,929**
	0	FPR	1	0,255	0,255	0,065	0,255	0,104	0,715
	1	Prec.	100	0,505	0,005	0,672	0,505	0,532	0,929
	2	**Rec.**	**97**	**0,507**	**0,005**	**0,672**	**0,507**	**0,534**	**0,929**
	2	F_1	**97**	**0,507**	**0,005**	**0,672**	**0,507**	**0,534**	**0,929**
	1	AUC	90	0,499	0,004	0,667	0,499	0,525	0,929
KNN_K1	1	TPR	95	0,744	0,005	0,741	0,744	0,74	0,952
	0	FPR	1	0,567	0,008	0,554	0,567	0,559	0,903
	1	Prec.	97	0,744	0,005	0,742	0,744	0,74	0,952
	1	Rec.	95	0,744	0,005	0,741	0,744	0,74	0,952
	1	F_1	95	0,744	0,005	0,741	0,744	0,74	0,952
	1	**AUC**	**38**	**0,741**	**0,005**	**0,739**	**0,741**	**0,738**	**0,953**
KNN_K3	1	TPR	29	0,718	0,009	0,709	0,718	0,709	0,962
	0	FPR	1	0,495	0,016	0,467	0,495	0,475	0,891
	2	**Prec.**	**100**	**0,718**	**0,009**	**0,712**	**0,718**	**0,709**	**0,963**
	1	Rec.	29	0,718	0,009	0,709	0,718	0,709	0,962
	1	F_1	29	0,718	0,009	0,709	0,718	0,709	0,962
	1	AUC	86	0,716	0,009	0,709	0,716	0,707	0,963
KNN_K5	1	TPR	95	0,711	0,013	0,696	0,711	0,695	0,966
	0	FPR	1	0,468	0,024	0,421	0,468	0,438	0,884
	2	**Prec.**	**100**	**0,711**	**0,013**	**0,697**	**0,711**	**0,696**	**0,966**
	1	Rec.	95	0,711	0,013	0,696	0,711	0,695	0,966
	2	F_1	**100**	**0,711**	**0,013**	**0,697**	**0,711**	**0,696**	**0,966**
	1	AUC	95	0,711	0,013	0,696	0,711	0,695	0,966
KNN_K7	**2**	**TPR**	**96**	**0,702**	**0,016**	**0,683**	**0,702**	**0,682**	**0,966**
	0	FPR	1	0,449	0,03	0,387	0,449	0,41	0,88
	2	**Prec.**	**96**	**0,702**	**0,016**	**0,683**	**0,702**	**0,682**	**0,966**
	2	**Rec.**	**96**	**0,702**	**0,016**	**0,683**	**0,702**	**0,682**	**0,966**
	2	F_1	**96**	**0,702**	**0,016**	**0,683**	**0,702**	**0,682**	**0,966**
	1	AUC	83	0,701	0,017	0,68	0,701	0,68	0,966
KNN_K10	**1**	**TPR**	**81**	**0,691**	**0,022**	**0,664**	**0,691**	**0,664**	**0,966**
	0	FPR	1	0,426	0,04	0,35	0,426	0,377	0,873
	1	Prec.	97	0,689	0,022	0,665	0,689	0,663	0,967
	1	**Rec.**	**81**	**0,691**	**0,022**	**0,664**	**0,691**	**0,664**	**0,966**
	1	F_1	**81**	**0,691**	**0,022**	**0,664**	**0,691**	**0,664**	**0,966**
	1	AUC	97	0,689	0,022	0,665	0,689	0,663	0,967
J48	**2**	**TPR**	**88**	**0,738**	**0,006**	**0,728**	**0,738**	**0,727**	**0,934**
	0	FPR	1	0,505	0,014	0,473	0,505	0,484	0,839
	2	Prec.	97	0,738	0,006	0,73	0,738	0,727	0,934
	2	**Rec.**	**88**	**0,738**	**0,006**	**0,728**	**0,738**	**0,727**	**0,934**
	2	F_1	**88**	**0,738**	**0,006**	**0,728**	**0,738**	**0,727**	**0,934**
	2	AUC	88	0,738	0,006	0,728	0,738	0,727	0,934

Tabela A.4: Melhor desempenho do experimento Descritivo Multiclasse para cada algoritmo de classificação separado por configuração, (1) Nem n-grams nem stemmer utilizado; (2) sem n-grams e com stemmer; (3) com n-grams e com stemmer.

Configuração	Votos	Técnica	Atributos	TPR	FPR	Prec.	Rec.	F_1	AUC
1	0	Naïve Bayes	82	0,481	0,004	0,662	0,481	0,511	0,928
	1	**KNN_K1**	**99**	**0,741**	**0,005**	**0,74**	**0,741**	**0,738**	**0,952**
	0	KNN_K3	90	0,713	0,009	0,705	0,713	0,703	0,963
	0	KNN_K5	95	0,701	0,013	0,683	0,701	0,684	0,966
	0	KNN_K7	55	0,691	0,016	0,667	0,691	0,67	0,966
	1	KNN_K10	21	0,676	0,018	0,64	0,676	0,649	0,967
	0	J48	88	0,744	0,006	0,732	0,744	0,733	0,937
2	0	Naïve Bayes	89	0,491	0,004	0,671	0,491	0,523	0,928
	1	**KNN_K1**	**98**	**0,741**	**0,005**	**0,739**	**0,741**	**0,738**	**0,952**
	0	KNN_K3	90	0,713	0,009	0,706	0,713	0,703	0,963
	0	KNN_K5	48	0,701	0,012	0,683	0,701	0,685	0,965
	0	KNN_K7	55	0,691	0,016	0,667	0,691	0,67	0,966
	1	KNN_K10	22	0,676	0,018	0,64	0,676	0,649	0,967
	0	J48	61	0,741	0,006	0,729	0,741	0,729	0,937
3	0	Naïve Bayes	97	0,507	0,005	0,672	0,507	0,534	0,929
	1	**KNN_K1**	**38**	**0,741**	**0,005**	**0,739**	**0,741**	**0,738**	**0,953**
	0	KNN_K3	100	0,718	0,009	0,712	0,718	0,709	0,963
	1	KNN_K5	100	0,711	0,013	0,697	0,711	0,696	0,966
	1	KNN_K7	96	0,702	0,016	0,683	0,702	0,682	0,966
	1	KNN_K10	81	0,691	0,022	0,664	0,691	0,664	0,966
	0	J48	88	0,738	0,006	0,728	0,738	0,727	0,934

Tabela A.5: Experimento Previsivo Multiclasse: a última versão na qual uma determinada mudança ocorreu foi utilizada como teste e as demais versões como dados de treino.

Técnica	Votos	Máximo	Atributos	TPR	FPR	Prec.	Revoc.	F_1	AUC
Naïve Bayes	0	TPR	100	0,201	0,064	0,320	0,201	0,214	0,699
	0	FPR	92	0,176	0,066	0,323	0,176	0,184	0,698
	0	Prec.	53	0,150	0,019	0,387	0,150	0,191	0,685
	0	Rec.	100	0,201	0,064	0,320	0,201	0,214	0,699
	1	F_1	**65**	**0,200**	**0,039**	**0,344**	**0,200**	**0,236**	**0,692**
	1	AUC	74	0,184	0,056	0,328	0,184	0,208	0,704
KNN_K1	0	TPR	34	0,318	0,089	0,338	0,318	0,236	0,646
	0	FPR	22	0,314	0,102	0,387	0,314	0,225	0,639
	0	Prec.	1	0,239	0,013	0,564	0,239	0,243	0,657
	0	Rec.	34	0,318	0,089	0,338	0,318	0,236	0,646
	1	F_1	**13**	**0,316**	**0,075**	**0,406**	**0,316**	**0,247**	**0,652**
	1	AUC	60	0,316	0,085	0,399	0,316	0,240	0,663
KNN_K3	1	TPR	57	0,301	0,084	0,488	0,301	0,242	0,634
	0	FPR	25	0,282	0,105	0,287	0,282	0,204	0,611
	0	Prec.	54	0,298	0,077	0,498	0,298	0,241	0,649
	1	Rec.	57	0,301	0,084	0,488	0,301	0,242	0,634
	1	F_1	57	0,301	0,084	0,488	0,301	0,242	0,634
	1	**AUC**	**12**	**0,283**	**0,066**	**0,399**	**0,283**	**0,232**	**0,657**
KNN_K5	0	TPR	17	0,283	0,088	0,386	0,283	0,213	0,622
	0	FPR	28	0,269	0,112	0,199	0,269	0,194	0,624
	0	Prec.	93	0,267	0,081	0,506	0,267	0,220	0,643
	0	Rec.	17	0,283	0,088	0,386	0,283	0,213	0,622
	1	F_1	**57**	**0,282**	**0,086**	**0,502**	**0,282**	**0,231**	**0,635**
	1	AUC	75	0,269	0,076	0,449	0,269	0,226	0,657
KNN_K7	0	TPR	56	0,272	0,090	0,503	0,272	0,218	0,641
	0	FPR	26	0,259	0,114	0,202	0,259	0,184	0,629
	0	Prec.	91	0,260	0,077	0,510	0,260	0,216	0,650
	0	Rec.	56	0,272	0,090	0,503	0,272	0,218	0,641
	2	F_1	**13**	**0,260**	**0,049**	**0,238**	**0,260**	**0,225**	**0,671**
	2	AUC	13	0,260	0,049	0,238	0,260	0,225	0,671
KNN_K10	**2**	**TPR**	**100**	**0,270**	**0,085**	**0,497**	**0,270**	**0,225**	**0,666**
	0	FPR	26	0,251	0,107	0,201	0,251	0,182	0,637
	0	Prec.	69	0,257	0,079	0,515	0,257	0,212	0,647
	2	Rec.	100	0,270	0,085	0,497	0,270	0,225	0,666
	2	F_1	100	0,270	0,085	0,497	0,270	0,225	0,666
	2	AUC	100	0,270	0,085	0,497	0,270	0,225	0,666
J48	0	TPR	90	0,310	0,079	0,383	0,310	0,254	0,607
	0	FPR	32	0,300	0,115	0,301	0,300	0,219	0,669
	0	Prec.	44	0,302	0,073	0,688	0,302	0,248	0,621
	0	Rec.	90	0,310	0,079	0,383	0,310	0,254	0,607
	1	F_1	46	0,299	0,052	0,418	0,299	0,255	0,638
	1	**AUC**	**16**	**0,296**	**0,084**	**0,249**	**0,296**	**0,221**	**0,692**

A.3 Lista de Mudanças

A Tabela A.6 contém uma lista das mudanças de Ec *number* com pelo menos 10 exemplos ao longo das 44 versões do Swiss-Prot que foram utilizadas nas análises desse trabalho.

Tabela A.6: Lista de mudanças e versões em que ocorreram

-.-.-.-;1.-.-.-;2;3;4;5;6;8;9;10;11;12;13;14;15;16;25;27;30;33;34;35;36;44	2.1.1.61;2.8.1.-;15
-.-.-.-;1.1.-.-;7;9;10;42	2.1.1.72;2.1.1.-;39
-.-.-.-;1.1.1.-;2;3;7;9;11;15;16;23;25	2.1.1.73;2.1.1.37;2
-.-.-.-;1.1.1.63;13	2.1.2.-;2.1.2.13;26
-.-.-.-;1.10.3.9;36	2.1.2.1;2.1.2.-;43
-.-.-.-;1.11.1.15;5;15	2.3.-.-;2.3.1.-;14;15;16
-.-.-.-;1.11.1.7;14	2.3.1.-;-.-.-.-;13;14;17;42
-.-.-.-;1.13.-.-;3;7;13;14;18	2.3.1.-;2.3.1.178;13
-.-.-.-;1.13.11.24;5;28;29	2.3.1.-;2.3.1.181;13
-.-.-.-;1.14.11.-;8;9;10;12;13;14;15;16;18;21;22;29;34	2.3.1.-;2.3.1.191;39
-.-.-.-;1.14.11.27;16;25	2.3.1.-;2.3.1.193;33
-.-.-.-;1.14.13.-;3;42;43;44	2.3.1.-;2.3.1.88;9;16;37
-.-.-.-;1.16.3.1;5;25;39	2.3.1.29;-.-.-.-;35
-.-.-.-;1.2.1.-;2;16	2.3.1.41;2.3.1.179;14
-.-.-.-;1.3.1.-;3;7;9;13;15;16;25;43	2.3.1.41;2.3.1.180;14
-.-.-.-;1.3.3.11;6	2.3.1.48;-.-.-.-;33;36
-.-.-.-;1.3.99.-;5;10;16;18;28	2.3.1.51;2.3.1.15;16
-.-.-.-;1.5.-.-;15	2.3.2.-;2.3.2.16;39
-.-.-.-;1.5.3.-;15	2.3.2.-;2.3.2.17;39
-.-.-.-;1.8.-.-;18;31	2.3.2.-;2.3.2.18;39
-.-.-.-;1.8.4.11;10	2.3.3.13;2.3.3.14;31
-.-.-.-;1.97.1.12;38	2.4.-.-;2.4.1.-;23;30;31
-.-.-.-;2.-.-.-;2;3;4;5;6;9;11;12;14;15;16;18;19;21;33;43	2.4.1.-;-.-.-.-;2;9
-.-.-.-;2.1.1.-;2;3;4;6;7;8;9;10;11;13;14;15;16;17;22;25;27;28;30;31;32;33;35;39;40;41;43	2.4.1.-;2.4.-.-;14
-.-.-.-;2.1.1.163;33	2.4.1.-;2.4.1.131;35
-.-.-.-;2.1.1.43;4;5;6;9;11;13;14;15;16;32	2.4.1.-;2.4.1.141;9
-.-.-.-;2.1.1.56;5;7;10;13;14;16;29;37;44	2.4.1.-;2.4.1.256;35
-.-.-.-;2.1.1.57;16;29	2.4.1.-;2.4.1.257;35
-.-.-.-;2.1.1.74;15	2.4.1.-;2.4.1.265;37
-.-.-.-;2.3.1.-;2;3;4;5;6;7;8;9;10;11;14;15;16;17;28;29;34;35;39;43	2.4.1.-;2.4.1.267;37
-.-.-.-;2.3.1.184;13;16	2.4.1.11;2.4.1.21;2
-.-.-.-;2.3.1.20;2;20	2.4.1.130;2.4.1.258;35
-.-.-.-;2.3.1.47;16;35	2.4.1.131;2.4.1.-;35
-.-.-.-;2.3.1.48;2;6;8;9;12;14;15;43;44	2.4.1.21;2.4.1.242;6
-.-.-.-;2.3.2.-;8;22	2.4.1.91;2.4.1.115;7
-.-.-.-;2.3.2.3;16;21	2.4.2.-;2.4.2.43;27
-.-.-.-;2.3.2.4;15;22	2.5.1.-;2.2.1.9;15
-.-.-.-;2.4.1.-;2;6;7;9;13;16;29;34	2.5.1.-;2.5.1.32;13
-.-.-.-;2.4.1.119;9	2.5.1.-;2.5.1.39;29
-.-.-.-;2.4.1.132;35	2.5.1.-;2.5.1.73;18
-.-.-.-;2.4.2.-;6;9;10;39	2.5.1.-;2.5.1.77;30
-.-.-.-;2.5.1.-;2;5;6;7;9;11;15;28;32;44	2.5.1.-;2.5.1.96;40
-.-.-.-;2.5.1.41;15	2.5.1.10;-.-.-.-;29
-.-.-.-;2.5.1.72;15	2.5.1.1;-.-.-.-;29
-.-.-.-;2.6.99.2;10	2.5.1.29;-.-.-.-;29
-.-.-.-;2.7.-.-;3;4;6;8;13;16;19	2.5.1.31;2.5.1.89;39
-.-.-.-;2.7.1.-;2;3;5;8;9;11;12;13;14;15;16;23;28;29;36	2.5.1.8;2.5.1.75;18
-.-.-.-;2.7.1.37;2;3;4;5	2.6.1.-;2.6.1.87;26
-.-.-.-;2.7.1.49;18;40	2.6.1.11;2.6.1.13;16

-.-.-.-;2.7.11.-;8;13;14;29;34	2.6.1.22;-.-.-.-;14
-.-.-.-;2.7.11.1;8;10;11;12;13;16;27;32;33;38	2.6.1.44;-.-.-.-;14
-.-.-.-;2.7.11.22;8;11;13;29	2.7.-.-;2.7.1.-;5;10
-.-.-.-;2.7.11.23;12;13;31	2.7.-.-;2.7.11.-;13
-.-.-.-;2.7.11.26;36;37	2.7.1.-;2.7.1.158;10
-.-.-.-;2.7.4.3;5;15	2.7.1.-;2.7.1.159;10
-.-.-.-;2.7.7.-;5;7;8;11;13;14;15;16;18;33;35;42	2.7.1.-;2.7.1.161;14
-.-.-.-;2.7.7.21;6	2.7.1.-;2.7.1.170;36
-.-.-.-;2.7.7.48;2;3;4;5;7;8;10;11;14;23;24;29;33;40	2.7.1.-;2.7.1.37;2;3;4;5;7
-.-.-.-;2.7.7.49;3;9;10;14;24;40	2.7.1.-;2.7.11.-;8
-.-.-.-;2.7.7.7;2;6;7;9;10;13;14;16;24;28;40	2.7.1.-;2.7.11.1;8
-.-.-.-;2.7.7.77;44	2.7.1.-;2.7.11.13;8
-.-.-.-;2.8.1.-;2;8;15	2.7.1.-;2.7.11.16;8
-.-.-.-;2.8.1.4;18	2.7.1.-;2.7.12.2;8
-.-.-.-;2.9.1.-;7;11	2.7.1.-;2.7.8.-;9
-.-.-.-;3.-.-.-;3;4;5;7;8;9;11;12;13;14;15;16;30	2.7.1.112;2.7.10.1;8
-.-.-.-;3.1.-.-;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;19;23; -.-.-.-;3.1.-.-;25;27;28;29;30;31;32;33;34;36;41;43;44	2.7.1.112;2.7.10.2;8
-.-.-.-;3.1.1.-;8;9;11;12;13;14;16;29;31;34;41	2.7.1.112;2.7.12.1;8
-.-.-.-;3.1.1.29;2;21;35;42	2.7.1.116;2.7.11.5;8
-.-.-.-;3.1.1.4;2;14;16	2.7.1.117;2.7.11.18;8
-.-.-.-;3.1.1.53;7	2.7.1.123;2.7.11.17;8
-.-.-.-;3.1.11.-;2;7;31;35;37	2.7.1.159;-.-.-.-;14
-.-.-.-;3.1.13.-;11;15;32	2.7.1.37;-.-.-.-;4;8
-.-.-.-;3.1.2.-;6;8;11;12;15;16	2.7.1.37;2.7.11.1;8
-.-.-.-;3.1.2.15;15;16;24	2.7.1.37;2.7.11.11;8
-.-.-.-;3.1.21.-;6;8;13;14;16;25;27	2.7.1.37;2.7.11.12;8
-.-.-.-;3.1.22.-;2;13	2.7.1.37;2.7.11.13;8
-.-.-.-;3.1.26.-;3;9;11;16;28;32;42	2.7.1.37;2.7.11.21;8
-.-.-.-;3.1.26.4;3;9;10;14;16;24;28;40	2.7.1.37;2.7.11.22;8
-.-.-.-;3.1.26.5;2	2.7.1.37;2.7.11.23;8
-.-.-.-;3.1.27.-;8;9;11;13	2.7.1.37;2.7.11.24;8
-.-.-.-;3.1.3.-;3;5;6;8;9;10;11;12;13;14;15;16;18;25;29;34;37;38;42;43	2.7.1.37;2.7.11.25;8
-.-.-.-;3.1.3.16;2;3;5;6;8;9;10;13;16	2.7.1.37;2.7.11.30;8
-.-.-.-;3.1.3.33;8;44	2.7.1.37;2.7.12.2;8
-.-.-.-;3.1.3.48;2;3;6;7;8;9;13;14;15;16;17;30	2.7.1.66;3.6.1.27;3
-.-.-.-;3.1.3.5;6;13	2.7.1.68;2.7.1.149;13
-.-.-.-;3.1.3.7;9;15;37	2.7.1.69;2.7.1.-;5
-.-.-.-;3.1.4.-;3;6;13;16;43	2.7.1.99;2.7.11.2;8
-.-.-.-;3.1.4.16;14	2.7.10.2;2.7.10.-;14
-.-.-.-;3.2.-.-;7;8;10;14;16	2.7.11.-;2.7.11.1;15;16;17;28
-.-.-.-;3.2.1.-;2;5;6;9;10;11;12;13;15;16;25	2.7.11.1;-.-.-.-;10;13;19;28;32;34
-.-.-.-;3.2.2.-;9;16;20;22	2.7.11.1;2.7.12.1;16
-.-.-.-;3.4.-.-;2;5;6;7;8;9;10;12;15;16;23;25;27;30;38;42;44	2.7.3.-;2.7.13.3;8;15
-.-.-.-;3.4.19.12;4;24;31	2.7.4.-;2.7.4.22;14
-.-.-.-;3.4.21.-;2;3;5;7;8;9;11;13;16;22;25;27;34;43	2.7.4.14;2.7.4.25;39
-.-.-.-;3.4.22.-;2;3;4;5;6;7;8;11;14;16;24;31;33;34;36	2.7.7.-;2.7.7.49;5
-.-.-.-;3.4.22.29;2;3;11	2.7.7.-;2.7.7.66;15
-.-.-.-;3.4.23.-;7;8;9;10;12;13;24;28;40	2.7.7.-;2.7.7.75;37
-.-.-.-;3.4.24.-;3;8;9;11;13;15	2.7.7.-;2.7.7.79;41
-.-.-.-;3.5.-.-;3;9;12	2.7.7.-;2.7.7.80;41
-.-.-.-;3.5.1.-;7;9;16;27;37;44	2.7.7.21;-.-.-.-;30
-.-.-.-;3.5.1.98;14	2.7.7.22;2.7.7.13;13;15
-.-.-.-;3.5.2.17;13;16	2.7.7.25;2.7.7.72;30
-.-.-.-;3.5.4.-;2;4;7;9;13;16;24;42	2.7.7.48;-.-.-.-;3;4;7;14
-.-.-.-;3.5.4.16;16	2.7.7.50;-.-.-.-;14
-.-.-.-;3.6.-.-;4;10;15;23	2.7.7.6;-.-.-.-;11;12;15

-.-.-.-;3.6.1.-;2;3;4;5;6;7;8;10;12;13;14;15;16;20;40	2.7.8.-;2.7.8.28;30
-.-.-.-;3.6.1.15;4;7;8;9;13;14;15;16;29	2.7.8.-;2.7.8.30;27
-.-.-.-;3.6.1.19;37;39	2.7.8.-;2.7.8.33;35
-.-.-.-;3.6.1.3;13;15;43	2.8.1.-;2.8.1.11;41
-.-.-.-;3.6.1.55;41	2.8.1.-;2.8.1.8;13
-.-.-.-;3.6.3.-;5;7;10;11;13;14;15;16;19;27;43	2.9.1.-;2.9.1.2;18
-.-.-.-;3.6.3.14;5;7;12;14	3.-.-.-;3.1.-.-;7
-.-.-.-;3.6.3.17;5;10	3.-.-.-;3.5.1.-;23;27
-.-.-.-;3.6.3.44;8;14	3.1.-.-;-.-.-.-;14;16;18;34
-.-.-.-;3.6.4.-;23;25;43	3.1.-.-;3.1.13.1;34
-.-.-.-;3.6.4.12;23;31;37	3.1.-.-;3.1.3.-;40
-.-.-.-;3.6.4.13;23;29;33	3.1.-.-;3.5.1.96;10
-.-.-.-;3.6.5.-;2;7;9;12;13;15;16;28	3.1.1.-;3.1.1.89;41
-.-.-.-;4.-.-.-;6;12;15;16;18;28;38	3.1.1.1;3.1.1.85;36
-.-.-.-;4.1.1.-;2;3;6;13;16;29;42	3.1.1.21;-.-.-.-;40
-.-.-.-;4.1.99.12;13	3.1.11.-;-.-.-.-;25
-.-.-.-;4.2.-.-;9	3.1.11.-;3.1.-.-;7;36
-.-.-.-;4.2.1.-;4;6;15;16;24;42	3.1.13.-;-.-.-.-;15;32
-.-.-.-;4.2.1.109;15;20;27	3.1.2.-;3.1.2.28;29
-.-.-.-;4.2.1.130;41;42	3.1.2.-;3.4.19.12;24
-.-.-.-;4.2.3.12;2	3.1.2.15;3.4.19.-;24
-.-.-.-;4.2.99.-;4;23	3.1.2.15;3.4.19.12;24;25
-.-.-.-;4.2.99.18;4;16;20	3.1.26.-;3.1.26.4;4
-.-.-.-;5.-.-.-;3;9;10;16	3.1.26.11;3.1.-.-;33
-.-.-.-;5.1.-.-;9;14;16	3.1.26.4;3.1.26.13;18
-.-.-.-;5.1.3.-;14;31	3.1.3.-;2.7.4.-;2
-.-.-.-;5.1.99.-;42	3.1.3.-;3.1.3.16;13;16;42
-.-.-.-;5.2.1.8;2;7;8;11;13;19;40	3.1.3.-;3.1.3.78;15
-.-.-.-;5.3.1.-;10;15;26	3.1.3.-;3.1.3.84;36
-.-.-.-;5.3.1.23;2;9;15;16;27	3.1.3.2;-.-.-.-;14;16
-.-.-.-;5.4.99.-;6;8;16;32	3.1.3.2;3.1.3.5;6
-.-.-.-;5.4.99.28;33	3.1.3.48;-.-.-.-;14
-.-.-.-;6.-.-.-;6;7;9;14;15	3.1.4.-;3.1.26.12;15
-.-.-.-;6.3.2.-;2;3;4;5;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21; -.-.-.-;6.3.2.-;23;24;25;27;29;30;31;32;33;34;36;37;39;41;42;43;44	3.1.4.14;1.7.-.-;8;9
-.-.-.-;6.3.4.-;5;13;32	3.1.4.16;3.1.-.-;33
-.-.-.-;6.5.1.-;31	3.1.4.17;3.1.4.35;6;15;29
-.-.-.-;6.5.1.1;16	3.1.4.41;-.-.-.-;18
-.-.-.-;6.5.1.3;11;31	3.13.1.-;4.4.1.21;5
-.-.-.-;6.6.1.1;2	3.2.1.-;3.2.1.171;41
1.-.-.-;1.1.1.-;3;4;15;16;22;25	3.2.1.-;4.2.2.-;15
1.-.-.-;1.1.1.298;27	3.2.1.18;-.-.-.-;6;15
1.-.-.-;1.14.99.44;36	3.2.2.-;-.-.-.-;14;28
1.-.-.-;1.2.7.7;2	3.2.2.-;3.2.2.27;18
1.-.-.-;1.2.7.8;2;6	3.2.2.-;3.2.2.28;18
1.-.-.-;1.3.1.-;40	3.2.2.16;-.-.-.-;4
1.-.-.-;1.3.1.89;40	3.4.-.-;3.4.21.89;35
1.-.-.-;1.3.99.22;5	3.4.11.14;3.4.11.4;2
1.1.1.-;1.1.1.290;9;12	3.4.11.5;-.-.-.-;14
1.1.1.-;1.1.1.300;18;26	3.4.21.-;-.-.-.-;7;14;15;16;23;38;43
1.1.1.-;1.1.1.305;26	3.4.21.-;3.4.21.105;7
1.1.1.-;1.1.1.307;41	3.4.21.-;3.4.21.107;38;39
1.1.1.179;-.-.-.-;14	3.4.21.-;3.4.21.90;22
1.1.1.204;1.17.1.4;5	3.4.21.98;3.4.21.91;2
1.1.1.284;-.-.-.-;14	3.4.22.-;-.-.-.-;3;4;14
1.1.1.5;1.1.1.304;24	3.4.22.-;3.4.22.56;11
1.1.1.63;-.-.-.-;14	3.4.22.-;3.4.22.66;13

1.1.99.16;1.1.5.4;18	3.4.22.-;3.4.22.68;14;41;43
1.1.99.5;1.1.5.3;15	3.4.22.-;3.4.22.69;21;24
1.10.2.2;-.-.-.-;9;13;14	3.4.22.29;3.6.1.15;13
1.10.99.1;1.10.9.1;41	3.4.23.-;-.-.-.-;7;9;14;26
1.11.1.-;1.11.1.15;5	3.4.23.-;3.4.23.16;13
1.11.1.-;1.11.1.20;33	3.4.23.-;3.4.23.50;21
1.11.1.6;-.-.-.-;36	3.4.24.-;-.-.-.-;16;26;29;32
1.11.1.7;1.11.1.21;36	3.4.24.-;3.4.22.-;4
1.11.1.9;1.11.1.12;2;7	3.4.24.-;3.4.24.40;38
1.13.-.-;1.13.11.53;31	3.4.24.57;-.-.-.-;32
1.13.-.-;1.13.11.54;39	3.4.24.57;2.7.11.1;32
1.13.11.12;1.13.11.58;36	3.4.25.-;3.4.25.2;23;31
1.13.11.32;1.13.12.16;18	3.4.99.-;3.4.21.-;11;13
1.13.11.53;1.13.11.54;39	3.4.99.-;3.4.24.-;7;9
1.14.-.-;1.14.13.-;14;15;30;36;41	3.5.-.-;6.3.-.-;18
1.14.11.-;1.14.11.27;10	3.5.1.-;3.5.1.97;13
1.14.12.17;-.-.-.-;14	3.5.1.1;3.4.19.5;14
1.14.13.-;1.14.13.127;36	3.5.4.-;3.5.4.31;37
1.14.14.1;1.14.13.-;16	3.5.4.4;3.5.4.2;33
1.14.15.3;-.-.-.-;16	3.6.1.-;-.-.-.-;3;9;12;14;16;21
1.14.15.4;-.-.-.-;16	3.6.1.-;3.6.1.22;6
1.14.99.-;1.3.8.2;40	3.6.1.-;3.6.1.54;39
1.14.99.7;1.14.13.132;40	3.6.1.-;3.6.4.-;23
1.17.4.3;1.17.7.1;16	3.6.1.-;3.6.4.12;23
1.18.-.-;1.8.7.2;27	3.6.1.-;3.6.4.13;23
1.2.1.-;1.2.1.70;5	3.6.1.11;-.-.-.-;35
1.2.1.16;1.2.1.79;29	3.6.1.15;3.6.1.19;34
1.2.1.1;1.1.1.284;6	3.6.1.15;3.6.4.13;36
1.2.1.3;1.2.1.36;5	3.6.1.19;3.6.1.-;40
1.20.4.-;-.-.-.-;14	3.6.1.3;-.-.-.-;14
1.3.1.-;1.3.1.87;36	3.6.1.3;3.6.4.3;16
1.3.3.1;1.3.1.14;33	3.6.1.50;3.6.5.5;2
1.3.3.1;1.3.5.2;16	3.6.3.14;-.-.-.-;13;15;16;40
1.3.3.1;1.3.98.-;33	3.6.3.15;-.-.-.-;15
1.3.98.-;1.3.98.1;36	3.6.3.16;3.6.-.-;16
1.4.1.-;1.4.1.21;7	3.6.3.17;-.-.-.-;10
1.4.3.6;1.4.3.21;15	3.6.4.13;-.-.-.-;34
1.4.98.1;1.4.9.1;40	3.7.1.-;3.7.1.14;38
1.4.99.3;1.4.98.1;36	3.8.1.4;1.97.1.10;2
1.5.1.29;1.5.1.42;38	4.-.-.-;2.8.1.10;40
1.5.1.35;1.2.1.19;13	4.-.-.-;4.1.99.17;40
1.5.3.-;-.-.-.-;17	4.1.1.-;1.1.1.-;11
1.6.4.-;1.8.1.-;17	4.1.1.21;5.4.99.18;27
1.6.5.3;-.-.-.-;13;16	4.1.1.21;6.3.4.18;27
1.6.8.-;1.5.1.-;8;17	4.1.2.-;4.1.2.43;15
1.6.99.3;-.-.-.-;13	4.1.3.-;4.1.3.40;12
1.7.1.-;1.7.1.13;13	4.1.3.-;6.3.5.8;2
1.7.99.6;1.7.2.4;35	4.2.-.-;4.2.1.-;16
1.8.4.6;1.8.4.11;10	4.2.1.-;4.2.1.108;13
1.8.4.6;1.8.4.12;10	4.2.1.-;4.2.1.113;13
1.9.3.1;-.-.-.-;13;16	4.2.1.-;4.2.1.126;36
2.-.-.-;2.10.1.1;37	4.2.1.52;4.-.-.-;38
2.-.-.-;2.4.-.-;13;14	4.2.1.70;5.4.99.-;5
2.-.-.-;2.4.2.-;16;25	4.2.1.70;5.4.99.12;5
2.-.-.-;2.8.1.12;41	4.2.2.-;4.2.2.23;37;41
2.1.-.-;2.1.1.-;14	4.2.3.12;4.-.-.-;18
2.1.1.-;-.-.-.-;14;15;31;32	4.2.99.-;4.2.99.20;15

2.1.1.-;2.1.1.163;33	4.2.99.18;-.-.-.-;14
2.1.1.-;2.1.1.166;27	4.3.1.5;4.3.1.24;14
2.1.1.-;2.1.1.170;27;44	4.4.-.-;2.8.1.9;40
2.1.1.-;2.1.1.176;29	4.4.1.-;2.8.1.7;2
2.1.1.-;2.1.1.177;39	4.4.1.16;-.-.-.-;14
2.1.1.-;2.1.1.178;29	5.-.-.-;5.3.1.-;3;15
2.1.1.-;2.1.1.182;28	5.1.3.-;5.1.3.24;38
2.1.1.-;2.1.1.183;28	5.2.1.8;-.-.-.-;14
2.1.1.-;2.1.1.185;39	5.3.1.-;5.3.1.28;27
2.1.1.-;2.1.1.186;39	5.3.1.16;-.-.-.-;14
2.1.1.-;2.1.1.189;29;40	5.3.1.24;-.-.-.-;14
2.1.1.-;2.1.1.190;29;40	5.4.99.-;5.4.99.19;33
2.1.1.-;2.1.1.191;29	5.4.99.-;5.4.99.20;33
2.1.1.-;2.1.1.192;29	5.4.99.-;5.4.99.22;33
2.1.1.-;2.1.1.193;29	5.4.99.-;5.4.99.23;33
2.1.1.-;2.1.1.194;29	5.4.99.-;5.4.99.24;33
2.1.1.-;2.1.1.198;30	5.4.99.-;5.4.99.25;33
2.1.1.-;2.1.1.199;30	5.4.99.-;5.4.99.26;33
2.1.1.-;2.1.1.200;33	5.4.99.-;5.4.99.27;33
2.1.1.-;2.1.1.201;33	5.4.99.-;5.4.99.29;33
2.1.1.-;2.1.1.206;33	5.4.99.6;5.4.4.2;2
2.1.1.-;2.1.1.207;36	5.5.1.-;5.5.1.19;43
2.1.1.-;2.1.1.211;38	6.-.-.-;2.3.1.-;2;13
2.1.1.-;2.1.1.233;41	6.-.-.-;6.3.2.-;2;4;10;15;16;24;44
2.1.1.-;2.1.1.61;29	6.1.1.-;6.1.1.27;22
2.1.1.194;2.1.1.224;39	6.1.1.16;6.3.1.13;25
2.1.1.31;2.1.1.221;39	6.1.1.6;-.-.-.-;21
2.1.1.31;2.1.1.228;39	6.3.2.-;2.7.7.63;13
2.1.1.32;2.1.1.216;39	6.3.2.-;6.3.2.19;2
2.1.1.36;2.1.1.220;39	6.3.2.13;6.3.2.-;13
2.1.1.48;2.1.1.181;28	6.3.2.13;6.3.2.7;13
2.1.1.48;2.1.1.184;28	6.3.2.19;6.3.2.-;3;7;10;24
2.1.1.52;2.1.1.172;27	6.3.4.-;6.3.4.19;37
2.1.1.52;2.1.1.173;27	6.3.5.1;6.3.1.5;2
2.1.1.52;2.1.1.174;27	6.3.5.8;2.6.1.85;13
2.1.1.55;2.1.1.223;40	6.4.1.2;-.-.-.-;14;16

Apêndice B

Artigo Publicado

Numa etapa exploratória inicial do presente trabalho, foi proposta uma modelagem para as mudanças de EC *number* em função dos parâmetros prefixo comum, especializações e generalizações (detalhada na Seção 4.2.1). Essa modelagem deu origem a uma ferramenta de visualização interativa chamada Advise, que permite ter um panorama geral das anotações EC ao longo de diversas versões do Swiss-Prot (como, por exemplo, tendência das anotações se tornarem mais específicas com o decorrer do tempo, conjuntos de entradas cujas anotações se tornaram mais gerais, potenciais correções, dentre outros). O artigo que descreve tal ferramenta, intitulado *Advise: Visualizing the dynamics of enzyme annotations in UniProt/Swiss-Prot*, foi publicado no *IEEE Symposium on Biological Data Visualization (BioVis), 2012* realizado em Seattle, EUA.

A geração de dados biológicos experimentou um crescimento sem precedentes nas últimas décadas, o que criou grandes desafios para a visulização de dados biológicos. Para enfrentá-los, pesquisadores das comunidades de Visualização e Bioinformática devem se engajar no projeto, implementação, aplicação e avaliação de novas técnicas e ferramentas de visualização, o que ajuda a entender os dados altamente volumosos e complexos disponíveis. O Biovis está inserido nesse contexto e é parte da *IEEE VisWeek*, que é o principal fórum de visualização atual e reúne o meio acadêmico, governo e indústria com interesse comum em ferramentas, técnicas e teorias para visualização de dados.

ADVISE: VISUALIZING THE DYNAMICS OF ENZYME ANNOTATIONS IN UNIPROT/SWISS-PROT

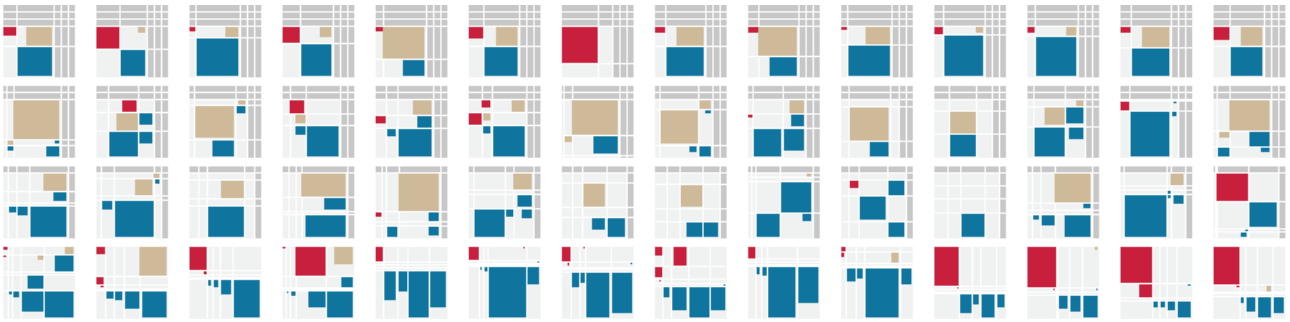
Sabrina A. Silveira*
Universidade Federal de Minas Gerais

Artur O. Rodrigues†
Universidade Federal de Minas Gerais

Raquel C. de Melo-Minardi‡
Universidade Federal de Minas Gerais

Carlos Henrique da Silveira§
Universidade Federal de Itajubá

Wagner Meira Jr.¶
Universidade Federal de Minas Gerais



ABSTRACT

In this paper, we propose an interactive visualization called ADVISe (Annotation Dynamics Visualization), which tackles the problem of visualizing evolutions in enzyme annotations across several releases of the UniProt/SwissProt database. More specifically, we visualize the dynamics of Enzyme Commission numbers (EC numbers), which are a numerical and hierarchical classification scheme for enzymes based on the chemical reactions they catalyze. An EC number consists of four numbers separated by periods and represents a progressively finer classification of the catalyzed reaction. The proposed interactive visualization gives a macro view of the changes and presents further details on demand, such as frequencies of change types segmented by levels of generalization and specialization as well as by enzyme families. Users can also explore entry metadata. With this tool, we were able to identify trends of specialization, database growth and exceptions in which EC numbers were deleted, divided or created and revisions of past annotation errors.

Availability: A video introducing ADVISe is available at *http://vimeo.com/arturhoo/advise* and the source code can be downloaded from *https://github.com/arturhoo/ADVISe*.

Keywords: Information visualization, Bioinformatics, Database dynamics, Enzymes, EC number, UniProt, SwissProt, Annotation, Processing.

*e-mail: sabrinas@dcc.ufmg.br
†e-mail: artur@dcc.ufmg.br
‡e-mail: raquelcm@dcc.ufmg.br
§e-mail: carlos.silveira@unifei.edu.br
¶e-mail: meira@dcc.ufmg.br

IEEE Symposium on Biological Data Visualization 2012
October 14 - 15, Seattle, Washington, USA
978-1-4673-4730-3/12/$31.00 ©2012 IEEE

1 INTRODUCTION

In recent decades, there has been a significant increase in the biological data generated by experimental techniques such as the new generation of DNA sequencing technologies, protein sequencing and protein structure determination. Much of these data are organized and publicly available to the scientific community in biological databases via the Internet. According to [14], these repositories store not only raw biological data but also relevant information such as literature data, protein function and the relationship between a protein and its encoding gene, among other metadata.

Because biological databases are growing at very high rates, most of these metadata are automatically assigned. In many cases, the roles of most genes in various organisms have been reported by homology propagation, without performing any laboratory experiments [4]. To ensure the reliability of these annotations, studies of the reliability of the entries and measures of confidence should be developed. Many studies have drawn attention to error rates in biological database annotations [6, 9, 8, 12, 16, 11].

In fact, the automatic identification of these errors remains an open problem, and several challenges must be overcome. In the absence of laboratory experiments to verify automatically assigned annotations, it will remain impossible to establish a definite conclusion. Many studies have presented comparisons of a diversity of methods of functional annotation, demonstrating that they are widely incompatible and constraining their accuracy.

A major step toward automatic error detection is the description of how and to what extent biological database entry annotations evolve. In other words, we must fully understand why some entries appear to be more stable while others remain more volatile as well as the factors that determine these contrasting behaviors.

The research and development of models and algorithms, coupled with constantly improving visualization resources, represent a promising approach toward understanding how biological databases evolve. Interactive visualizations can be particularly powerful for depicting voluminous, high-dimensional and complex datasets from a macro/micro perspective and to help users unveil trends and exceptions in those datasets.

1.1 Enzyme annotations

By the late 1950s, during a period in which the number of known enzymes was increasing rapidly, it had become evident that the nomenclature of enzymology was becoming unmanageable. In many cases, the same enzymes became known by several different names, while conversely, the same name was occasionally given to different enzymes [21]. Many of the names conveyed little or no idea of the nature of the reactions catalyzed, and similar names were sometimes given to enzymes of quite different types. To address this situation, the General Assembly of the International Union of Biochemistry (IUB) decided, in consultation with the International Union of Pure and Applied Chemistry (IUPAC), to set up an International Commission on Enzymes. Its objective was to consider the classification and nomenclature of enzymes and co-enzymes, their units of activity and standard methods of assay and the symbols used in the description of enzyme kinetics. The Commission prepared a report in 1961 that was promptly adopted and has since been widely used in scientific journals, textbooks, and so on. The size of the Enzyme Commission number (EC number) list has increased steadily since the publication of the first report, and many corrections have been made.

The EC number is a numerical classification scheme for enzymes based on the chemical reactions they catalyze. Each enzyme code consists of four numbers separated by periods. Those numbers represent a hierarchical, progressively finer classification of the catalyzed reaction. For example, the code: 3.4.21.4 represents the following information:

3: hydrolase, which means the enzyme breaks a chemical bond with a water molecule.

3.4: peptidase, which means the broken bond is a peptide bond, i.e., a bond between amino acid residues in a protein chain.

3.4.21: endopeptidase, which breaks an intra-chain peptide bond in which a serine residue participates in the mechanism of catalysis.

3.4.21.4: trypsin, which indicates an enzyme that cleaves mainly at the carboxyl side of the amino acid residues lysine or arginine.

When a new enzyme is annotated, one can add from one to four levels to the EC number, depending on the level of detail of the existing knowledge. In the best scenario, everything is known about the catalyzed reaction as well as the specific substrates and products involved. However, in many cases, when not all of the details about the catalytic activity are known, partial EC numbers, in which the unknown levels are indicated with hyphens, are used to annotate enzymes. The EC number "3.4.21.-", for example, indicates that the specific enzyme substrates are not known, although information about the reaction catalyzed is available.

In this paper, we tackle the problem of analyzing enzyme annotation dynamics and propose a technique to visualize the evolution of these annotations across several releases of the UniProt/SwissProt database. This paper is organized as follows: in section 2, we describe how we modeled the problem. Section 3 details the dataset presented in the visualization. In section 4, we discuss previous related studies, and in section 5, we describe in detail the basis of the technique proposed as well as its capabilities. Finally, we discuss several insights that we obtained in section 6 and conclude the work and present perspectives in section 7.

2 Problem Modeling

Based on the numerical and hierarchical natures of the Enzyme Commission number, we proposed a model to characterize the EC changes observed over several versions of UniProt/SwissProt. Our initial focus was on the visualization of the types of changes that occur and the frequency with which they occur. Furthermore, it is important to know the hierarchical level in which a change occurs because an alteration at a higher level (leftmost) is more severe than at a lower level. Thus, we decided to segment changes by their common prefix length together with the number of generalizations and specializations associated with a specific EC number.

An example of an EC number change characterized by our model is shown below.

$$3.1.3.2 \rightarrow 3.1.3.5$$

This change occurred in 77 hydrolases of release 5 to 6. The common prefix length is 3 (the first three levels from left to right remained the same), there was 1 generalization (number 2 was deleted) and 1 specialization (number 5 was inserted). This change means that an acid phosphatase is now classified as a 5'-nucleotidase.

More examples of EC moves characterized by our prefix/generalization/specialization model are provided in Table 1.

3 Dataset

In this work, we use the biological database UniProt [5], which aims to provide a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation as well as the incorporation, interpretation, integration and standardization of data from a large number of disparate sources. The UniProt Knowledgebase (UniProtKB) is the most comprehensive catalog of protein sequence and functional annotation. As stated by [5], the UniProtKB is an expertly curated database and a central access point for integrated protein information with cross-references to multiple sources.

In accordance with [1], UniProtKB consists of two sections: UniProtKB/SwissProt and UniProtKB/TrEMBL. SwissProt contains manually annotated records with information extracted from the literature and curator-evaluated computational analysis. Annotation is performed by biologists with specific expertise to achieve accuracy. TrEMBL contains computationally analyzed records enriched with automatic annotation and classification. Because SwissProt is considered the gold standard for protein annotation, in this work, we use its data to observe and analyze the changes in EC annotation.

The major releases available in the repositories of the UniProt database at the beginning of this study (March 2009) were downloaded. We analyzed releases 1 (when SwissProt was integrated to UniProt) through 15 (the current release when this study was initiated).

To determine if an EC number change occurred, we examined a database entry EC annotation in two consecutive releases; therefore, the mentioned releases were studied in pairs, and the intersection of identifiers across two consecutive releases was taken.

The total number of entries as well as the number of entries annotated with an EC number, and their percentage in the 15 releases are provided in Table 2. Table 3 shows the number of entries in the set intersection of each release pair.

4 Related Work

We will review different contexts where information visualization techniques have been successfully used in visual analytic processes. In [18], the authors investigated the dynamics of Wikipedia articles through an exploratory data analysis tool that was effective in revealing patterns within a given set of changes in article texts. In [20], a color scheme approach was proposed to present edit histories of Wikipedia administrators. Furthermore, many authors [10, 13, 15, 19] have studied visualizations to facilitate control and understand software source code evolution or to map collaborative efforts of various developers.

Table 1: Example of EC numbers across consecutive database releases and our prefix/generalization/specialization model.

Previous EC number	Actual EC number	UniProt id	Releases	Common prefix length	Degrees of generalizations	Degrees of specializations
-.-.-.-	-.-.-.-	Q9K5T1	1 to 2	0	0	0
3.1.4.14	1.7.-.-	P41407	7 to 8	0	4	2
1.1.1.-	1.-.-.-	P52895	5 to 6	1	2	0
5.3.-.-	5.3.1.27	P42404	14 to 15	2	0	2
2.5.1.64	2.5.1.-	P17109	13 to 14	3	1	0
4.1.1.22	4.1.1.22	P95477	1 to 2	4	0	0

Table 2: Releases 1 to 15 of UniProt/SwissProt.

Release	Release date (MM/DD/YYYY)	% of entries with EC	Number of entries with EC	Total number of entries
1	12/15/2003	37	52,434	141,681
2	07/05/2004	38	57,931	153,871
3	10/25/2004	38	61,229	163,235
4	02/01/2005	38	63,221	168,297
5	05/10/2005	38	69,164	181,571
6	09/13/2005	38	74,468	194,317
7	02/07/2006	39	80,874	207,132
8	05/30/2006	40	89,245	222,289
9	10/31/2006	40	97,508	241,242
10	03/06/2007	40	105,225	260,175
11	05/29/2007	40	108,876	269,293
12	07/24/2007	40	111,230	276,256
13	02/26/2008	43	151,694	356,194
14	07/22/2008	43	168,849	392,667
15	03/24/2009	44	189,234	428,650

Table 3: Release pairs and number of entries in the intersection.

Release pair	Number of entries in $\cap$
1-2	141,249
2-3	151,318
3-4	162,812
4-5	166,933
5-6	181,005
6-7	193,382
7-8	207,069
8-9	222,181
9-10	241,189
10-11	260,065
11-12	269,152
12-13	276,011
13-14	356,036
14-15	392,597

In this work, we are interested in the existence and quantification of specific events of change in enzyme hierarchical annotations. To the best of our knowledge, there are no other works that propose a visualization of this type of data.

5 ADVISE

The main objectives of the proposed visualization were the following:

1. to provide a panoramic macro view of the evolution of EC number annotations;
2. to permit users to explore the complete set of changes, including entry metadata, and the formulation and resolution of general questions about EC number changes.

Concerning the first objective, we wanted to present, in a single perspective, the EC changes segmented by all of the possible combinations of events considering the three parameters of the model (common prefix length, number of generalizations and specializations) across all of the database releases.

5.1 Multivariate display

We have a multivariate problem in which the fundamental task is to simultaneously compare multiple instances of several variables and to permit users to identify similarities and differences among them. Small Multiples of Tufte [17] or Trellis Displays of Cleveland [2, 3] are a straightforward approach to present our data. These approaches consist of splitting the data into multiple graphs that are presented close to each other in the screen, permitting easier examination of the data in a given graph and relatively simple comparison of values and patterns among graphs.

According to Few [7], individual graphs within multiple graphs display a subset of a dataset originally divided according to a categorical variable, and the several graphs differ only in terms of the data displayed. Every graph ideally shares the same type, shape and size and, consequently, the same categorical and quantitative scales. The scales in each graph must start and end with the same values (otherwise the accurate comparison is more difficult). Graphs can be arranged horizontally or vertically or as a matrix in a meaningful order.

5.1.1 Basic frame

With the above in mind, we proceed with our explanation of the proposed visual representation. The basic graph of the proposed Small Multiple representation, which we will refer to as frame, is presented in Figure 1. It is a two-dimensional plot in which we present the number of specializations in the x-axis and the number of generalizations in the y-axis. Both x and y-axes vary in the interval [0,4].

Note some remarkable positions in the frame:

Position (0,0): entries with no changes in the corresponding pair of versions.

Diagonal: entries with the same level of generalizations and specializations, potentially error corrections. They are presented in beige in the Quadmap.

Lower right matrix: entries with more levels of specializations than generalizations; in other words, knowledge about the catalyzed reaction has increased. They are presented in blue in the Quadmap.

Upper left matrix: entries with more levels of generalizations than specializations; in other words, knowledge about the catalyzed reaction has decreased. They are presented in red in the Quadmap.

Invalid positions: if a change retains a common prefix of size 3, it is impossible to have 2 degrees of generalization. These types of events are presented in a dark shade of gray.

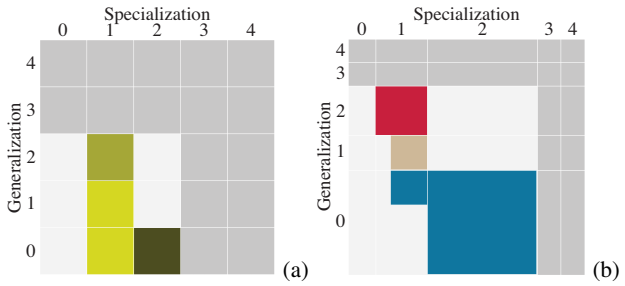


Figure 1: Basic frames for the proposed small multiple visualization. In (a), we present the Heatmap version and in (b), the Quadmap. In (a), the darker the green, the higher the value represented. Likewise, in (b), the bigger the rectangle area, the higher the value. Red represents entries above the diagonal. In blue, we depict entries bellow the diagonal and in beige, we represent diagonal entries. In (a) and (b), dark gray depicts disabled changes, that is, changes that are not possible due to the common prefix length represented by this frame (2 in this case); empty positions are presented in light gray.

Several frames like the one shown are then arranged in a Small Multiple fashion as in Figure 2. On the x-axis, we present the consecutive pairs of releases. The y-axis presents the possible common prefixes in [0,4].

5.1.2 Heatmap

In the first version of the graph, we use a Heatmap representation in which color is a pre-attentive attribute that encodes the frequency of a given change configuration.

The aim of this representation was to bring forth an overview of the complete data, evidencing trends and exceptions across the 15 releases. An interesting feature of this representation is that values in the lower right triangular matrix represent specializations and, in the upper left triangular matrix, generalizations. Consequently, it is easy to recognize global trends toward generalization or specialization in enzyme reaction annotations.

5.1.3 Quadmap

Heatmaps present relevant trends in terms of generalization and specialization occurrences, but we observe two possible drawbacks to that approach.

First, color is not a pre-attentive attribute that is able to precisely encode quantitative data. One can perceive that an intense color represents a higher value than a less intense color. However, it is very difficult to precisely estimate the values from color intensities.

The second drawback is that our Heatmap presents too much blank space. According to Tufte [17], the data density of a graph is the proportion of the total size of the graph that is dedicated to displaying data. Tufte prefers high data density graphs because the human perceptual system is capable of detecting subtle patterns, trends and exceptions. Therefore, we decided to propose a second, complementary view, with the aim of reducing blank (non-data) space as well as improving quantity estimation.

The Quadmap representation was inspired in two-dimensional scatter plots where the points, which we will refer to as positions, are rectangles in which area represents frequency. Although area is not the most precise visual attribute to encode quantity, it is more precise than color. Note, in Figure 1, that it is easier to estimate quantities in the Quadmap (b) than in the Heatmap (a).

It is important to highlight that the axes in Quadmaps are different from one frame to the other, going against the rule of preservation of axis and scale in Small Multiples. This occurs because rectangle sizes distort tics in axes so, to identify the diagonal, lower right and upper left matrix we coloured these elements in beige, blue and red respectively. Nevertheless, we believe this option helps to emphasize trends and exceptions by using colored pixels to represent quantities more precisely than in Heatmaps.

5.2 Analytical interaction and navigation

5.2.1 Filtering, scales and normalization options

The efficaciousness of the information visualization techniques hinge on their ability to clearly and accurately represent information and on the capacity to fathom underlying information through interaction. Indeed, no matter how rich the display is, questions will arise, making interaction a necessary instrument in the pursuit of answers. Furthermore, contrasting different perspectives can lead to different insights. The proposed visualization provides predefined filters and different scaling and normalization options:

1. **Logarithmic or linear scale on the frequencies:** rectangle areas in Quadmap or Heatmap colors are computed according to a logarithmic scale or absolute value of frequencies.

2. **Normalization of frequencies globally or by frame:** global normalization leads to a more realistic view of frequencies, while local (or frame) normalization, despite contradicting Small Multiple rules, emphasizes a part-to-whole relationship into a given frame.

3. **Filter by only changes or presentation of the complete data set**: only entries that suffered changes are showed or the whole dataset (including stable entries). The data are very unbalanced because we have many more stable entries than changes. In conclusion, when we visualize the complete dataset, the changes are de-emphasized.

5.2.2 Hierarchical navigation

A particularly interesting way to create dense graphics is through what Tufte refers to as micro/macro readings [17]. These graphics convey one layer of information on a micro scale and another layer on a zoomed-out, macro scale. A favorable consequence of this technique is that information is consumed hierarchically. The viewer may scan from a distance to observe a global trend and, later, scrutinize closely to examine individual components of that trend. Our multivariate view is a macro view of the entire set of changes in the dataset. Users can click on each frame and see it zoomed in

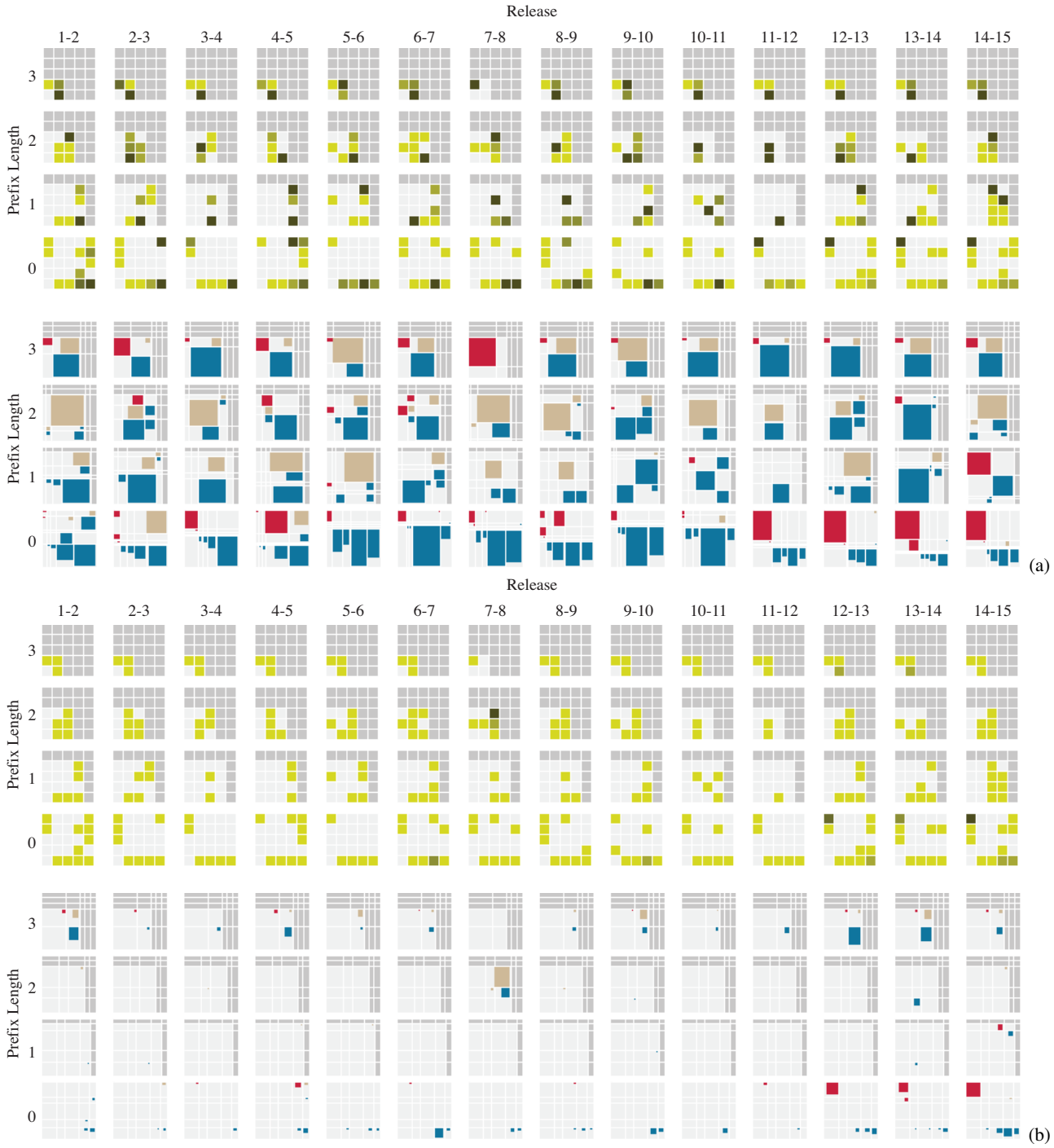


Figure 2: (a) Multivariate view of Heatmap and Quadmap with linear scale, only changes presented and local normalization. (b) Multivariate view of Heatmap and Quadmap with linear scale, only changes presented and global normalization. In (a), local normalization highlights changes that are numerous inside the frames and in (b), global normalization highlights changes that are numerous across all of the releases. Basic frames and axis for Heatmap and Quadmap were detailed in Figure 1.

a micro view. In other words, as observed in Figure 3, users can select a specific release and common prefix length and view a detailed description of the respective frame.

Additionally, users can click on the positions in the micro view and see interactive histograms of each type of change. Through these histograms, users can identify the enzyme families that are subject to that change. These histograms are composed by small rectangles representing each change, and by clicking on individual rectangles, users can view details about that specific entry.

5.3 Implementation

ADVISe was implemented in Processing[1], release 1.5.1. The dataset accessed by our visualization tool was downloaded from UniProt and filtered using Java Development Kit 6[2] to get the data we were interested in: EC number annotation and line types Reference Position (RP), Organism Classification (OC) e Keyword (KW) from UniProt text files. These data were processed by some Python[3] scripts (version 2.6.5) and stored in a MySQL[4] database version 5.1.61.

6 Discussion

In this section, we describe the insights we obtained from ADVISe.

6.1 Trends

6.1.1 Stable enzyme annotations

The most common event spread over the entire dataset is located at the bottom left corner of each frame, position (0,0), and represents pairs of observed EC numbers that remained unchanged in a pair of releases. In this case the two EC numbers involved were equal (i.e., 3.1.3.2 to 3.1.3.2) or there was no EC number (i.e., -.-.-.- to -.-.-.-).

In Figure 4 (a), we present a more realistic view of the dataset, aggregating stable entries (position (0, 0) at each frame) and changes in other positions with global normalization and a linear scale. We can observe a global predominance of entries with no generalization or specialization and prefix length 0. These entries usually have undefined EC numbers (-.-.-.-) that have remained so. Note that the area of this specific position is clearly growing across releases, reflecting the growth in the UniProt/SwissProt database over the fifteen analyzed releases.

In Figure 4 (b), we show the same data normalized by frame, revealing that stable entries are predominant in almost every frame. Exceptions do exist and will be discussed in section 6.2.

6.1.2 Generalization versus Specialization

In the Heatmap of Figure 2 (a), we can observe that the lower triangular matrices have more entries than the upper triangular matrices and thus, in the entire dataset, there were more specializations than generalizations. In the Quadmap of Figure 2 (a), in which we present only changes in linear scale and local normalization, we can observe a predominance of blue rectangles representing this trend. Once again, exceptions are apparent, and some will be discussed further in section 6.2.

Figure 2 (a) also emphasizes that the line representing no generalizations in the bottom row of frames (common prefix length 0) in the multivariate matrix is a frequent type of change. It reveals an interesting trend of specialization for entries without annotation (-.-.-.-) because they tend to receive EC levels in each release.

[1]http://processing.org/

[2]http://www.oracle.com/technetwork/java/index.

[3]http://www.python.org/

[4]http://www.mysql.com/

6.2 Exceptions

6.2.1 Annotation deletion

The four positions indicated by red rectangles on the bottom row of Figure 2 (b), in which the parameters are common prefix length 0, 4 degrees of generalization and no specialization in releases 12-13, 13-14 and 14-15, represent a drastic change in which the four levels of EC numbers were deleted. Table 4 shows the frequencies associated with each position.

Table 4: Frequency of four-level EC number deletion from releases 11 to 15.

Pair of releases	Frequencies
11-12	146
12-13	1,357
13-14	1,006
14-15	1,976

EC numbers must be assigned to protein catalytic subunits. This implies that in large protein complexes, only one or a few of the subunits will be annotated with an EC number. Indeed, proteins can have non-catalytic functions such as transport of substances or an immunological or structural role. In some cases, automatic annotation can assign EC numbers to a whole complex, including non-catalytic subunits. Positions that symbolize such cases in ADVISe represent corrections in which the curators completely removed the EC numbers because the related subunits are not enzymes. We present three examples of UniProt/SwissProt entries that experienced four-level EC number deletion from releases 12 to 13.

- Identifier Q6FSJ2, which was annotated as 1.10.2.2 in version 12, is subunit 7 of cytochrome b-c1 but is not the subunit with reductase activity.
- Identifier Q8LX28, which was annotated as 3.6.3.14 in version 12, is subunit 8 of ATP synthase, which is part of the membrane proton channel.
- Identifier Q6AY96, which was annotated as 2.7.11.1 in version 12, is a subunit of a transcription factor but is not the subunit with serine/threonine kinase activity.

6.2.2 Deleted EC numbers

In Figure 2 (b), a total of 1,900 EC number changes are represented by the position with common prefix length 2, 2 degrees of generalization and 2 degrees of specialization in releases 7 to 8. The three most numerous changes depicted in this position are, respectively, 2.7.1.37 to 2.7.11.1 (918 entries), 2.7.1.112 to 2.7.10.1 (215 entries) and 2.7.1.112 to 2.7.10.2 (165 entries). As stated by IUBMB, EC number 2.7.1.37 was deleted and divided in 2005 into 2.7.11.1, 2.7.11.8, 2.7.11.9, 2.7.11.10, 2.7.11.11, 2.7.11.12, 2.7.11.13, 2.7.11.21, 2.7.11.22, 2.7.11.24, 2.7.11.25, 2.7.11.30 and 2.7.12.1. Similarly, EC number 2.7.1.112 was deleted and divided into 2.7.10.1 and 2.7.10.2. In such cases, transferase annotations, and more specifically, EC numbers beginning with 2.7 (transferring phosphorus-containing groups), underwent a revision caused by a change in the EC number classification system and not a change in enzyme function annotation.

A similar phenomenon occurred at the position with a common prefix length 1, 2 degrees of generalization and 3 degrees of specialization in releases 14 to 15 (212 entries). This position can be better visualized in the Quadmap of Figure 2 (b) and represents the EC number change 2.5.1.- (transferring alkyl or aryl groups other than

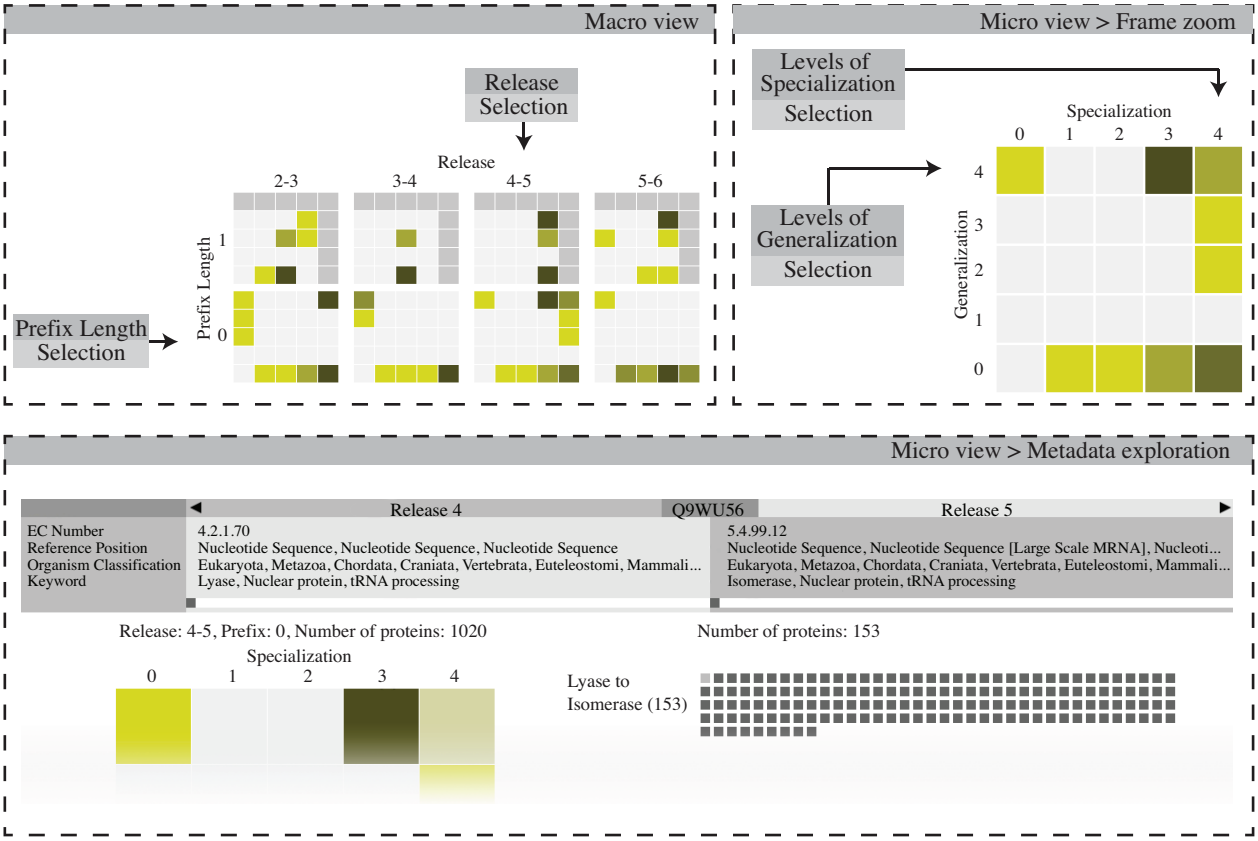


Figure 3: Navigation scheme.

methyl groups) to 2.2.1.9 (2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid synthase). The EC number 2.5.1.64 was created in 2003 and deleted in 2008, when it was divided into 2.2.1.9 and 4.2.99.20. In this case, the annotation changes are due to the creation of a new EC (2.2.1.9); in other words, there was a change in the EC number classification system.

6.2.3 Created EC numbers

In some cases, enzymes were integrated into the UniProt/SwissProt database when their catalytic activity was already known but there were no appropriate EC numbers defined by IUBMB to describe this specific catalytic activity. For example, in Figure 2 (b), the position with common prefix length 3, no generalizations and 1 degree of specialization in releases 12 to 13 represents a total of 637 EC number changes. A representative EC number change depicted by this position is 2.8.1.- (sulfurtransferases) to 2.8.1.8 (EC created in 2006 to represent lipoyl synthase), with 117 entries. The UniProt entry Q7UH37 exhibited this change. It was integrated to UniProt on 10 May 2004, and its associated function was lipoyl synthase. However, there was not an EC number related to lipoyl synthase at that time, and this entry remained with the same incomplete EC number, 2.8.1.-, until release 13 (26 Feb 2008), when it was annotated with EC number 2.8.1.8.

6.2.4 Annotation errors

Another exception we detected is presented in Figure 2 (b) by the red position with common prefix length 1, 3 degrees of generalization and 2 degrees of specialization in releases 14 to 15. This position represents a single type of change that occurred 261 times. The EC number change was 2.1.1.61, which was created in 1982 and is associated with tRNA (5-methylaminomethyl-2-thiouridylate-methyltransferase) activity, to 2.8.1.-, which is associated with sulfurtransferase activity. The EC number 2.1.1.61 was not deleted, and thus, the EC number change was a correction to annotate the associated entries with a more appropriate catalytic function.

7 Conclusions and Future Works

In this paper, we proposed ADVISe, an interactive tool to visualize the dynamics of enzyme annotation evolution, and specifically, EC numbers, across several releases of the UniProt/SwissProt database. We modeled the changes of consecutive releases with the parameters of common prefix length and levels of generalization and specialization. The proposed interactive visualization gives a macro view of the changes and presents further details on demand such as frequencies of types of changes segmented by levels of generalizations and specializations as well as by enzyme family. Users can further explore entry metadata. By visual inspection, we were able to identify trends of specialization and database growth as well as detect several exceptions in which EC numbers were deleted, divided or created or annotation errors were corrected.

In future work, we intend to implement a consensus view to summarize each line and generate a frame that is representative of the trends related to each common prefix length. As a consequence, we believe we will be able to spot relevant exceptions relative to the pattern. We will highlight these exceptions automatically to simplify the visual analytical process. Furthermore, we want to investigate methods to allow users to annotate insights from specific positions of the frames so that we can collect relevant data from expert users for further studies. Last, but not least, we are planning

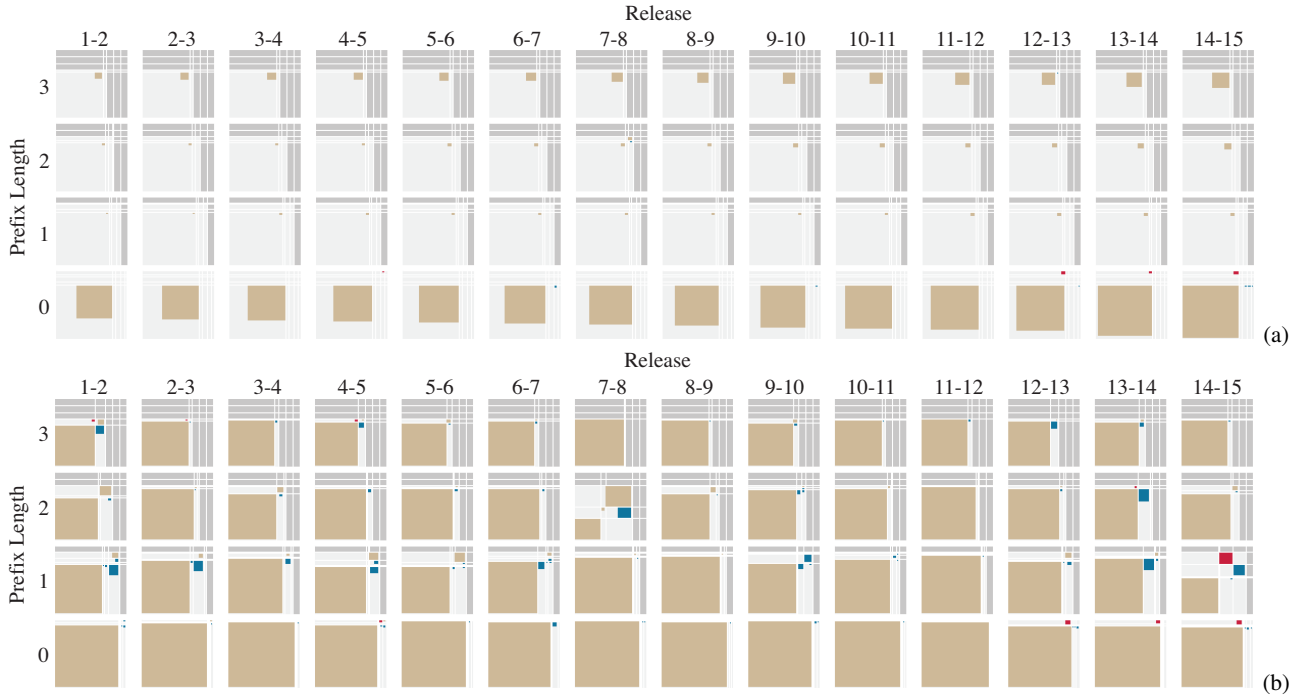


Figure 4: (a) Multivariate view of Quadmap with linear scale, stable entries and changes presented and global normalization. (b) Multivariate view of Quadmap with linear scale, stable entries and changes presented and local normalization. Basic frames and axis for Quadmap were detailed in Figure 1.

to systematically measure user insights and impressions about the proposed visualization.

Acknowledgements

This work was supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Financiadora de Estudos e Projetos (FINEP) and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

References

[1] R. Apweiler, M. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, R. Antunes, D. Barrell, B. Bely, M. Bingley, D. Binns, et al. The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, 38:D142–D148, 2010.

[2] R. Becker, W. Cleveland, M. Shyu, and S. Kaluzny. Trellis display: a framework for visualizing 2d and 3d data. Technical report, 1994.

[3] R. Becker, W. Cleveland, M. Shyu, and S. Kaluzny. Trellis display: User's guide. Technical report, 1994.

[4] S. Brenner et al. Errors in genome annotation. *Trends in Genetics*, 15(4):132–133, 1999.

[5] U. Consortium et al. Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res*, 40:D71–D75, 2012.

[6] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends in Genetics*, 17(8):429–431, 2001.

[7] S. Few. *Now you see it*. 2009.

[8] W. Gilks, B. Audit, D. de Angelis, S. Tsoka, and C. Ouzounis. Percolation of annotation errors through hierarchically structured protein sequence databases. *Mathematical biosciences*, 193(2):223–234, 2005.

[9] M. Green and P. Karp. Genome annotation errors in pathway databases due to semantic ambiguity in partial ec numbers. *Nucleic acids research*, 33(13):4035–4039, 2005.

[10] R. Holt. Gase: visualizing software evolution-in-the-large. In *Proceedings of the Third Working Conference on Reverse Engineering*, pages 163–167, 1996.

[11] S. Hung, J. Wasmuth, C. Sanford, and J. Parkinson. Detect - a density estimation tool for enzyme classification and its application to plasmodium falciparum. *Bioinformatics*, 26(14):1690–1698, 2010.

[12] C. Jones, A. Brown, and U. Baumann. Estimating the annotation error rate of curated go database sequence annotations. *BMC bioinformatics*, 8(1):170, 2007.

[13] M. Lanza. The evolution matrix: recovering software evolution using software visualization techniques. In *Proceedings of the 4th International Workshop on Principles of Software Evolution*, 2001.

[14] A. Lesk and J. Wiley. *Database annotation in molecular biology*. Wiley Online Library, 2005.

[15] F. V. Rysselberghe. Studying software evolution information by visualizing the change history. In *Proceedings of the 20th IEEE International Conference on Software Maintenance*, pages 328–337, 2004.

[16] A. Schnoes, S. Brown, I. Dodevski, and P. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605, 2009.

[17] E. Tufte. *Envisioning information*. 1990.

[18] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 575–582, New York, NY, USA, 2004. ACM.

[19] L. Voinea, A. Telea, and J. van Wijk. Cvsscan: visualization of code evolution. In *Proceedings of the 2005 ACM Symposium on Software Visualization*, 2005.

[20] M. Wattenberg, A. B. Vigas, and K. Hollenbach. Visualizing activity on wikipedia with chromograms. In *In Proceedings of INTERACT*, pages 272–287, 2007.

[21] E. Webb et al. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Number Ed. 6. Academic Press, 1992.

Apêndice C

Artigo Submetido

O artigo intitulado *ENZYMAP: Exploiting protein metadata for modeling and predicting annotation changes in UniProt/Swiss-Prot*, que contém os resultados dessa tese, foi submetido à revista Bioinformatics[1]. Tal revista é uma publicação oficial da *The International Society for Computational Biology* (ISCB) e é líder no campo da Bioinformática, sendo ainda um periódico qualis A1 segundo a *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES). Seu fator de ipacto é atualmente 5,468.

Os autores acreditam no potencial de contribuição do presente trabalho à comunidade científica, daí a escolha de um renomado periódico para submissão do artigo. Isso é importante na divulgação da metodologia inovadora empregada, que busca prever possíveis alterações de EC em entradas do Swiss-Prot através de uma estratégia baseada em aprendizagem supervisionada alimentada com mudanças de EC já experimentadas por outras entradas da base modeladas em termos dos metadados as caracterizam.

[1] `http://bioinformatics.oxfordjournals.org/`

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

ENZYMAP: Exploiting protein metadata for modeling and predicting annotation changes in UniProt/Swiss-Prot

S. A. Silveira [1,2*], R. C. de Melo-Minardi [1*], C.H. da Silveira [3], M.M. Santoro [2] and W. Meira Jr. [1*]

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The volume and diversity of biological data are increasing at very high rates. Vast amounts of protein sequences and structures, protein and genetic interactions and phenotype studies have been produced. The majority of data generated by high-throughput devices is automatically annotated because manually annotating them is not possible. Thus, efficient and precise automatic annotation methods are required to ensure the quality and reliability of both the biological data and associated annotations.

Results: We proposed ENZYMAP, a technique to characterize and predict EC number changes based on metadata from UniProt/Swiss-Prot using a supervised learning approach. We evaluated ENZYMAP experimentally and showed that predicting EC changes using selected metadata is possible. Finally, we compared ENZYMAP and DETECT with respect to their predictions and checked both against the UniProt/Swiss-Prot annotations. ENZYMAP was shown to be more accurate than DETECT, coming closer to the actual changes in UniProt/Swiss-Prot.

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

Contact: sabrinas@dcc.ufmg.br, raquelcm@dcc.ufmg.br and meira@dcc.ufmg.br

1 INTRODUCTION

In recent decades there has been a surge in the amount of biological data available. According to Fritz et al., 2011, new DNA sequencing technologies allowed a 1000-fold drop in sequencing costs since 1990 and made an increasing number of large data collection projects economically possible, leading to an exponential increase in the DNA sequence data available. Additionally, vast amounts of data, such as protein sequences and structures, gene-expression measurements, protein and genetic interactions and phenotype studies, have been generated (Howe et al., 2008). A significant portion of these data are organized and publicly available to the scientific community in biological repositories accessible through the Internet. In accordance with Lesk and Wiley, 2005, these repositories store not only biological raw data but also relevant information such as protein function, literature information and the relationship between a protein and its encoding gene, among other metadata, also called annotation.

Considering the existing and the increasing volumes of biological data, a common approach involves selected data sets of high relevance being manually curated by experts while most data are automatically annotated (Mewes *et al.*, 2011). In the majority of cases, the roles of genes have been reported by sequence similarity propagation without experimental evidence (Furnham *et al.*, 2009; Brenner *et al.*, 1999). Glycoprotein G of the Nipah virus (entry Q9IH62 in Swiss-Prot) illustrates the drawbacks of this approach. When considering residue similarity, it is very similar (more than 50%) to hemagglutinin-neuraminidases, an enzyme group associated with viral attachment and fusion to the host cell. The structures of Glycoprotein G of the Hendra and Nipah viruses were solved (PDB id 2VSK and 2VSM, respectively), revealing the six-blade β propeller structural motif typical of these hydrolases (Bowden *et al.*, 2008). A structural alignment with a legitimate neuraminidase from Parainfluenza Virus Type III (1V3D), which also belongs to the same Paramyxoviridae family of Henipavirus, resulted in a RMSD lower than 2.0 Å(Lawrence *et al.*, 2004). Thus, an automated system based on such similarities may erroneously classify the function of Glycoprotein G of Henipavirus as having neuraminidase activity. In fact, up to release 14 of UniProt/Swiss-Prot (Consortium et al., 2012) (July 2008), entry Q9IH62 was considered an enzyme. However, despite all these sequence and structural similarities, Henipavirus Glycoproteins G are now known to not be enzymes and to have only hemagglutinin activity, performing protein-protein interactions with host receptors (Bowden *et al.*, 2008). At the time we wrote this article, the PDB[1] still classified them as hydrolases. In summary, the scientific community still has concerns regarding the quality and reliability of the data and annotations from the large, publicly available databases.

As we mentioned, biological repositories almost always store some metadata that characterize and provide biological context to the raw data. In this work, we investigate the extent to which these metadata may be used to detect problems in the database. In particular, we want to verify whether the UniProt/Swiss-Prot metadata are good indicators that an annotation change will occur and determine how we can systematically perform such predictions.

*To whom correspondence should be addressed

[1] http://www.rcsb.org/pdb/

In this work, we propose a supervised learning approach to characterize and predict annotation changes in temporal data, which we named *ENZYmatic Metadata Annotation Predictor* (ENZYMAP). More precisely, we are interested in predicting enzyme function annotation based on UniProt/Swiss-Prot entry metadata. This proposal allows us to suggest corrections to annotations from biological repositories and may be used together with other annotation methods, improving the quality and reliability of these data. Our approach uses data already available to enhance the repository and thus does not demand new expensive bench experiments. Furthermore, there is a huge volume of data that cannot be analyzed manually, hence the importance of reliable automatic annotation methods.

1.1 Enzyme Annotations

In this work, enzyme function annotation refers to the Enzyme Commission (EC) number (Webb *et al.*, 1992), which is a numerical classification scheme for enzymes based on the chemical reactions they catalyze. Each enzyme code consists of four numbers separated by periods. Those numbers represent a hierarchical, progressively finer classification of the catalyzed reaction. For example, the code *3.4.21.4* represents the following information: (*3*) hydrolase, indicating that the enzyme breaks a chemical bond involving a water molecule; (*3.4*) peptidase, indicating that the broken bond is a peptide bond, i.e., a bond between residues in a protein chain; (*3.4.21*) endopeptidase, indicating that an intra-chain peptide bond is broken and that a serine residue participates in the mechanism of catalysis; and (*3.4.21.4*) trypsin, indicating an enzyme that cleaves mainly at the carboxyl side of lysine or arginine residues.

The EC classification system is known to have some drawbacks. Green and Karp, 2005 reported a systematic annotation error in genome and pathway databases resulting from the misinterpretation of partial EC numbers. The key issue is that different enzymes that catalyze different reactions within the same class can be assigned the same partial EC number but the same partial EC number does not mean that the enzymes have the same activities. Egelhofer *et al.*, 2010 stated that the same reaction can be correctly annotated with different EC numbers. For example, the reaction catalyzed by the sterol 14-demethylase (1.14.13.70) is correctly assigned to 1.14.13, but it could be assigned to 1.14.21. These two sub-subclasses are similar and could be merged without a loss of information. In addition, according to Egelhofer *et al.*, 2010, the general principle that the enzyme class is defined by its chemical reaction is violated in some cases. For example, the reaction $ATP + H2O = ADP + phosphate$ is catalyzed by the enzymes adenosinetriphosphatase (3.6.1.3) and myosin ATPase (3.6.4.1). In 3.6.1.3, the ATPase activity is not connected to actin movement, but in 3.6.4.1 it is.

Nonetheless, we chose to analyze the EC number as an enzyme function annotation because it is a mature and widely adopted enzyme classification scheme yet a controlled vocabulary that is numerical and hierarchical, which makes it particularly complex and interesting for computational modeling and description.

2 RELATED WORK

Several studies have drawn attention to the error rates in biological database annotation. Here, we briefly review some of them.

Brenner *et al.*, 1999 compared annotations in *Mycoplasma genitalium* performed by three different groups and detected an error rate from 7% to 15% (depending on the gene analyzed and the group responsible for the analysis). Devos and Valencia, 2001 estimated the error rates in the genomes of *Mycoplasma genitalium*, *Haemophilus influenzae* and *Methanococcus jannaschii* by counting the number of discrepancies in sets of similar proteins and concluded that the error rates vary from 4% to 40% for the first genome and from 4% and 34% for the last two genomes. Both analyses were based on the discrepancies of annotations made by different research groups for very specific genomes, which allows the placement of a lower limit on the likely levels of misannotation according to Schnoes *et al.*, 2009.

A systematic annotation error in genome and pathway databases that results from the misinterpretation of partial EC numbers was reported in Green and Karp, 2005. This error results in the assignment of genes annotated with a partial EC number to many or all biochemical reactions that are annotated with the same partial EC. For example, in KEGG (Kanehisa *et al.*, 2012), out of 135 genes from *Escherichia coli* annotated with a partial EC number, 58 were incorrectly assigned to reactions.

Schnoes *et al.*, 2009 investigated the levels of misannotation for the molecular function in UniProtKB/Swiss-Prot, GenBank Non-redundant (NR), UniProtKB/TrEMBL and KEGG for 37 enzyme families with experimental evidence. Swiss-Prot presented error levels close to 0% for most families, whereas GenBank NR, TrEMBL and KEGG showed high levels of misannotation, from 5%-63%, across the six studied superfamilies. Furthermore, an analysis of the sequences from GenBank NR showed that the level of misannotation was close to 0% in 1999 but was approximately 40% in 2005, indicating that misannotation increased during that period.

Egelhofer *et al.*, 2010 investigated inconsistencies in the EC number classification scheme as they can lead to inconsistencies in enzyme annotation. The authors validated the data of 3,788 enzymatic reactions and found a greater than 80% agreement between their assignment and the EC scheme. These results can be used to make corrections and improve the EC number classification.

These works focused on the levels of misannotation, showing that they are significant in a variety of databases, even those with manual revision such as UniProt/Swiss-Prot. The following works are related to annotation prediction. The Density Estimation Tool for Enzyme ClassificaTion (DETECT), a probabilistic method for enzyme prediction based on both global and local sequence alignments, was presented in Hung *et al.*, 2010. It uses a Bayesian framework to integrate information from density estimation profiles generated for each EC number. Compared with BLAST, DETECT improved the enzyme annotation accuracy and, when applied to *Plasmodium falciparum*, identified potential annotation errors.

In Quester and Schomburg, 2011, the EnzymeDetector was implemented to automatically compare and evaluate the assigned enzyme functions from some annotation databases (NCBI RefSeq (Pruitt *et al.*, 2009), KEGG, PEDANT (Walter *et al.*, 2009), (Winsor *et al.*, 2011) (Winsor et al., 2011) and UniProt/Swiss-Prot) and to supplement them with its own function prediction. In the same work, the authors analyzed nine prokaryotic genomes and found approximately 70% inconsistencies in the enzyme predictions of the annotation resources used.

Funtree is presented in Furnham *et al.*, 2012 as a resource that combines structural, sequence, phylogenetic and functional data for structurally defined enzyme superfamilies. The authors stated that combining these data into a single resource enables the investigation of how novel enzyme functions have evolved, which can help predict the functions of uncharacterized enzymes.

In this work, we propose ENZYMAP, a supervised learning approach to characterize and predict annotation changes in temporal data based on annotation metadata. To the best of our knowledge, there are no other works that propose this type of approach to improve the quality of biological annotations over time.

3 MATERIALS AND METHODS

To characterize and predict the EC number changes, we performed three supervised learning experiments in this work: *Descriptive Multiclass*, which is intended to verify whether separating entries in UniProt/Swiss-Prot that suffered a specific change in the EC number from those that remained constant based on entry metadata is possible; *Predictive Multiclass*, which attempts to use all available data in the repository to predict an upcoming EC change; and *Predictive Common Source*, which segments EC changes by their common source (EC annotation before the change) to improve the latter experiment. In the next sections, we detail the data employed in these experiments, how the EC changes and metadata that describe such changes were modeled and the techniques used to construct our approach.

3.1 Data

The EC number annotations of entries from the biological database UniProtKB/SwissProt were studied in this work. A set of 44 major releases available in UniProt repositories in May 2012 were downloaded. Releases 1 to 44 were analyzed.

To determine whether a specific UniProt/Swiss-Prot entry has undergone an EC number change, checking that entry's EC number in two consecutive releases is necessary, and therefore the 44 releases were analyzed in pairs, taking the set intersection of identifiers in two consecutive releases. A total of 18,727,155 EC pairs were obtained from the entire data set. Among them, 55,908 are pairs with different EC numbers.

The total number of entries, the number of entries annotated with an EC number and their percentage in the 44 releases are provided in Table 1 and Figure 1 from the Supplementary Material. The number of entries in the set intersection of each release pair is shown in Table 2 and Figure 2 from the Supplementary Material.

3.1.1 Selected Metadata In addition to the EC number change data from the 44 UniProt/Swiss-Prot releases, in the Descriptive Multiclass, Predictive Multiclass and Predictive Common Source Experiments, we are interested in the entry metadata (annotation attributes) able to characterize such changes. The following line types from UniProt/Swiss-Prot text files were selected as candidate annotation attributes. First, Organism Classification (OC), which refers to the taxonomic classification of the source organism, was selected because it includes extensively studied organisms, which could potentially lead to good quality annotation. *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Caenorhabditis elegans* are examples of that; Second, Reference Position (RP) describes the extent of the work (reference) relevant to the entry. Entries with more specific references in RP (e.g., function) likely have better annotation than entries with general references (e.g., large scale genomic DNA). Third, KeyWord (KW), which provides information that can be used to generate indices of the sequence entries based on functional, structural, or other categories, represents a type of summary for each entry and contains relevant words related to that protein. An example of the selected line types is provided below for UniProt/Swiss-Prot id P66880, whose EC number is currently 3.1.3.5. Further information regarding the line types and UniProt text file format can be obtained from the UniProtKB User Manual[2].

```
OC   Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;
OC   Brucellaceae; Brucella.
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
KW   Complete proteome; Cytoplasm; Hydrolase; Metal-binding;
KW   Nucleotide-binding.
```

3.2 Problem Modeling

3.2.1 Initial Exploration Based on the numerical and hierarchical nature of the EC number, we proposed a model to characterize the EC changes observed across the releases of UniProt/Swiss-Prot. Knowing the hierarchical level in which a change occurs is important because an alteration at a higher level (leftmost) is more severe than that at a lower level. Thus, we decided to characterize the EC changes observed in release pairs by the following parameters: common prefix length, number of generalizations and specializations. Common prefix length refers to the number of levels that remained the same from left to right; the number of generalizations and specializations represent the number of deleted and added levels, respectively. Examples of EC changes described by our model are provided in the Supplementary Material, Table 3.

To tackle the problem of analyzing and visualizing such a large amount of data representing the evolution of enzyme annotations across several releases of the UniProt/Swiss-Prot database, we proposed ADVISe (Silveira *et al.*, 2012). It is a tool that provides a panoramic macro view of EC changes and presents further details on demand, such as the frequencies of change types segmented by a common prefix length, by the levels of generalization and specialization as well as by the UniProt/Swiss-Prot releases and by the enzyme families (leftmost EC level). Consequently, the trends of specialization, database growth and exceptions in which the EC numbers were deleted, divided or created and the revisions of past annotation errors can be identified using this tool. ADVISe also allows users to explore and compare the entry metadata used in our supervised learning approach.

3.2.2 Descriptive and Predictive Experiments The data modeling was the same for the three performed experiments: Descriptive Multiclass, Predictive Multiclass and Predictive Common Source.

The training data from the EC changes and non-EC changes (also called the control set) are required to characterize and predict the EC number changes using a supervised learning approach. The algorithm needs to learn from these data in a training phase to be able to subsequently separate a set of entries that underwent EC changes from a set in which the EC annotations remained the same. For example, the entry with UniProt/Swiss-Prot id Q9PKH4, which underwent the EC change 3.1.3.2 to 3.1.3.5 from release 5 to 6, is an example of EC change type $3.1.3.2 \rightarrow 3.1.3.5$, and the id P20611, for which the EC annotation remained the same from release 5 to 6, is an example of control set $3.1.3.2 \rightarrow 3.1.3.2$.

Here, we proposed an occurrence data matrix to model the EC changes and non-changes. In such a matrix, the columns represent features (terms obtained from the OC, RP and KW line types as explained in Section 3.3.1), and the rows represent instances of the change or control set. The position i, j of this matrix is one whenever the instance of index i (a given entry) has the annotation attribute corresponding to the column of index j and is zero otherwise. The last column represents the classes for each instance (row). The classes were modeled considering a source EC number (before the EC change) and a destination EC number (after the EC change), so an instance whose class is $3.1.3.2 \rightarrow 3.1.3.5$ had its EC annotation changed from 3.1.3.2 to 3.1.3.5. A fragment of an occurrence matrix showing the EC change $3.1.3.2 \rightarrow 3.1.3.5$, which occurred from release 5 to 6, and its control is provided in Table 1.

3.3 Technique

[2] http://web.expasy.org/docs/userman.html

Table 1. Fragment of occurrence matrix proposed to model the EC changes and non-changes.

Id	F1	F2	F3	F4	F5	Class
Q8TUG3	1	1	0	1	0	$3.1.3.2 \rightarrow 3.1.3.5$
O67004	1	1	0	1	0	$3.1.3.2 \rightarrow 3.1.3.5$
P34724	0	0	1	0	1	$3.1.3.2 \rightarrow 3.1.3.2$
P44009	0	1	0	1	1	$3.1.3.2 \rightarrow 3.1.3.2$

F1 = nucleotide-binding, F2 = magnesium, F3 = eukaryota, F4 = metal-binding, F5 = signal.

3.3.1 Generation of occurrence matrix To generate an occurrence data matrix to feed the supervised learning approach, for each type of EC change and for each release of the UniProt/Swiss-Prot in which such a change occurred, we parsed the text files of the entries that experienced the change and the files of the control group entries to extract the annotation attributes OC, RP and KW. We performed a text preprocessing on these metadata, which is a set of techniques applied in the text to reduce the data dimensionality and ambiguity. The following text preprocessing tasks were performed: *Normalization*, which is intended to remove punctuation from the text and convert the characters to lowercase; *Stop word removal*, which aims to remove stop words, which are extremely common words, such as pronouns and articles, and do not add information; *N-grams*, which is a contiguous sequence of n words from a given sequence of text that is used to capture some context present in the analyzed metadata and to match not only exact terms but also approximate ones; *Stemming*, which is an algorithm that reduces inflected words to their stem, such as the words stem, stems and stemming, which have the root stem. The employed stemmer was a Java implementation of the Porter stemming algorithm (Porter *et al.*, 1980), downloaded from the author's website[3].

We processed the metadata from line types OC, RP and KW, which resulted in a set of features for the classification task. Considering a given type of change, the release in which this change occurred and an entry that underwent such a change, we extracted metadata from this entry for all releases prior to the change (until the release immediately before the change).

3.3.2 EC change selection As in the Descriptive Multiclass Experiment we employed a ten-fold cross-validation to evaluate the performance of our supervised learning approach, change types with at least ten examples were selected (discarded and used types of EC changes in each release are presented in Figure 3 of the Supplementary Material). The total number of EC change types was 1,968. Among them, 508 EC change types had at least 10 examples. Here, examples of the change types are $3.1.3.2 \rightarrow 3.1.3.5$ and $4.-.-.- \rightarrow 4.1.99.17$. Q8TUG3 and O67004 are examples of change type $3.1.3.2 \rightarrow 3.1.3.5$.

For some change types, such as $-.-.-.- \rightarrow 5.2.1.8$ from release 39 to 40, there were many examples (288,932) in the control set, $-.-.-.- \rightarrow -.-.-.-$, which represent entries that were not annotated with an EC in release 39 and remained without an EC annotation in release 40. Thus, we set an upper limit to the number of examples in the control set; otherwise, performing the tasks of dimensionality reduction (detailed in Section 3.3.3) and classification would not be possible due to the computational cost. The upper limit chosen for the examples in the control set is the median of the number of EC change examples, which is 27, because it is more representative of the number of examples than the mean, which is 102.2 with a standard deviation of 224.6. Additional information is provided in Figure 4 of the Supplementary Material.

[3] http://tartarus.org/∼martin/PorterStemmer/java.txt

3.3.3 Dimensionality reduction through SVD Singular Value Decomposition (SVD) is a technique from linear algebra in which an m by n matrix A can be represented by the product $U\Sigma V^T$ where U is an m by m matrix and its columns are the left singular vectors of A; Σ is an m by n diagonal matrix with its values in descending order; and V is an n by n matrix and its columns represents right singular vectors of A. To compress the data used in the classification task, reducing the number of features and noise, yet maintaining relevant semantic relationships among the terms, matrix A can be approximated by matrix A_k (with rank k where k is less than the rank of A) as: $A_k = U_k \Sigma_k V_k^T$.

To achieve A_k, the first k singular values of A were taken, and thus the resulting matrix has k features: $A_k = U_k \Sigma_k V_k^T = U_k(\Sigma_k V_k^T) = U_k(D_k)$. According to Eldén, 2006, A_k can be computed using only matrix D_k, which is: $A_k = \Sigma_k V_k^T = D_k$.

A similar approach to compute D_k was adopted in Pires *et al.*, 2011. As stated by Deerwester *et al.*, 1989, the choice of k is an empirical matter; therefore approximations with k from 1 to 100 were generated, and the matrix that led to the best classification model was chosen. It is important to highlight that the applied dimensionality reduction via SVD may reduce the computational cost and memory requirements of the algorithms used in the classification task. SVD was used and discussed in a similar way in several studies (Berry *et al.*, 1995; del Castillo-Negrete *et al.*, 2007; Bécavin *et al.*, 2011; Deerwester *et al.*, 1989).

3.3.4 Classification In accordance with (Pang-Ning *et al.*, 2006), classification is a supervised learning technique that consists of associating one or several predefined labels or classes with data objects. A classification model may be viewed as a function f that maps a set of attributes x to a given class y. The classification task is represented in Figure 5 of the Supplementary Material and is performed as follows in each experiment:

Descriptive Multiclass Experiment: This step aimed to verify whether the annotation attributes OC, RP and KW are able to discriminate entries that underwent a specific change in the EC number from those in which the EC annotation remained the same. We generated classification models using data matrices (constructed from the entire dataset, that is, the 44 UniProt/Swiss-Prot releases) that we reduced via SVD using k from 1 to 100, and we selected the best classification model. We evaluated the model performance through a ten-fold cross-validation.

Predictive Multiclass Experiment: We used EC change types previously modeled in the Descriptive Experiment to construct a classification model and predict the EC changes. Here, we reserved the last release in which a change type occurred to test the model. We consider as modeled EC change types those that had F_1 score greater than 0.5 (we detailed the F_1 score in Section 3.3.5). Only those were used because the change types that were not characterized in the Descriptive Experiment (in which the entire data set was used and a cross-validation was performed) are not expected to be predicted.

Predictive Common Source Experiment: We segmented the data set from the Predictive Multiclass by the common source, and each source corresponds to a classifier. The common source here is the previous EC number (before the EC change) associated with an entry. For example, the EC changes $2.1.1.- \rightarrow 2.1.1.189$, $2.1.1.- \rightarrow 2.1.1.190$ and their control $2.1.1.- \rightarrow 2.1.1.-$ have the common source EC 2.1.1.-, and there is one classifier in which the possible classes are these three EC changes. We performed this experiment expecting that making correct predictions using a more specialized classifier would be easier than the multiclass experiment in which a single classifier has 664 classes.

We employed and compared the classification algorithms Naïve Bayes (John and Langley, 1995), K Nearest Neighbor (KNN) (Aha *et al.*, 1991) and C4.5, also called J48 (Quinlan, 1993). We chose these algorithms due to their low memory requirements and short execution time.

3.3.5 Classifier evaluation strategy We performed several experiments to choose the best classification model. We used the 100 matrices resulting from SVD with k (number of features or columns) varying from 1 to 100, and for each matrix, we applied three classification algorithms: Naïve Bayes,

KNN with $K = \{1, 3, 5, 7, 10\}$ and J48. To assess the performance of the classifiers, we used the metrics F_1 score (also called F measure) and Area Under the ROC Curve (AUC) (Fawcett, 2006).

The F_1 score is the harmonic mean of precision (p) and recall (r), and it tends toward the least of these elements ($F1 = 2\frac{p \times r}{p+r}$). Precision is the fraction of actually positive instances among those that were predicted as positive by the classifier ($p = \frac{TP}{TP+FP}$) and recall refers to the fraction of actually positive instances that were retrieved by the classifier ($r = \frac{TP}{TP+FN}$).

The Receiver Operating Characteristic (ROC) Curve is a method to evaluate classifiers in which the true positive rate ($TPR = \frac{TP}{TP+FN}$) is plotted on the y axis and the false positive rate ($FPR = \frac{FP}{FP+TN}$) is plotted on the x axis. Some points of ROC curves have a well-defined interpretation: (FPR = 1, TPR = 0) means that all predictions are wrong, and (FPR = 0, TPR = 1) means that all positive and negative instances are correctly predicted. The case in which FPR = 0 and TPR = 1 is the ideal classifier, and the Area Under ROC Curve is 1. Thus, the closer AUC is to one, the better the model.

In the Descriptive Multiclass Experiment and the Predictive Multiclass Experiment, to select the best result for a specific classification algorithm, which is the matrix that led to this result, we applied a voting scheme. One vote was assigned for each result with the greatest value for F_1 and similarly one vote was assigned for each result with the greatest value for AUC. Note that more than one result may present the maximum value for F_1 or AUC. If there was a tie, we chose the result obtained from the matrix with the smallest number of columns.

Similarly, after choosing the best result within a specific classification algorithm, we selected the best result among all techniques through the same voting scheme. In this case, if there was a tie, we chose the result with the best F_1. When comparing the results obtained from the different classification algorithms, those with similar AUC values may have quite different F_1 values (hence, different precision and recall). Therefore, we prioritize the best values of F_1 when there was a tie in the voting scheme.

In the Predictive Common Source Experiment, we chose the best result according to the best value for F_1 because in this experiment even classifiers with high values for AUC showed low values for F_1 and therefore for precision and recall.

3.3.6 Implementation All graphs were generated with R (R Core Team, 2012) software version 2.10.1 and SVD dimensionality reduction. We implemented the data collection and processing in Java Development Kit 6 and performed the classification task using algorithms from Weka Data Mining Software (Hall *et al.*, 2009) version 3.6.2. The EC changes collected were stored in a MySql database, release 5.5.24.

4 RESULTS

4.1 Descriptive Multiclass Experiment

In this section, we present the results of the descriptive step. This experiment aimed to verify whether the annotation attributes of OC, RP and KW are able to discriminate entries that experienced a specific change in their EC number from those that remained the same. We generated classification models using data matrices reduced via SVD with k from 1 to 100 and chose the best classification model as explained in Section 3.3.5. We evaluated the model performance through a ten-fold cross-validation.

Table 2 provides the best result for this experiment. The complete results are provided in the Supplementary Material, Tables 4, 5, 6 and 7. Except for Naïve Bayes, the classifiers predicted the EC changes as their precision, recall and F_1 were approximately 70% and AUC was greater than 90%. We chose the KNN with 1 nearest neighbor as the best result due to its high F_1 values, which was considered by our voting scheme. It is important to highlight that, in general, modeled classes ($F_1 > 0.5$) have more examples than unmodeled ones, as presented in Table 8 from the Supplementary Material.

In Table 9 from the Supplementary Material, the arithmetic and weighted means were calculated separately for the classes that represent EC changes (change set) and non-changes (control set). In general, the values were worse for the change set than the control set, which was expected because predicting an annotation that changes is more difficult than predicting an annotation that remains constant because the data set has more examples from the control set than the change set.

This experiment provided evidence that the annotation attributes OC, RP and KW are able to discriminate and characterize entries that experienced a specific EC number change because even in a multiclass classifier with 664 classes (a complex classification problem as the probability of correctly predicting a class at random is 1/664 or 0.15%), the values of 0.74 for F_1 and 0.95 for AUC indicate that our classifier is far from random (when F_1 and AUC are approximately 0.5).

Table 2. Best results for the Descriptive Multiclass Experiment and Predictive Multiclass Experiment.

Multiclass experiment	Algorithm	# of feat.	FPR	Prec.	Rec.	F_1	AUC
Descriptive	KNN_K1	38	0.01	0.74	0.74	0.74	0.95
Predictive	KNN_K1	13	0.08	0.41	0.32	0.25	0.65

# of feat. refers to the number of features or attributes (in the matrix that resulted in the best classification model). TPR corresponds to recall and was omitted.

4.2 Predictive Experiments

The test data set was formed by the last occurrence of a certain type of EC change and the training data set comprised the previous occurrences of the same type of change. Consider the change $-.-.-.- \rightarrow 2.3.1.48$, which occurred in releases 2, 6, 8, 9, 12, 14, 15, 43, and 44. The entries that experienced such a change in releases 2, 6, 8, 9, 12, 14, 15, and 43 are part of the training data set, and the entries that experienced the same change in release 44 are part of the test data set.

Here, we simulated a scenario in which all available information about a certain type of EC change was applied to predict an upcoming EC change of the same type, which means that our approach can predict only changes previously observed in the database.

4.2.1 Multiclass The aim of the Predictive Multiclass Experiment was to make predictions for the last occurrence of each EC change type using a single multiclass classifier that comprises all possible classes. This experiment was performed similarly to the descriptive one, except for the EC change types, as here only those modeled in the Descriptive Multiclass Experiment were analyzed (361 classes).

The experimental results are provided in Table 2. The arithmetic and weighted means calculated separately for the change and control sets are shown in Table 12 from the Supplementary

Material (complete results are in Tables 10 and 11, also from the Supplementary Material). The values of precision, recall, F_1 and AUC were significantly lower than those in the Descriptive Experiment. When the last release in which a change occurred was left for the test set, some examples are lost for the training set, which impacted the result quality.

Therefore, to improve the results, we need more train examples or a more specialized classification task (with fewer classes than in the Predictive Multiclass Experiment). As we do not have control over the changes occurrence and amount, the changes were segmented by their common source, and a more specialized classification task was performed as detailed below.

4.2.2 Common Source This experiment was performed as an attempt to improve the classification results of the Predictive Multiclass Experiment shown in Section 4.2.1. The data set was segmented by the common EC source, and each source corresponds to a specific classifier. There are 24 common EC sources and thus 24 classifiers that are more specialized than the previous general multiclass, increasing the chance of making correct predictions (as there are fewer options of classes for each classifier). As explained in Section 3.3.5, 100 matrices resulting from the SVD were processed by three classification algorithms: Naïve Bayes, KNN with $K = \{1, 3, 5, 7, 10\}$ and J48. This process was performed for each of the 24 common source data sets, and the best results were chosen according to the best values for F_1.

The result of this experiment is provided in Table 3. The mean of the 24 best results was calculated to summarize the results ($TPR = 0.876$, $FPR = 0.257$, $Precision = 0.908$, $Recall = 0.876$, $F_1 = 0.864$, $AUC = 0.807$). The mean had values of precision, recall and F_1 greater than 0.86. However, there was one common origin, -.-.-.-, that had a significantly worse result compared with the mean. In addition, this origin had a high value of weight as it contains 36 types of changes and 2,631 instances of EC number changes.

In Table 13 from the Supplementary Material, the arithmetic and weighted means were calculated separately for the change and control sets. In the weighted mean from the change set, the precision (0.756) is greater than the recall (0.274), which is also known as the true positive rate (TPR) or specificity. This result indicates that predicting a change is difficult, but if the classifier predicts an instance as a change, it has a great chance to make a correct prediction.

It is important to highlight that although some values of these metrics appear low for the change set, the data considered to be the correct answer (UniProt/Swiss-Prot EC annotations) can present some inconsistencies or even errors as we observed that changes in the EC annotation occur over time in this repository. Furthermore, these metrics calculated from the Weka results do not take into consideration partial results (when not all predicted EC levels are correct). Thus, to provide a fair comparison between UniProt/Swiss-Prot and ENZYMAP, the predicted annotations were compared with the Swiss-Prot annotation considering from 1 to 4 levels of the EC number.

To extend this comparison, the DETECT tool (Hung *et al.*, 2010) was used to make the EC predictions for the same Swiss-Prot entries used in our approach. Thus, predictions from the ENZYMAP, DETECT and Swiss-Prot annotations were compared. DETECT was chosen because it is a relatively new technique (2010) that is able to predict the EC number annotations based on global and local sequence alignments. It receives FASTA residue sequences separated by organism as input and then outputs EC number predictions. Although ENZYMAP and DETECT are essentially different (as ENZYMAP is based on entry metadata from UniProt/Swiss-Prot and DETECT is based on residue sequence), their EC predictions can be used in a complementary manner to improve annotations.

Table 3. Results of the Common Source Experiment.

Source	FPR	Prec.	Rec.	F_1	AUC	Algorithm	# of feat.	# of clas.
-.-.-.-	0.10	0.66	0.34	0.31	0.66	KNN_K1	1	36
1.1.1.-	0.00	1.00	1.00	1.00	1.00	KNN_K1	11	2
1.10.2.2	0.00	1.00	1.00	1.00	1.00	KNN_K5	2	2
1.9.3.1	0.33	0.70	0.70	0.70	0.68	KNN_K10	2	2
2.-.-.-	0.31	0.77	0.42	0.32	0.62	N. Bayes	1	3
2.1.1.-	0.24	0.91	0.90	0.91	0.93	KNN_K7	74	3
2.3.1.-	0.96	0.93	0.96	0.95	0.91	KNN_K10	100	2
2.4.-.-	0.00	0.98	0.97	0.97	0.98	J48	13	2
2.7.1.-	0.03	0.93	0.88	0.89	0.89	KNN_K3	89	2
2.7.3.-	0.00	1.00	1.00	1.00	1.00	J48	30	2
2.7.7.48	0.30	0.70	0.66	0.66	0.55	KNN_K3	40	2
2.7.7.6	0.01	0.96	0.93	0.94	0.96	N. Bayes	32	2
3.-.-.-	0.01	0.95	0.90	0.91	0.94	KNN_K1	5	2
3.1.-.-	0.96	0.93	0.96	0.95	0.61	KNN_K1	100	2
3.1.13.-	0.06	0.95	0.95	0.95	0.91	KNN_K10	65	2
3.1.2.15	0.00	1.00	0.96	0.98	0.00	KNN_K10	100	2
3.2.1.18	0.93	0.87	0.93	0.90	0.50	J48	10	2
3.4.22.-	0.00	1.00	1.00	1.00	1.00	KNN_K10	100	2
3.4.25.-	0.33	1.00	1.00	1.00	0.97	KNN_K10	41	2
3.6.3.14	0.05	0.94	0.94	0.94	0.95	N. Bayes	12	2
4.2.2.-	0.64	0.80	0.72	0.62	0.80	KNN_K1	2	2
5.-.-.-	0.00	1.00	1.00	1.00	1.00	KNN_K1	4	2
6.-.-.-	0.90	0.81	0.90	0.85	0.50	N. Bayes	100	2
6.4.1.2	0.00	1.00	1.00	1.00	1.00	KNN_K1	10	2

Each line corresponds to the best result (classifier) obtained for each source as we used the training and test data from 1 up to 100 features after SVD processing and the classification techniques of Nave Bayes, J48 and KNN with $K = \{1, 3, 5, 7, 10\}$. The last two columns refers to the number of features or attributes (in the occurrence matrix that resulted in the best classification model) and to the number of classes in each classifier. The TPR corresponds to the recall and was omitted.

4.3 ENZYMAP, DETECT and Swiss-Prot comparison

The same input data set used for the predictive experiments in Section 4.2, with 3,582 EC number changes, was given as input for DETECT 1.0[4]. Our technique made 3,582 EC predictions, whereas DETECT made 1,876; both prediction sets were compared with the annotations from UniProt/Swiss-Prot. Figure 1 presents the comparison among the techniques.

For the first level shown in Figure 1 (a), 56% of the predictions made by ENZYMAP agree with UniProt/Swiss-Prot, whereas this rate is 49% for DETECT. If we consider the two approaches

[4] http://www.compsysbio.org/projects/DETECT/

together, their intersection with UniProt/Swiss-Prot represents 72% of these repository annotations, which shows that combining both of them increases the coverage of the annotations.

For levels 2, 3 and 4, the percentage of predictions made by ENZYMAP that are correct is greater than those made by DETECT, and both techniques together account for more than 64% of the database annotations as shown in Table 4. However, for level 4, the percentage of predictions made by DETECT that are correct decreases significantly and reaches 32%, whereas for ENZYMAP, the rate is 49%. Here, predictions that agree with the UniProt/Swiss-Prot are considered to be correct. The more specific the annotation, the more difficult it is to predict, which can lead to a common type of error called overprediction (when the annotation procedure assigns more levels than it should) (Schnoes *et al.*, 2009). Thus, in this aspect ENZYMAP outperforms DETECT.

Table 4. Correct predictions made by both techniques and the percentage of UniProt/Swiss-Prot they cover for the 4 levels of EC.

	Level 1	Level 2	Level 3	Level 4
ENZYMAP (%)	56	53	49	49
DETECT (%)	49	48	45	32
Coverage (%)	72	70	65	64

The rows ENZYMAP and DETECT correspond to the percentage of predictions made by our approach and by DETECT that are in accordance with the UniProt/SwissProt annotations. The coverage represents the percentage of repository annotations covered by the techniques used in a complementary manner.

4.3.1 Case studies In the common source 2.4.-.-, our technique predicted that entry Q5NDL2 should be annotated as 2.4.1.-. It was considered as an error by Weka because the Swiss-Prot annotation was 2.4.-.- in the test dataset. However, in release 2012_07 from July 2012 (released after our analysis), this entry received EC 2.4.1.255 in Swiss-Prot. We performed our prediction using the training data prior to release 2011_02 from February 2011 (inclusive) and the test data from release 2011_03 (March 2011), indicating that our technique anticipated the third EC level for entry Q5NDL2 16 months before it occurred in Swiss-Prot. DETECT did not return a result for this entry.

Entry Q5FWH2 was predicted to be 6.3.2.- for the test data from release 44, and this entry really experienced the change $-.-.-.- \rightarrow 6.3.2.-$ from release 43 to 44. In this case, our approach correctly predicted three EC levels starting from a non-annotated entry. DETECT did not return a result for Q5FWH2.

In the common source 3.1.2.15, the values of the metrics were excellent, but AUC was zero. In this case, there were two classes, $3.1.2.15 \rightarrow 3.1.2.15$ and $3.1.2.15 \rightarrow 3.4.19.12$. Predictions were made using the training data prior to release 2010_08 from July 2010 (inclusive) and the test data from release 2010_09 (August 2010). Among the 74 examples of change, 71 were correctly predicted. Nevertheless, in release 2010_09, there were no test examples in the control set ($3.1.2.15 \rightarrow 3.1.2.15$), which explains why Weka returned zero for AUC.

DETECT and ENZYMAP predicted that entries O61694 and O94581, subunits of Cytochrome c oxidase of an insect and a yeast, respectively, should receive EC number 1.9.3.1, which refers to oxidoreductases acting on heme groups as electron donors and oxygen as acceptors. In UniProt/Swiss-Prot, an EC number is not assigned to these entries, indicating that they are not enzymes. The point is that Cytochrome c oxidase is a large transmembrane protein complex, with several subunits, which may introduce some ambiguity. The prediction is correct if we consider them to be part of the Cytochrome c oxidase enzymatic complex. However, these subunits (per se) may have no direct catalytic function. This case illustrates the difficulty of composing an unbiased annotation when the entry comes from multi-domain or multi-chain protein complexes with different functional units. Indeed, until release 15 (March 2009), Swiss-Prot assigned EC number 1.9.3.1 to these entries.

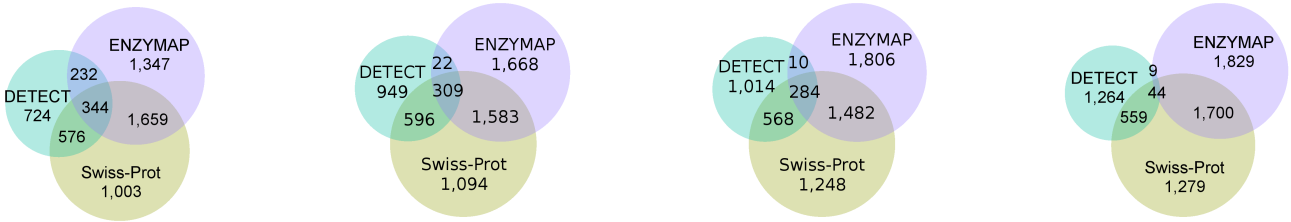
5 CONCLUSION

In this work, we proposed ENZYMAP, a technique to characterize and predict EC number changes in temporal data from UniProt/Swiss-Prot using a supervised learning approach. Our proposal is intended to be an automatic complementary method that helps improve the quality and reliability of enzyme annotations through the use of metadata that are already available in the repository, suggesting possible corrections and anticipating annotation changes.

To characterize and predict the EC number changes, we performed three experiments: *Descriptive Multiclass*, in which we concluded that the selected metadata were able to discriminate entries that experienced a specific change in the EC number from those that remained constant; *Predictive Multiclass*, which indicated that predicting the last occurrence of an EC change type using a multiclass classifier and having a scarce number of examples was not possible; and *Predictive Common Source*, which showed that predicting the last occurrence of an EC change type using more specialized classifiers even under the constraint of a scarce number of examples was possible. In addition, the predictions made by our proposal were compared with those made by the DETECT method, and both were checked against the Swiss-Prot annotations. The percentage of predictions made by ENZYMAP that were in accordance with Swiss-Prot was greater than the same percentage for DETECT for all 4 EC levels, and thus our technique outperformed DETECT in this aspect.

We envision a use case for ENZYMAP in which we employ the data available about EC changes in a set of UniProt/Swiss-Prot releases to predict upcoming changes, as performed in the Predictive Common Source Experiment. Thus, entries from the latest available release are given as input for a specific classifier trained with EC changes from all previous releases segmented by the common source, and this classifier returns EC predictions for each entry.

As future work, we intend to investigate whether it is possible to assign a reliability score to our predictions to help the user decide whether s/he should accept this prediction. In addition, we are considering using Formal Concept Analysis to elucidate for domain experts what were the most relevant metadata from Swiss-Prot to make the predictions, an information that is lost due to SVD use.



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

Fig. 1. We compared the EC number predictions made by ENZYMAP and DETECT and checked both against the UniProt/Swiss-Prot annotations. The number of predictions in which the techniques agree or disagree is presented in the diagrams. In (a), the first level of the EC number annotation is compared; In (b), (c) and (d), two, three and four levels of the EC number annotation are compared.

ACKNOWLEDGEMENTS

This work was supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and and Pró-Reitoria de Pesquisa da UFMG.

REFERENCES

Aha, D., Kibler, D., and Albert, M. (1991). Instance-based learning algorithms. *Machine learning*, **6**(1), 37–66.

Bécavin, C., Tchitchek, N., Mintsa-Eya, C., Lesne, A., and Benecke, A. (2011). Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. *Bioinformatics*, **27**(10), 1413–1421.

Berry, M., Dumais, S., and O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, pages 573–595.

Bowden, T., Aricescu, A., Gilbert, R., Grimes, J., Jones, E., and Stuart, D. (2008). Structural basis of nipah and hendra virus attachment to their cell-surface receptor ephrin-b2. *Nature structural & molecular biology*, **15**(6), 567–572.

Brenner, S. *et al.* (1999). Errors in genome annotation. *Trends in Genetics*, **15**(4), 132–133.

Deerwester, S., Dumais, S., Furnas, G., Harshman, R., Landauer, T., Lochbaum, K., and Streeter, L. (1989). Computer information retrieval using latent semantic structure. US Patent 4,839,853.

del Castillo-Negrete, D., Hirshman, S., Spong, D., and DAzevedo, E. (2007). Compression of magnetohydrodynamic simulation data using singular value decomposition. *Journal of Computational Physics*, **222**(1), 265–286.

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics*, **17**(8), 429–431.

Egelhofer, V., Schomburg, I., and Schomburg, D. (2010). Automatic assignment of ec numbers. *PLoS computational biology*, **6**(1), e1000661.

Eldén, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, **15**, 327–384.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, **27**(8), 861–874.

Furnham, N., Garavelli, J., Apweiler, R., and Thornton, J. (2009). Missing in action: enzyme functional annotations in biological databases. *Nature chemical biology*, **5**(8), 521–525.

Furnham, N., Sillitoe, I., Holliday, G., Cuff, A., Rahman, S., Laskowski, R., Orengo, C., and Thornton, J. (2012). Funtree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic acids research*, **40**(D1), D776–D782.

Green, M. and Karp, P. (2005). Genome annotation errors in pathway databases due to semantic ambiguity in partial ec numbers. *Nucleic acids research*, **33**(13), 4035–4039.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, **11**(1), 10–18.

Hung, S., Wasmuth, J., Sanford, C., and Parkinson, J. (2010). Detect - a density estimation tool for enzyme classification and its application to plasmodium falciparum. *Bioinformatics*, **26**(14), 1690–1698.

John, G. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, **40**(D1), D109–D114.

Lawrence, M., Borg, N., Streltsov, V., Pilling, P., Epa, V., Varghese, J., McKimm-Breschkin, J., and Colman, P. (2004). Structure of the haemagglutinin-neuraminidase from human parainfluenza virus type iii. *Journal of molecular biology*, **335**(5), 1343–1357.

Mewes, H., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K., *et al.* (2011). Mips: curated databases and comprehensive secondary data resources in 2010. *Nucleic acids research*, **39**(suppl 1), D220–D224.

Pang-Ning, T., Steinbach, M., and Kumar, V. (2006). *Introduction to data mining*. Pearson Education India.

Pires, D., de Melo-Minardi, R., dos Santos, M., da Silveira, C., Santoro, M., and Meira, W. (2011). Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, **12**(Suppl 4), S12.

Porter, M. *et al.* (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

Pruitt, K., Tatusova, T., Klimke, W., and Maglott, D. (2009). Ncbi reference sequences: current status, policy and new initiatives. *Nucleic acids research*, **37**(suppl 1), D32–D36.

Quester, S. and Schomburg, D. (2011). Enzymedetector: an integrated enzyme function prediction tool and database. *BMC bioinformatics*, **12**(1), 376.

Quinlan, J. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schnoes, A., Brown, S., Dodevski, I., and Babbitt, P. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, **5**(12), e1000605.

Silveira, S. A., Rodrigues, A. O., Melo-Minardi, R. C., da Silveira, C. H., and Meira Jr, W. (2012). Advise: Visualizing the dynamics of enzyme annotations in uniprot/swissprot. Biological Data Visualization (BioVis), 2012 IEEE Symposium.

Walter, M., Rattei, T., Arnold, R., Güldener, U., Münsterkötter, M., Nenova, K., Kastenmüller, G., Tischler, P., Wölling, A., Volz, A., *et al.* (2009). Pedant covers all complete refseq genomes. *Nucleic acids research*, **37**(suppl 1), D408–D411.

Webb, E. *et al.* (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Number Ed. 6. Academic Press.

Winsor, G., Lam, D., Fleming, L., Lo, R., Whiteside, M., Nancy, Y., Hancock, R., and Brinkman, F. (2011). Pseudomonas genome database: improved comparative analysis and population genomics capability for pseudomonas genomes. *Nucleic acids research*, **39**(suppl 1), D596–D600.

Referências Bibliográficas

Aha, D.; Kibler, D. e Albert, M. (1991). Instance-based learning algorithms. *Machine learning*, 6(1):37–66.

Ansorge, W. (2009). Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203.

Apweiler, R.; Bairoch, A. e Wu, C. (2004a). Protein sequence databases. *Current opinion in chemical biology*, 8(1):76–80.

Apweiler, R.; Bairoch, A.; Wu, C.; Barker, W.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M. et al. (2004b). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115.

Ashburner, M.; Ball, C.; Blake, J.; Botstein, D.; Butler, H.; Cherry, J.; Davis, A.; Dolinski, K.; Dwight, S.; Eppig, J. et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Ashburner, M.; Ball, C.; Blake, J.; Butler, H.; Cherry, J.; Corradi, J.; Dolinski, K.; Eppig, J.; Harris, M.; Hill, D. et al. (2001). Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–1433.

Bécavin, C.; Tchitchek, N.; Mintsa-Eya, C.; Lesne, A. e Benecke, A. (2011). Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. *Bioinformatics*, 27(10):1413–1421.

Becker, P. e Correia, J. (2005). The toscanaj suite for implementing conceptual information systems. *Formal Concept Analysis*, pp. 324–348.

Benson, D.; Karsch-Mizrachi, I.; Lipman, D.; Ostell, J. e Sayers, E. (2011). Genbank. *Nucleic acids research*, 39(suppl 1):D32–D37.

Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J. e Sayers, E. W. (2009). GenBank. *Nucleic acids research*, 37(Database issue):D26–31.

Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. e Bourne, P. (2000). The protein data bank. *Nucleic acids research*, 28(1):235.

Berry, M.; Dumais, S. e O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, pp. 573–595.

Binns, D.; Dimmer, E.; Huntley, R.; Barrell, D.; O'Donovan, C. e Apweiler, R. (2009). Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045.

Blake, J.; Bult, C.; Kadin, J.; Richardson, J. e Eppig, J. (2011). The mouse genome database (mgd): premier model organism resource for mammalian genomics and genetics. *Nucleic acids research*, 39(suppl 1):D842.

Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.; Estreicher, A.; Gasteiger, E.; Martin, M.; Michoud, K.; O'Donovan, C.; Phan, I. et al. (2003). The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365.

Bowden, T.; Aricescu, A.; Gilbert, R.; Grimes, J.; Jones, E. e Stuart, D. (2008). Structural basis of nipah and hendra virus attachment to their cell-surface receptor ephrin-b2. *Nature structural & molecular biology*, 15(6):567–572.

Brenner, S. et al. (1999). Errors in genome annotation. *Trends in Genetics*, 15:132–132.

Buneman, P.; Chapman, A. e Cheney, J. (2006). Provenance management in curated databases. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pp. 539–550. ACM.

Camon, E.; Barrell, D.; Dimmer, E.; Lee, V.; Magrane, M.; Maslen, J.; Binns, D. e Apweiler, R. (2005). An evaluation of go annotation retrieval for biocreative and goa. *BMC bioinformatics*, 6(Suppl 1):S17.

Cimiano, P.; Hotho, A. e Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24(1):305–339.

Codd, E. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387.

Commission, E. (1961). Report of the commission on enzymes. *IUB Symposium Series*, 20.

Consortium, U. (2011). Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.*, 39:214–219.

Consortium, U. et al. (2012). Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res*, 40:D71–D75.

Dall'Olio, G.; Bertranpetit, J. e Laayouni, H. (2010). The annotation and the usage of scientific databases could be improved with public issue tracker software. *Database: The Journal of Biological Databases and Curation*, 2010.

Deerwester, S.; Dumais, S.; Furnas, G.; Harshman, R.; Landauer, T.; Lochbaum, K. e Streeter, L. (1989). Computer information retrieval using latent semantic structure. US Patent 4,839,853.

del Castillo-Negrete, D.; Hirshman, S.; Spong, D. e D'Azevedo, E. (2007). Compression of magnetohydrodynamic simulation data using singular value decomposition. *Journal of Computational Physics*, 222(1):265–286.

Devos, D. e Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics*, 17(8):429–431.

Egelhofer, V.; Schomburg, I. e Schomburg, D. (2010). Automatic assignment of ec numbers. *PLoS Computational Biology*, 6(1):e1000661.

Eldén, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, 15:327–384.

Elmasri, R. e Navathe, S. (2008). *Fundamentals of database systems*, volume 2. Pearson Education India.

Engel, S.; Balakrishnan, R.; Binkley, G.; Christie, K.; Costanzo, M.; Dwight, S.; Fisk, D.; Hirschman, J.; Hitz, B.; Hong, E. et al. (2010). Saccharomyces genome database provides mutant phenotype data. *Nucleic acids research*, 38(suppl 1):D433.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fritz, M.; Leinonen, R.; Cochrane, G. e Birney, E. (2011). Efficient storage of high throughput dna sequencing data using reference-based compression. *Genome research*, 21(5):734–740.

Furnham, N.; Garavelli, J.; Apweiler, R. e Thornton, J. (2009). Missing in action: enzyme functional annotations in biological databases. *Nature chemical biology*, 5(8):521–525.

Furnham, N.; Sillitoe, I.; Holliday, G.; Cuff, A.; Rahman, S.; Laskowski, R.; Orengo, C. e Thornton, J. (2012). Funtree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic acids research*, 40(D1):D776–D782.

Garg, A. e Roth, D. (2003). Margin distribution and learning algorithms. In *Proc. of the International Conference on Machine Learning (ICML)*, pp. 210–217.

Gilks, W.; Audit, B.; De Angelis, D.; Tsoka, S. e Ouzounis, C. (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18(12):1641.

Gilks, W.; Audit, B.; de Angelis, D.; Tsoka, S. e Ouzounis, C. (2005). Percolation of annotation errors through hierarchically structured protein sequence databases. *Mathematical biosciences*, 193(2):223–234.

Google (2012). Google scholar. `http://scholar.google.com/`.

Green, M. e Karp, P. (2005). Genome annotation errors in pathway databases due to semantic ambiguity in partial ec numbers. *Nucleic acids research*, 33(13):4035.

Gruber, T. et al. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 43(5):907–928.

Han, J. e Kamber, M. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.

Harris, M.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C. et al. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(Database issue):D258.

Howe, D.; Costanzo, M.; Fey, P.; Gojobori, T.; Hannick, L.; Hide, W.; Hill, D.; Kania, R.; Schaeffer, M.; St Pierre, S. et al. (2008). Big data: The future of biocuration. *Nature*, 455(7209):47–50.

Hung, S.; Wasmuth, J.; Sanford, C. e Parkinson, J. (2010). Detect - a density estimation tool for enzyme classification and its application to plasmodium falciparum. *Bioinformatics*, 26(14):1690–1698.

John, G. e Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pp. 338–345. Morgan Kaufmann Publishers Inc.

Jones, C.; Brown, A. e Baumann, U. (2007). Estimating the annotation error rate of curated go database sequence annotations. *BMC bioinformatics*, 8(1):170.

Kanehisa, M. e Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27.

Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M. e Tanabe, M. (2012). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114.

Lawrence, M.; Borg, N.; Streltsov, V.; Pilling, P.; Epa, V.; Varghese, J.; McKimm-Breschkin, J. e Colman, P. (2004). Structure of the haemagglutinin-neuraminidase from human parainfluenza virus type iii. *Journal of molecular biology*, 335(5):1343–1357.

Lehninger, A.; Nelson, D. e Cox, M. (2008). *Lehninger principles of biochemistry*. Lehninger Principles of Biochemistry. W.H. Freeman.

Leinonen, R.; Akhtar, R.; Birney, E.; Bower, L.; Cerdeno-Tárraga, A.; Cheng, Y.; Cleland, I.; Faruque, N.; Goodgame, N.; Gibson, R. et al. (2011). The european nucleotide archive. *Nucleic acids research*, 39(suppl 1):D28.

Leinonen, R.; Diez, F.; Binns, D.; Fleischmann, W.; Lopez, R. e Apweiler, R. (2004). Uniprot archive. *Bioinformatics*, 20(17):3236.

Lesk, A. (2005). *Database annotation in molecular biology*. Wiley Online Library.

Lieber, J.; Napoli, A.; Szathmary, L. e Toussaint, Y. (2006). First elements on knowledge discovery guided by domain knowledge (kddk). In *Proceedings of the 4th international conference on Concept lattices and their applications*, pp. 22–41. Springer-Verlag.

Lindig, C. e Götzmann, D. (2007). Colibri-java–formal concept analysis implemented in java. Online: <http://code.google.com/p/colibri-java/> (acesso em 24.08.12).

Luscombe, N.; Greenbaum, D. e Gerstein, M. (2001). What is bioinformatics? a proposed definition and overview of the field. *Methods of information in medicine*, 40(4):346–358.

Ma, H.; Sorokin, A.; Mazein, A.; Selkov, A.; Selkov, E.; Demin, O. e Goryanin, I. (2007). The edinburgh human metabolic network reconstruction and its functional analysis. *Molecular systems biology*, 3(1).

Mewes, H.; Ruepp, A.; Theis, F.; Rattei, T.; Walter, M.; Frishman, D.; Suhre, K.; Spannagl, M.; Mayer, K.; Stümpflen, V. et al. (2011). Mips: curated databases and comprehensive secondary data resources in 2010. *Nucleic acids research*, 39(suppl 1):D220–D224.

Murzin, A.; Brenner, S.; Hubbard, T. e Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540.

Naumoff, D.; Xu, Y.; Glansdorff, N. e Labedan, B. (2004). Retrieving sequences of enzymes experimentally characterized but erroneously annotated: the case of the putrescine carbamoyltransferase. *BMC genomics*, 5(1):52.

NC-IUBMB (1999). Nomenclature committee of the international union of biochemistry and molecular biology (nc-iubmb), enzyme supplement 5 (1999). *Eur J Biochem*, 264(2):610–50.

Ogasawara, O.; Mashima, J.; Kodama, Y.; Kaminuma, E.; Nakamura, Y.; Okubo, K. e Takagi, T. (2012). Ddbj new system and service refactoring. *Nucleic Acids Research*.

Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M. e Thornton, J. (1997). Cath-a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109.

Otwinowski, Z. e Minor, W. (1997). [20] processing of x-ray diffraction data collected in oscillation mode. *Methods in enzymology*, 276:307–326.

Pegg, S.; Brown, S.; Ojha, S.; Seffernick, J.; Meng, E.; Morris, J.; Chang, P.; Huang, C.; Ferrin, T. e Babbitt, P. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry*, 45(8):2545.

Pires, D.; de Melo-Minardi, R.; dos Santos, M.; da Silveira, C.; Santoro, M. e Meira, W. (2011). Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12(Suppl 4):S12.

Porter, M. et al. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Pruitt, K.; Tatusova, T.; Klimke, W. e Maglott, D. (2009). Ncbi reference sequences: current status, policy and new initiatives. *Nucleic acids research*, 37(suppl 1):D32–D36.

PubMed (2012). Pubmed. `http://www.ncbi.nlm.nih.gov/pubmed/`.

Qin, Y.; Polacek, N.; Vesper, O.; Staub, E.; Einfeldt, E.; Wilson, D. e Nierhaus, K. (2006). The highly conserved lepa is a ribosomal elongation factor that back-translocates the ribosome. *Cell*, 127(4):721–733.

Quester, S. e Schomburg, D. (2011). Enzymedetector: an integrated enzyme function prediction tool and database. *BMC bioinformatics*, 12(1):376.

Quinlan, J. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

Rawlings, N.; Barrett, A. e Bateman, A. (2012). Merops: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, 40(D1):D343–D350.

Rost, B.; Yachdav, G. e Liu, J. (2004). The predictprotein server nucleic acids res,. 32. *W321 W*.

Scheer, M.; Grote, A.; Chang, A.; Schomburg, I.; Munaretto, C.; Rother, M.; Söhngen, C.; Stelzer, M.; Thiele, J. e Schomburg, D. (2011). Brenda, the enzyme information system in 2011. *Nucleic acids research*, 39(suppl 1):D670.

Schmidt, S.; Sunyaev, S.; Bork, P. e Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? *Trends in biochemical sciences*, 28(6):336–341.

Schnoes, A.; Brown, S.; Dodevski, I. e Babbitt, P. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605.

Silveira, S. A.; Rodrigues, A. O.; Melo-Minardi, R. C.; da Silveira, C. H. e Meira Jr, W. (2012). Advise: Visualizing the dynamics of enzyme annotations in uniprot/swissprot. Biological Data Visualization (BioVis), 2012 IEEE Symposium on.

Stein, L. (2003). Integrating biological databases. *Nature Reviews Genetics*, 4(5):337–345.

Suzek, B.; Huang, H.; McGarvey, P.; Mazumder, R. e Wu, C. (2007). Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282.

Tan, P.; Steinbach, M.; Kumar, V. et al. (2006). *Introduction to data mining*. Pearson Addison Wesley Boston.

Tang, E.; Suganthan, P. e Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, 65(1):247–271.

Tufte, E. (1990). *Envisioning information*.

Tweedie, S.; Ashburner, M.; Falls, K.; Leyland, P.; McQuilton, P.; Marygold, S.; Millburn, G.; Osumi-Sutherland, D.; Schroeder, A.; Seal, R. et al. (2009). Flybase: enhancing drosophila gene ontology annotations. *Nucleic Acids Research*, 37(suppl 1):D555.

Valtchev, P.; Grosser, D.; Roume, C. e Hacene, M. (2003). Galicia: an open platform for lattices. In *Using Conceptual Structures: Contributions to 11th Intl. Conference on Conceptual Structures (ICCS'03)*, pp. 241–254.

Walter, M.; Rattei, T.; Arnold, R.; Güldener, U.; Münsterkötter, M.; Nenova, K.; Kastenmüller, G.; Tischler, P.; Wölling, A.; Volz, A. et al. (2009). Pedant covers all complete refseq genomes. *Nucleic Acids Research*, 37(suppl 1):D408–D411.

Winsor, G.; Lam, D.; Fleming, L.; Lo, R.; Whiteside, M.; Nancy, Y.; Hancock, R. e Brinkman, F. (2011). Pseudomonas genome database: improved comparative analysis and population genomics capability for pseudomonas genomes. *Nucleic acids research*, 39(suppl 1):D596–D600.

Wu, C.; Yeh, L.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z.; Kourtesis, P.; Ledley, R.; Suzek, B. et al. (2003). The protein information resource. *Nucleic Acids Research*, 31(1):345.

Yevtushenko, S. (2003). Conexp. Online:<http://sourceforge.net/projects/conexp> (acesso em 24.08.12).

Zeeberg, B.; Feng, W.; Wang, G.; Wang, M.; Fojo, A.; Sunshine, M.; Narasimhan, S.; Kane, D.; Reinhold, W.; Lababidi, S. et al. (2003). Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28.