DEIVE CIRO DE OLIVEIRA

ESCORE DE INCERTEZA EM BANCOS DE DADOS CATEGÓRICOS

Belo Horizonte
01 de dezembro de 2011

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

ESCORE DE INCERTEZA EM BANCOS DE DADOS CATEGÓRICOS

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

DEIVE CIRO DE OLIVEIRA

Belo Horizonte
01 de dezembro de 2011

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escore de Incerteza em Bancos de Dados Categóricos

DEIVE CIRO DE OLIVEIRA

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

Ph. D. Marcelo Matos Santoro – Co-orientador
Universidade Federal de Minas Gerais

Belo Horizonte, 01 de dezembro de 2011

Resumo

Temos observado um grande crescimento no volume de dados biológicos, em particular de natureza biomolecular, armazenados em bancos de dados como Genbank, KOGG SCOP, PDB e Uniprot, os quais são acessados livremente através da internet e tem tido um impacto enorme nas atividades de pesquisa e desenvolvimento. Esse crescimento é explicado pelo desenvolvimento de novas e menos dispendiosas técnicas de obtenção daqueles dados, assim como menor custo e maior disponibilidade de meios de armazenamento e comunicação. Uma distinção importante entre esses bancos é com relação à forma de geração e manutenção da base de dados. Alguns bancos têm seus dados gerados a partir de métodos computacionais (*in silico*) e não são submetidos a processos de revisão. Outros bancos, denominados curados, adotam supervisão especializada nos processos de geração e revisão de características, a qual pode ser feita pelos usuários que acessam esses bancos através da internet. Os bancos de dados curados alcançam alto padrão de qualidade em termos de anotação mas possuem um alto custo comparado a processos automatizados. Neste contexto, metodologias e técnicas para auxiliar no processo de revisão são relevantes, pois tornam a curagem mais eficiente e reduzem o custo de realizá-la. Este trabalho tem por objetivo investigar, desenvolver e avaliar tais metodologias e técnicas e apresenta duas contribuições principais. A primeira é uma metodologia para caracterizar temporalmente modificações em um banco de dados categórico. Essa metodologia é aplicada ao UniprotKB/Swiss-prot, quantificando as taxas de modificações específicas em palavras-chave do UniprotKB/Swiss-prot. Também são apresentadas as modificações nas associações existentes entre as palavras-chave, sob perspectiva temporal. A segunda contribuição é uma metodologia para auxílio no processo de revisão em bancos de dados categóricos. Um exemplo de aplicação é a revisão do campo palavras-chave do banco de dados UniprotKB/Swiss-prot, onde pudemos observar que a metodologia proposta é efetiva.

Abstract

We have been witnessing a significant growth in the volume of biological data, in particular biomolecular data that are stored in databases such as Genbank, KOGG SCOP, PDB, and Uniprot, which are made available through the internet and have been causing a major impact in research and development activities. Such growth is explained by the development of novel and less costly data gathering techniques, as well as, lower costs and higher availability of storage and communication resources. A key feature that distinguishes those databases is regarding the procedure to generate and to maintain those databases. Several databases are created using automated procedures (in silico) and the resulting data is not curated by an expert. Other databases, named curated, employ specialized supervision for both generation and revision of characteristics, which may be performed by the users that access the databases through the internet. The curated databases present a much higher quality with respect to annotations, but are very costly when compared to automatic processes. In this scenario, research on novel methodologies and techniques that help on the revision process are relevant, since they make it more efficient and less costly. This work aims to investigate, develop, and evaluate these methodologies and techniques and has two main contributions. The first is a methodology for temporally characterizing the modifications in a categorical database. This methodology is applied to the UniprotKB/Swiss-prot, and quantified the record changes in keywords from this database. We also characterize the modifications on the keywork associations, under a temporal perspective. The second contribution is a methodology for improving the revision process. An example of application scenario is the revision of the field keywords from the UniprotKB/Swiss-prot database, where we can clearly see that proposed methodology is effective.

".....De perto ninguém é normal."

Caetano Veloso

Agradecimentos

Agradeço a meu pai Adelson, minha mãe Maria, minha irmã Dili e a todos meus familiares pelo apoio incondicional nos momentos positivos e negativos.

Agradeço a minha esposa Carolline pelo amor e sobretudo pela tolerância num período que a ausência foi maior do que a presença.

Agradeço aos orientadores Wagner Meira e Marcelo Santoro pela serenidade nos momentos tensos, pela severidade nos momentos necessários e sobretudo pela confiança durante todo o período do curso.

Agradeço aos colegas Bernardo "Pardal", Wagner Magalhães, Ricardo Andrez, Crysttian Arantes, Carlos Henrique da Silveira, Raquel Minardi, Douglas Pires, Valdete Almeida, Priscila Grynberg. Todos participaram e colaboraram com este trabalho assiduamente através de revisões, discussões, consultas.

Agradeço aos colegas da UNIFAL-MG(Varginha) pela amizade e pelo constante diálogo. Saibam que a Bioinformática é multidisciplinar por natureza. O desenvolvimento da multidisciplinaridade acentuei no doutoramento, mas exercitei também em Varginha.

Agradeço aos pesquisadores/docentes/colaboradores do programa e que participaram de minha formação acadêmica. De distintas maneiras voces ajudaram na execução deste trabalho.

Agradeço a FAPEMIG pelo apoio financeiro parcial do trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Desde o final da década de 80, o crescimento dos conjuntos de dados biológicos disponíveis à comunidade científica tem características exponenciais. Pode-se citar a internet como fator importante no aumento deste acesso. Além disso o aumento da capacidade de geração destes dados também influiu neste crescimento. Isto se torna mais acentuado no contexto de dados biomoleculares, onde a obtenção (geração de dados via tecnologias experimentais) tornou-se um processo de larga escala. Este crescimento se deve à redução de custos de tecnologias e pelo uso de novas metodologias mais eficientes. Frente a esta grande massa de dados, novos desafios surgem. Entre eles podem ser citados a necessidade de integração de distintas fontes de informação assim como a garantia de integridade e qualidade. A existência de milhares de bases de dados que tratam de entidades relacionadas apresenta o desafio de integração. Como estabelecer o relacionamento entre estas entidades em repositórios de dados distintos? Não só a integração é necessária, mas também a garantia de integridade dos dados. A integridade constitui um conjunto de regras sintáticas de utilização do banco de forma a evitar erros. Mesmo assim adotar políticas de integridade não garantem a eliminação da totalidade de erros no banco. Neste sentido o desafio abrange a localização destes erros incrementando a qualidade existente no banco.

O termo qualidade, quando tratamos de dados, é bastante amplo. A qualidade das bases de dados tem sido uma preocupação emergente na literatura em várias áreas de aplicação, inclusive no campo biológico. Neste trabalho a qualidade é relacionada com incerteza. Definimos incerteza como a garantia de correção de um dado. Uma série de aspectos podem influenciar na incerteza sobre um dado. Dois podem ser listados: a forma de obtenção e a revisão. Em dados de natureza biomolecular a incerteza é inerente aos seus processos de obtenção ou geração. Independente do método aplicado, todos estão sujeitos a taxas de erros que resultam em incerteza. Essas taxas são baixas (Harismendy et al., 2009), porém com uma quantidade grande de dados, realidade atual, o volume de erros é maior em termos absolutos. Ademais existe também o aspecto de revisão. Frente ao desenvolvimento de novos conhecimentos, dados obtidos via metodologias menos eficientes (mais antigas) podem ter a necessidade de serem reavaliados o que implica na aplicação

de processos de revisão.

Em termos da geração, também citado como anotação no contexto biológico, os conjuntos de dados podem ser divididos em: bancos de dados de anotação automática e de anotação manual. Os primeiros são gerados por processos computacionais de forma automatizada. Embora sejam metodologias de menor custo, em virtude do aumento da escala (tratamento de um grande conjunto de dados rapidamente), elas apresentam um nível de incerteza maior quando comparadas com bancos de anotação manual. Alguns exemplos de bancos cujo conteúdo é gerado automaticamente são UniprotKB/TREMBL e Ensembl. Bases com anotação manual são geradas com a supervisão integral e/ou parcial de especialistas. Eles são conhecidos como bancos de dados curados. O custo da supervisão especializada é alto se comparado a bancos de dados de anotação automática, porém os resultados são melhores. Reflexo disto está na utilização de bancos curados para ajuste, treinamento, e validação de processos de geração automática. São exemplos de bancos curados o Structural Classification of Proteins (SCOP) e o UniprotKB/Swiss-prot.

Processos de revisão podem ser aplicados em bancos de dados. Estes processos, mediante uma avaliação, procuram atualizar o conteúdo da base. Esta revisão pode se dar pelo incremento de informação e/ou correção ou adequação de um determinado conteúdo. Exemplos de revisão são a reanotação de genomas, a revisão de anotações em dados de sequências protéicas, a correção ou adição de modelos variantes em bancos de estruturas biomoleculares. Tais revisões em alguns bancos podem se dar sob demanda (p. ex.: constatação de um erro via administrador e/ou usuário) ou de forma sistemática (p. ex.: inspeção periódica dos bancos). Tanto quanto o processo de anotação, a revisão manual é onerosa. Em bancos de dados de acesso público, a revisão sob demanda é facilitada dada a interação dos muitos usuários. O UniprotKB/Swiss-prot é um banco de dados público curado que implementa a revisão periódica dos registros.

A incerteza sobre os dados armazenados gera demandas. Dentre as que permeiam este trabalho, podem ser citadas:

1. **Como caracterizar e quantificar informações espúrias existentes na base?** Inferir taxas de anotações incorretas é fundamental para avaliar a qualidade de uma base de dados. Espera-se que bases de alta qualidade tenham frequências baixas de erros. Questões associadas passam pela caracterização do que é um erro em um banco de dados e de como avaliar sua frequência. Vamos destacá-las como questões de inferência.

2. **Como identificar os registros que precisam ser revisados?** Avaliar a frequência de erros em um conjunto de dados passa pela detecção da incorreção no banco. Mas como antecipar a detecção de erros? Como estabelecer quais registros possuem a maior chance de apresentarem erros e portanto prioritários em políticas de revisão? Estas questões serão relacionadas como de manutenção do conjunto de dados.

3. **Como agregar a informação de qualidade dos dados na aplicação de métodos de extração de conhecimento?** Uma vez que o conjunto de dados possui menor ou maior incerteza associada, como aplicar métodos de extração de conhecimento que considerem esta incerteza e a agreguem em seus resultados? Estas questões estão relacionadas com a extração de informação sob incerteza.

A literatura tenta sanar estas questões segundo diferentes especificidades. Algumas são:

- **Qualidade em Bancos de Dados Biológicos (BDBs): Demanda 1 - Inferência**

  Uma vez definido o que é um erro, o problema de inferência sobre sua taxa em um conjunto de dados pode ser tratado de duas formas: métodos *a priori* ou *a posteriori* geração dos dados. Métodos *a priori* consideram taxas de erros antes do processo de geração do dado. Podem ser citados testes de hipóteses quando define-se um nível de significância, métodos de classificação e agrupamento quando afere-se via validação sua capacidade e eficiência. Métodos *a posteriori* tratam de como estimar a taxa de modificações em um conjunto de dados após seu processo de geração. Para tanto, uma vertente é, a partir de metodologias indiretas, fazer uma estimação sobre a taxa de erros de anotação. Outra estratégia mais simples, porém de maior custo, é a estimação via observação da frequência de ocorrência das modificações.

- **Qualidade em Bancos de Dados Biológicos (BDBs): Demanda 2 - Manutenção**

  Conhecer as taxas globais de erros é importante, porém não é muito informativo para fins de manutenção dos dados. A detecção de erros ou atualizações exige uma análise local de cada registro. Ou seja, a propensão de um determinado registro sofrer alterações. Alguns métodos trabalham no caso de erros sintáticos, porém os de difícil detecção são os de característica semântica. Visto que alguns bancos têm revisão periódica e não são públicos, a inexistência de critérios que agilizem a descoberta de erros ou a necessidade de atualizações inviabiliza a manutenção destas bases. A proposição de critérios automáticos para revisão é o objetivo principal deste trabalho.

- **Qualidade em Bancos de Dados Biológicos (BDBs): Demanda 3 - Extração de Informação**

  Assumindo conhecimento sobre a incerteza de um dado em seu nível global ou local, como utilizar estas informações na aplicação de métodos de descoberta e extração de informação? Os algoritmos clássicos de mineração trabalham com pré-suposições de determinismo dos dados. Em um contexto com incerteza esta pré-suposição é inválida. Uma área emergente de pesquisa é a mineração de dados sob

incerteza. Ela estende os métodos determinísticos para dados com informação prévia de incerteza. Outra área conexa é a mineração de incerteza sobre os padrões emersos sobre os dados. Avaliar como os padrões (associações, agrupamentos, classificações) variam temporalmente é extremamente importante para quantificar e identificar modificações generalizadas em um conjunto de dados.

Sejam a inferência, a manutenção, a extração de informação em cenário de incerteza, quase todas estas demandas necessitam de subsídio de informações sobre modificações ou erros. Cita-se a exceção de procedimentos de inferência de erros pré-anotação dos dados (Testes de hipóteses). Como investigar a qualidade dos dados com disponibilidade parcial ou nula destas informações? Algumas iniciativas que tentam suprir esta ausência serão tratadas no capítulo seguinte. Grande parte destas iniciativas depende das modificações do banco de dados ao longo do tempo.

Alguns bancos de dados biológicos já estão atingindo 2, 3 decádas desde sua criação. Muitos destes bancos sofreram modificações em termos de conteúdo, acabando por gerar diferentes versões ao longo do tempo. Alguns deles já se preocupam em disponibilizar aos usuários os históricos de modificações das entidades armazenadas. Esta disponibilidade se dá através de dados brutos ou dados já estruturados, considerando o tempo. A existência destes dados possibilita a proposição e/ou validação de metodologias para ajudar na tarefa de revisão. Além disso, é possivel reportar sobre mudanças dos padrões existentes no banco ao longo do tempo.

Seja em bancos de dados curados ou automáticos, os processos de anotação e revisão sofrem modificações ao longo do tempo. Ferramentas de auxilio à anotação e a maior disponibilidade de dados experimentais são fatores que contribuem para esta evolução. **Um problema é como caracterizar, quantificar e identificar momentos significativos destas mudanças?** Avaliar consistências, mudanças ao longo do tempo e identificar pontos de mudanças generalizadas informam sobre a necessidade de adequação da base de dados e dos métodos pós-geração dentre outros.

Considere como exemplo, a existência de uma significativa mudança de padrões em um banco de dados. Uma vez que um método utilize o banco de dados para validação em um momento prévio às mudanças generalizadas nos padrões, seus resultados não estarão validados sob a estrutura de padrões atuais. Considere uma extração de padrões realizada antes da modificação. O resultado da extração não é válido com o banco de dados atual. Uma forma muito utilizada no contexto biológico para representação de conhecimento (padrões) são as ontologias. Ontologias são redes de relações entre atores ou entidades. Vários esforços são dedicados à integração de diferentes ontologias, mas só é possível tratar de integração, assumindo que estas ontologias são consistentes durante o tempo, ou identificar períodos de consistência. Outro exemplo pode ser visto na aplicação de algoritmos de classificação e agrupamento. Pode ser necessário revalidar o método construído a partir da base alterada para se obter seus verdadeiros indicadores de eficiência

coerentes com a base atual. Contribuições neste sentido são válidas sob a dimensionalidade metodológica (método de estudo de mudanças de padrões no banco) e de resultados (subsídio para futuras decisões relacionadas ao banco de dados analizado)

O banco de dados UniprotKB/Swiss-prot disponibiliza versões antigas, o que possibilita seu estudo sob perspectiva temporal. Em termos de conteúdo ele é um banco curado e armazena dados de anotação protéica. A versão mais recente já conta com mais de meio milhão de proteínas anotadas. Estudos sobre as mudanças do UniprotKB/Swiss-prot podem considerar todos os atributos do banco. Uma outra estratégia, mais parcimoniosa, é avaliar somente um grupo de atributos ou anotações representativas de todo um registro (ou entrada). Esta decisão pode ter implicações práticas na redução de custo computacional para se realizar a análise. Dependendo da técnica de análise empregada e o conjunto de dados, pode ser computacionalmente inviável tratar o problema de estudo da evolução do banco. Dentre as anotações do UniprotKB/Swiss-prot, uma em particular trata de sumarizar a proteína. Esta anotação é o campo *k*eywords (palavras-chave). O aspecto de resumo faz com que as palavras-chaves sejam um campo representativo dentre os existentes em um registro UniprotKB/Swiss-prot. Elas atuam ainda como índices, tornando-se informação relevante na realização de uma consulta, na representação de características importantes das proteínas, etc. Tais propriedade tornam-as excelentes atributos candidatos sob os quais investigar a evolução do UniprotKB/Swiss-prot.

Este trabalho tem como objeto de estudo as palavras-chave das proteínas do UniprotKB/Swiss-prot. Seu objetivo é a construção de uma metodologia para incremento de qualidade dos dados de palavras-chave UniprotKB/Swiss-prot através de estabelecimento de prioridades no processo de revisão. Esta metodologia, subsidiada por modificações passadas, realiza a previsão sobre modificações futuras no banco. O dois principais resultados são: estudos sobre mudanças temporais no UniprotKB/Swiss-prot e a proposição de uma metodologia para auxílio na revisão, ambos restritos ao campo palavras-chave. O capítulo 2 apresenta a revisão de literatura sobre inferência de incerteza em banco de dados biológicos assim como na melhoria e auxílio de anotação e revisão. O capítulo 3 explicita os objetivos e a 4 concentra as metodologias utilizadas no trabalho. Os dois capítulos seguintes apresentam os estudos sobre mudanças nas palavras-chave do UniprotKB/Swiss-prot. Em 5 trata sobre mudanças na revisão (correção e atualizações) e a 6 sobre mudanças nos relacionamentos entre palavras-chave. O capítulo 7 é apresentado o escore de prioridade de revisão para bancos de dados categóricos. Como fechamento, na seção 8 são apresentadas as considerações finais.

Capítulo 2

Fundamentos

2.1 Bancos de Dados Biológicos - (BDB)

Nas últimas três décadas, sob diversos aspectos, houve um incremento de informação biológica disponível à comunidade científica. Este crescimento tem diferentes características (de linear a exponencial) seja em acessibilidade e quantidade. A popularização e maior usabilidade de sistemas gerenciadores de bancos de dados foram fatores que fomentaram o aumento do número de conjuntos de dados publicados. O número de usuários de internet que cresceu ao longo destas décadas, sobretudo após 1990, também foi um fator importante na disseminação destes bancos de dados biológicos. Embora importantes, a estruturação e acessibilidade acabam por ser necessidades quando tratamos grandes conjuntos de dados. Neste sentido, o avanço nos processos de geração dos dados biológicos acaba por ser determinante neste crescimento. A maior escala na geração de dados foi devida tanto à maior eficiência das tecnologias (p. ex.: baixo custo de hardware (Hennessy e Patterson, 2003), sequenciadores de nova geração (Kato, 2009)) quanto dos novos métodos propostos na literatura (p. ex.: algoritmos de menor custo computacional (Khreisat, 2007), especialização de métodos estatísticos (Cogger, 2010)).

A biologia molecular tem sido a área que demanda estruturação e mecanismos de acesso, uma vez que suas massas de dados tem crescido exponencialmente. Dados genômicos, mais simples, quando comparados a outros dados biológicos (ex: proteômica, interatômica), já se aproximam da escala de giga (10^9) em número de entidades armezenadas em seus bancos (Genomenet, 2011). Dentre os conteúdos, um recorte trata de proteínas.

Proteínas são complexos biológicos responsáveis por diferentes tipos de atividades metabólicas (Berg et al., 2008). Em organismos, sua síntese se dá a partir da informação genética do indivíduo (ácido nucleico - DNA ou RNA). Um ácido nucleico é um polímero de nucleotídeos que se diferenciam segundo as bases nitrogenadas que o compoem. A sequência de bases define o ácido. Por sua vez, uma codificação do ácido nucleico, após processos celulares, sintetiza um polipeptídeo. Uma cadeia polipeptídeca é formada por ligações covalentes entre aminoácidos. Após a formação, interações mais fracas entre

os átomos que compõem a cadeia polipeptídeca vão definir a estrutura tridimensional da molécula. As proteínas são compostos com um ou mais polipeptídeos podendo estar complexados com compostos inorgânicos. Em termos de atuação, elas são responsáveis por uma série de processos biológicos. A especificidade de atuação vai depender do ambiente e da conformação estrutural da molécula. Dada sua importância, um conjunto de bancos de dados dedica-se a armazenar entidades proteicas. Estes bancos diferenciam-se segundo conteúdo (neste caso específico, características da proteína armazenada). Alguns focam em informações estruturais (p. ex.: PDB (Bernstein et al., 1977)), outros centram em informações taxonômicas (p. ex.: SCOP (Hubbard et al., 1997)) e em relações entre proteínas (p. ex.:KEGG (Kanehisa et al., 2010)). É importante destacar que vários destes bancos hoje têm seus conteúdos integrados.

O Uniprot Knowledgebase (UniprotKB) (Consortium, 2011) é uma coleção de bancos de dados relacionados a proteínas. O registro de uma proteína é denominado entrada (*entry*). Entre as características principais armazenadas estão dados referentes à sequência de aminoácidos, nome da proteína e sua descrição, informação taxonômica e de citação na literatura. Os registros também armazenam dados sobre ontologias biológicas relacionadas, classificações da proteína armazenada e referências a outros bancos. Além disso, o UniprotKB armazena informações sobre a qualidade de sua anotação. Isto se dá pela retenção de dados sobre evidência da característica anotada (Informação *apriori* sobre a qualidade do dado). Em geral estas evidências são separadas segundo resultados experimentais ou computacionais.

A quase totalidade das sequências proteicas presentes no UniprotKB são derivadas de bancos públicos de ácido nucléicos como: EMBL-Bank (banco de sequência nucleotídica do Laboratório Europeu de Biologia Molecular) (Kulikova et al., 2007), GenBank (banco de sequência genética do NIH-Instituto Nacional de Saúde dos Estados Unidos)(Benson et al., 2009), DDBJ (Bancos de Dados japonês de DNA). Esta derivação se dá através da tradução dos códigos genéticos em sequências de proteínas hipotéticas. Uma vez traduzidas, elas são integradas ao UniprotKB.

O UniprotKB é composto por dois bancos com esquemas de dados iguais, porém diferindo segundo o processo de anotação e revisão. O primeiro é o UniprotKB/TrEMBL. Ele tem suas anotações e caracterizações funcionais geradas a partir de processos computacionais de larga escala. As entradas do UniprotKB/TrEMBL não passam por revisão. Uma vez integradas ao TrEMBL, aguardam por anotação e posterior integração a um segundo banco chamado UniprotKB/Swiss-prot.

O Swiss-prot é um banco de anotações de sequências de proteínas criado em 1986. A partir de 2002 ele foi integrado ao UniprotKB. O UniprotKB/Swiss-prot é alimentado com dados do TrEMBL. Uma vez inserida no Swiss-prot, a entrada é excluída do TrEMBL. Para integrar o UniprotKB/Swiss-prot, a entrada passa por um processo de anotação manual realizada por uma equipe especializada em curadoria. Esta anotação é feita por meio de buscas de referências às sequências na literatura além da utilização de ferramen-

tas computacionais. As fontes sobre a sequência a ser anotada podem identificá-la com ou sem caracterização bioquímica. Dependendo da referência, a entrada pode ser caracterizada também segundo sua qualidade. As caracterizações são *Potential*, *Probable* e *by similarity*. *Potential* indica evidência da anotação. *Probable* refere-se a anotações com ao menos um resultado experimental comprovado. Portanto evidência mais forte que a caracterização anterior. *by Similarity* refere-se à propagação de anotação por similaridade de sequências. No Swiss-prot, não só o processo de anotação é manual, mas também uma revisão periódica realizada em todas as entradas do banco.

Embora o número de entradas cresca exponencialmente nos dois bancos, o TrEMBL cresce numa escala cerca de 30 vezes maior que o Swiss-prot. Isto é reflexo da curagem manual e revisão realizada no Swiss-prot que é muito mais onerosa. O UniprotKB/Swiss-prot chega a um total de 526969 de entradas enquanto o TrEMBL já atinge 14555721 (dados referentes à versão de abril de 2011). No entanto, devido à curadoria manual, o Swiss-prot é considerado padrão ouro em termos de qualidade de anotação.

A alta similaridade entre sequências de proteína leva a conformações estruturais similares e consequentemente atividades metabólicas semelhantes. Uma vez que a informação central do UniprotKB é a sequência, a não consideração da similaridade poderia levar a um banco altamente redundante. Buscando minimizar a redundância, o UniprotKB/Swiss-prot aglomera várias proteínas com alta similaridade em uma única entrada. Embora todas as proteínas tenham um identificador no banco, elas são associadas a um identificador primário. Este identificador não se modifica ao longo do tempo. Na entrada são listadas variações como *splices* alternativos, polimorfismos ou conflitos entre as proteínas. Com o passar do tempo as entradas podem ser mescladas ou divididas com o objetivo de reduzir a redundância.

As revisões periódicas realizadas no UniprotKB/Swiss-prot muitas vezes modificam o banco. Estas modificações podem ser complementos (adição de uma nova publicação, uma nova variante da proteína armazenada, dentre outros), bem como a correção de anotações (adequação, substituição ou exclusão de uma anotação). Com isto, o banco é modificado ao longo do tempo não só em estrutura, como é comum, mas também em termos de conteúdo. Desta forma, a mesma entrada pode ter versões distintas ao longo do tempo. O UniprotKB/Swiss-prot, embora seja atualizado periodicamente, mantém acessíveis versões desde sua criação em 1986 até a mais atual de 2011 (UniprotKB, 2011a). Além de disponibilizar estas versões antigas, o UniprotKB criou um banco que considera as diferentes versões da entrada ao longo do tempo. O Unisave (Leinonen et al., 2006), como é chamado, oferece ao usuário a possibilidade de consulta e de comparação, por entradas, de todas as versões desde sua criação. Tendo acesso ao histórico do UniprotKB, é possível verificar as taxas de modificações entre as versões. A disponibilidade deste tipo de dado permite inferir sobre a qualidade da anotação realizada neste banco.

2.2 Qualidade em Bancos de Dados

O estudo sobre a qualidade dos dados é uma área emergente com questões em aberto (Lee et al., 2009). A definição de qualidade de dados exige conhecimentos multidisciplinares assim como abrange várias dimensões. Podem ser citadas: duplicação, inconsistências, dados perdidos, desatualização, anomalias, incorreções dentre outros. A aplicação de métodos de mineração para detectar, quantificar, explicar e corrigir deficiências na qualidade de dados é objetivo central de uma área de pesquisa denominada *Data Mining Quality* (Hipp et al., 2001; Guillet e Hamilton, 2007). Este trabalho trata da investigação de qualidade de dados considerando excepcionalidade e incerteza. A seguir serão discutidos estes conceitos.

Segundo (Ch et al., 2007), exceções, também referidas como *outliers*, são padrões nos dados que não se adequam a uma característica esperada. Considerando um modelo probabilístico (Ross, 2006), exceções são dados com baixa verossimilhança, pouco frequente ou raros. Seguindo este definição e assumindo como exemplo um modelo normal, em que X segue uma distribuição normal com média μ e variância σ^2 ($N(\mu, \sigma^2)$), dados da cauda da distribuição são exceções. Ou seja, para todo $|X| < c$ onde $c > 0$ é um limite, X é considerado uma exceção. Neste exemplo a exceção é definida de forma determinísta. Entretanto esta definição pode ser relativizada. No caso deste exemplo, a proximidade de X em relação à μ aumenta sua verossimilhança. Assim, pode-se afirmar que um dado é excepcional ou não, mas também avaliar seu grau de excepcionalidade.

O tema da detecção de exceções é muito estudado na literatura (Chandola et al., 2009; Ch et al., 2007; Hodge e Austin, 2004), mas poucos trabalhos dedicam-se à sua aplicação a dados categóricos quando comparados a dados quantitativos (He et al., 2005). Muitas técnicas podem ser adaptadas do espaço quantitativo para o qualitativo, porém existem restrições. Métodos baseados em agrupamento são algumas das adaptações. Estes se tornam relativos, dada a dependência da função de similaridade entre as entidades. Modificações na forma de se aferir estas similaridade podem alterar significativamente o resultado. Os métodos baseados em classificação dependem da disponibilidade de dados de treinamento e validação. Considere casos de classificação em dados quantitativos. A localização no hiper espaço sem a informação sobre a classe não informa sobre a propensão da entidade ser uma exceção. As abordagens estatísticas são aplicáveis a dados categóricos (distribuição multinomial), embora, na prática, os modelos paramétricos mais simples possam não capturar a distribuição dos dados.

Um grupo de técnicas que trata exceções em dados categóricos é baseado na construção de regras. Assumindo regras que representam padrões no conjunto de dados, a exceção é toda instância que não obedece a uma regra. Existem vários algoritmos para extração de regras de associação. Os mais utilizados são baseados no princípio Apriori (Agrawal e Srikant, 1994). A busca de exceções via regras tem como desvantagem a possibilidade de não identificá-las sob determinadas parametrizações. Entretanto são desnecessárias pré-

vias identificações das exceções (rotulação) e das etapas de treinamento. Alguns trabalhos aplicam esta metodologia a detecção de invasão em sistemas (Barbara et al., 2001; Otey et al., 2003). Na mesma linha, (Mahoney e Chan, 2003) mescla a mineração de regras com amostragem probabilística para identificar exceções. Esta abordagem é aplicada em detecção de fraudes (Brause et al., 1999; Yairi et al., 2001). Ainda podemos citar (Narita e Kitagawa, 2008; He et al., 2002) que propõem critérios de excepcionalidade construídos a partir de frequências dos registros (*itemsets*) e em regras de associação geradas em um conjunto de dados. A pré-suposição básica deste trabalho é admitir que entidades excepcionais, especificamente em dados categóricos, terão maiores níveis de incerteza.

O conceito de incerteza pode ter interpretações distintas. Podemos citar como exemplos de medidas de incerteza: a probabilidade, a variabilidade em eventos (Ross, 2006) e a entropia (Schneier, 1995). A incerteza está presente em diferentes ambientes de aplicação. Podemos citar instrumentos de observação experimental como: sequenciadores (Kato, 2009) (incerteza em cada base gerada), equipamentos de elucidação de estruturas proteicas (Pevsner, 2009)(incerteza na definição da posição de um átomo). Além disso, métodos estatísticos de inferência (inferência via intervalos de confiança) e testes de hipótese (erros tipo I e II associados à análise realizada) (Casella e Berger, 2001) também apresentam incerteza em seus resultados. Da mesma forma, algoritmos de classificação e análise de agrupamentos também têm incerteza inerente aos resultados (Witten e Frank, 2005), geralmente avaliada por validação. Como em vários casos a incerteza é pequena, o determinismo dos resultados é assumido. Mesmo sendo pequena, como estes dados são utilizados para construção de conhecimento de mais alto nível, a incerteza pode ser propagada e aumentada. Tomando como exemplo a aplicação de um algoritmo de classificação, o problema de propagação é agravado, uma vez que os métodos de validação, frente a dados incertos, são incapazes de avaliar corretamente a eficiência de um classificador.

Dada a presença de incerteza, algumas questões surgem para tratamento destes dados. Como considerá-la no armazenamento de dados? Uma área que tenta sanar esta questão são modelos de bancos de dados probabilísticos (Green e Tannen, 2006). Uma etapa posterior ao armazenamento é a extração de informação sobre conjuntos de dados com incerteza. É muito comum a aplicação de mineração considerar os dados determinístas. A área de aprendizado de máquina sob incerteza (Uncertain Learning) (Aggarwal e Yu, 2009; Ngai et al., 2006) extrapola os métodos determinístas de extração de conhecimento para cenários probabilísticos.(Brenner, 1999)

A incorporação de incerteza no armazenamento e análise de dados passa por uma questão anterior. Como quantificá-la? Isto pode ser feito via observação (Schnoes et al., 2009) ou metodologias indiretas (Jones et al., 2007). A observação foca na ocorrência do evento que fornece a informação sobre a incerteza. A ocorrência de erros na anotação de genes (Brenner, 1999), a variabilidade de taxonomias em um banco de dados (EBI, 2011), a variação no código genético (Mooney, 2005) são exemplos deste tipo de estimação. Metodologias indiretas buscam a estimação sem a observação do evento informativo, mas

sim de um evento correlacionado. Quando trata-se da inferência sobre incerteza em bancos de dados, é possível tratá-la em níveis globais e locais:

1. **Incerteza no nível global**

   Vamos considerar a incerteza como uma probabilidade de ocorrência de erros em um registro qualquer de uma base de dados. Pode-se não condicionar a taxa de erros a um grupo ou ao próprio registro. O não condicionamento apresenta a informação de incerteza em nível global. Um exemplo no contexto UniprotKB/Swiss-prot é tomar uma proteína qualquer e avaliar sua chance de conter erros de anotação. Esta informação de incerteza é de nível global. Quando aplicado um método de anotação automática, esta probabilidade pode ser dada por informações de precisão e acurácia. Em (Jones et al., 2007) estima-se, via uma metodologia indireta, a taxa de erros de anotação global em um banco de dados proteico. Ao considerar a busca por erros, a probabilidade global é pouco informativa. Neste caso, estratégias para investigação de incerteza em nível local devem ser utilizadas.

2. **Incerteza no nível local**

   A informação local mapeia algum tipo de característica à qual a incerteza possa ser condicionalmente dependente. No caso de um banco de dados, a identificação de atributos que indiquem maior ou menor incerteza, independente de sua definição. Suponha um subconjunto (taxon, classe estrutural etc.) do UniprotKB/Swiss-prot. Inferir a probabilidade de ocorrência de erro de anotação em uma proteína deste grupo é obter uma informação em nível local. A análise de variância (Faraway, 2004) trabalha com o conceito de incerteza global e local para avaliação de experimentos. Trabalhos que tratam a detecção de erros como (Gilks et al., 2002) utilizam esta abordagem, uma vez que mensuram medidas para tanto. Como a motivação central deste trabalho é propor critérios de prioridade de revisão, a estratégia aqui proposta infere sobre o nível de incerteza local (probabilidade de ocorrência de erro em uma entidade), condicionando-a à excepcionalidade da entidade. Alguns deste trabalhos são apresentados na seção seguinte.

2.3 Trabalhos relacionados

Iniciativas para avaliar a qualidade de bancos de dados biológicos são apresentadas na literatura. Tanto para quantificar a incerteza em nível global (qualidade do banco como um todo) quanto para avaliar a incerteza em nível local (qualidade dos registros).

Bancos de dados proteicos de acesso público apresentam taxas de erros de anotação significativas, quando consideradas taxas globais. Isto pode ser explicado, dado que grande parte das anotações são geradas por análise de similaridade (cerca de 70% no UniprotKB/Swiss-prot na versão 2011_5 (UniprotKB, 2011b)), e que esta é uma fonte

importante de erros (Gilks et al., 2002). Em (Jones et al., 2007), estima-se que, no Gene Ontology (GO), 49% das sequências anotadas automaticamente tem anotações incorretas. A qualidade de bancos de dados curados, frente a anotados automaticamente também é evidenciada. O trabalho de (Schnoes et al., 2009) mostra a taxa de erros de anotação em funções moleculares variando entre 6% e 63% em bancos de geração automática. Em 27% das famílias estudadas a taxa de erros estava acima de 80%. No Swiss-prot este valor foi próximo de 0% para um conjunto representativo de sequências.

Dada a disponibilidade de versões prévias de alguns bancos de dados, é possível realizar a exploração das relações e suas modificações ao longo tempo. Em (Hartung et al., 2008) esta avaliação é realizada no Gene Ontolology (GO) e disponibilizada no sistema Onex para consulta de histórico de modificações (Hartung et al., 2009).

Uma vez que grande parte das anotações proteícas de mais alto nível são categóricas, há a necessidade de métodos específicos para tratamento destes dados. Em (Artamonova et al., 2005, 2007) são filtradas exceções baseadas em regras de anotações categóricas do UniprotKB/Swiss-prot. Estas exceções, definidas de forma determinísta, em 97% das instâncias estavam relacionadas com erros. Entretanto, como propõem (Hipp et al., 2001; He et al., 2002; Narita e Kitagawa, 2008), a excepcionalidade pode ser relativa. Neles são apresentadas medidas de excepcionalidade em dados categóricos.

Neste trabalho, em última instância, a metodologia central é quantificar e incrementar a qualidade do banco de dados UniprotKB/Swiss-prot. A melhoria do banco se dará pela adoção de um critério de revisão orientado pela chance de ocorrência de erros em um determinado registro (maior chance de erro implica em maior prioridade de revisão). Não será tratada a temática de correção. Como a chance de erros é inferida sobre escopo local (para cada entidade), a pré-suposição adotada é condicionar o erro à excepcionalidade apresentada por um registro. Diferente de (Artamonova et al., 2005, 2007) a relação excepcionalidade erro será relativizada.

Capítulo 3

Objetivos

3.1 Contexto

O UniprotKB/Swiss-Prot disponibiliza versões prévias de dados. Com o conhecimento destas versões é possível caracterizar as mudanças que o banco de dados sofreu sob a dimensão temporal. Estas modificações podem se dar em termos da dinâmica de incorporação e estrutura do conhecimento na base de dados. Elas podem identificar uma evolução no banco refletida pelo aperfeiçoamento de técnicas para construção do conhecimento agregado a base. Estudos desta natureza auxiliam na investigação sobre a estabilidade de ontologias, organizações de dados amplamente utilizadas no contexto biológico.

A caracterização da dinâmica de modificação do UniprotKB/Swiss-prot não se restringe ao aspecto descritivo, mas também pode servir de subsídio para incremento da qualidade no banco. Podemos avaliar as modificações do banco ao longo do tempo e utilizá-las como características informativas na detecção de modificações futuras. Processos de detecção de erros e estabelecimento de critérios de prioridade de revisão podem ser propostos a partir de informações passadas. Contribuições metodológicas nesta linha são de grande validade sobretudo em bancos onde não exista uma política sistemática de revisão ou prioridade de revisão.

A investigação das modificações, tanto sob aspecto descritivo e preditivo, pode ser feita em quaisquer atributos presentes em um banco com versões passadas disponíveis. Atributos de maior representatividade são escolhas mais adequadas, uma vez que possuem a sumarização da entidade armazenada (proteínas neste trabalho). No caso do UniprotKB/Swiss-prot, adota-se a premissa que as palavras-chaves são atributos representativos.

Considerando este contexto, este trabalho trata da caracterização da dinâmica do UniprotKB/Swiss-prot e proposição de uma estratégia sistemática de curagem através de um critério de prioridade de revisão das entidades proteínas armazenadas no banco.

3.2 Objetivo Geral

Este trabalho objetiva caracterizar e quantificar a incerteza global existente no UniprotKB/Swiss-prot e inferir a incerteza em nível local (instância) condicionada à excepcionalidade do registro. Esta informação subsidiará a construção de uma metodologia para determinar prioridades na revisão de registros da base. Todo o estudo será conduzido considerando o conjunto de palavras-chave do UniprotKB/Swiss-prot, embora seja possível a aplicação da metodologia a bases com perfil similar à estudada (atributos categóricos e disponibilidade de versões prévias).

3.3 Objetivos Específicos

Os objetivos específicos deste trabalho são:

1. a caracterização e quantificação das mudanças temporais que se dividem em duas tarefas:
   a) caracterização e quantificação das modificações, ao longo do tempo, de registros quanto ao conteúdo de palavras-chave (incerteza em nível global);
   b) estudo da dinâmica de mudança dos padrões de relações entre palavras-chave (incerteza sobre os padrões existentes no banco).
2. proposta de um critério de prioridade de revisão em bancos de dados categóricos. O critério é orientado pela chance de ocorrência de erros em um registro (incerteza em nível local). A inferência sobre a incerteza se dá condicionando-a à excepcionalidade do respectivo registro.

Capítulo 6

Caracterização das modificações em relações entre palavras-chave no UniprotKB/Swiss-prot

6.1 Introdução

O capítulo 5 apresenta as séries descritivas sobre o comportamento de incorporação ou exclusão de dados ao banco. Porém, estas estatísticas isoladas não demonstram mudanças em seu padrão. Para tanto, uma análise especializada é proposta, considerando um conjunto de regras candidatas a ter relevância semântica, mineradas em todas as versões estudadas. Neste conjunto, investiga-se a existência de mudanças generalizadas no padrão das regras, além da identificação do ponto onde estas mudanças são mais intensas. Para estas análises, foram utilizadas a combinação da mineração de regras de associação seguida da avaliação da variação das confianças nas regras mineradas. Para o estudo destas séries foi utilizado o modelo de Cadeias de Markov com estados ocultos. A dinâmica de análise foi apresentada nas seções 4.3 e 4.4.

6.2 Dados e Métodos

A disponibilidade das 15 versões das palavras-chave do UniprotKB/Swiss-prot permitem o estudo das regras de forma individual ou em conjunto. A representação do conhecimento presente no banco será feita através de conjuntos de regras de associação. Para capturar um conjunto amplo de regras com baixa redundância, foi aplicado o algoritmo de mineração ChARM. Cada uma das versões do banco foi minerada integralmente, ao contrário do capítulo anterior, em que os estudos foram realizados sobre conjuntos interseção de palavras-chave e registros em versões subsequentes. Sob parametrização de suporte mínimo 0.001 e confiança irrestrita, foram obtidas as regras para cada versão. Considerando todas as versões, foram obtidas 356608 regras (considerando repetições de

regras), distribuídas segundo a figura 6.1. O crescimento das regras, observado em quase todas as versões se deve ao aumento de palavras-chave (conjunto dicionário) e do número de registros destacado na tabela 5.1. A exceção é observada na versão 12 para 13, onde mesmo havendo um crescimento do número de registros, há estabilização do número de palavras-chave (903 tanto nas versões 12 quanto 13 apresentado na tabela 5.1). Como a mineração é realizada sob suporte 0.001, o número de *itemsets* fechados minerados na versão 13 será menor do que a 12 o que também reduz o número de regras.

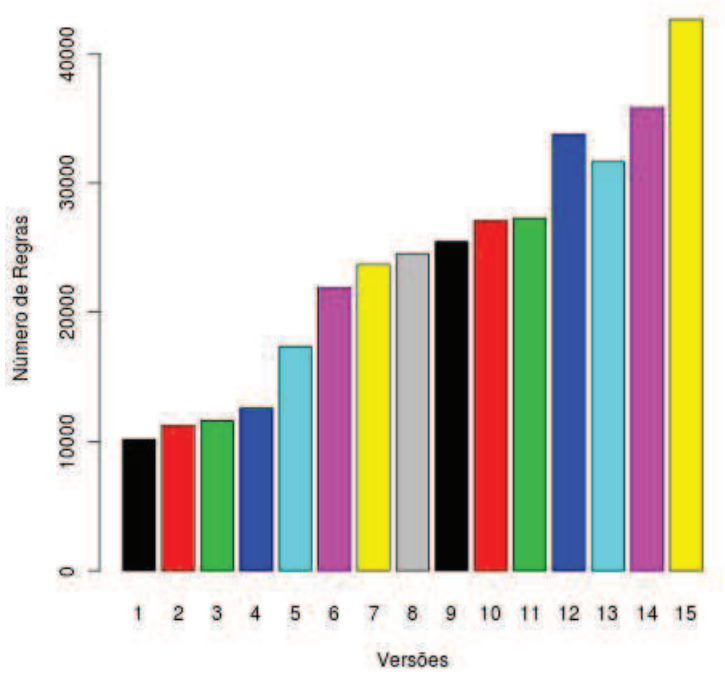

Figura 6.1: Crescimento do número de regras mineradas por versão do UniprotKB/Swiss-prot

Foram filtradas 75418 regras de associação distintas presentes nos conjuntos de mineração. Para analisar o conjunto de relações entre palavras-chave perenes, foi filtrado um conjunto de 2949 regras persistentes (regras que aparecem no resultado de mineração de todas as versões em estudo). Este número representa uma redução de mais de 96% em relação ao conjunto original. O conjunto de regras resultante se mostra candidato a refletir o conhecimento presente na base. Para cada regra, existe uma sequência de valores de suporte e confiança. A avaliação bidimensional respectiva a estas duas variáveis pode ser feita porém aqui será abordada a análise sobre as séries de confiança.

A partir do conjunto de regras que atendem as características de filtro, foram ajustadas cadeias de Markov com estados ocultos. O ajuste foi feito em cada uma das séries de confianças das regras filtradas. Com base nos modelos estimados, foram inferidas informações sobre as características do conjunto de regras.

6.3 Resultados e Discussão

As regras que figuraram em todas as minerações, sem consideração dos filtros paramétricos, apresentam um perfil similiar em termos de valores de confiança. A distribuição destes valores é bimodal. No domínio, regras com confianças extremas, ou seja, com grandes e pequenos valores se mostram mais frequentes. Entretanto as distribuições, apresentadas na figura 6.3, sofrem uma leve modificação ao longo do tempo. Alguns padrões vão emergindo a partir da incorporação de informação na base. Isto pode ser observado, pelo crescimento de 159% no número de regras com confiança superior a 0.9, em todo o período de análise. Este crescimento é apresentado na figura 6.2. Este salto sai de 256 regras, na versão 1/43 para 422 na versão 15/57 (intervalo em vermelho no histograma da figura 6.3). Durante o período de 6 anos avaliado pela análise, novo conhecimento foi agregado nos processos de anotação. Esta modificação no conhecimento pode estar refletida na mudança de distribuição de confiança das regras.

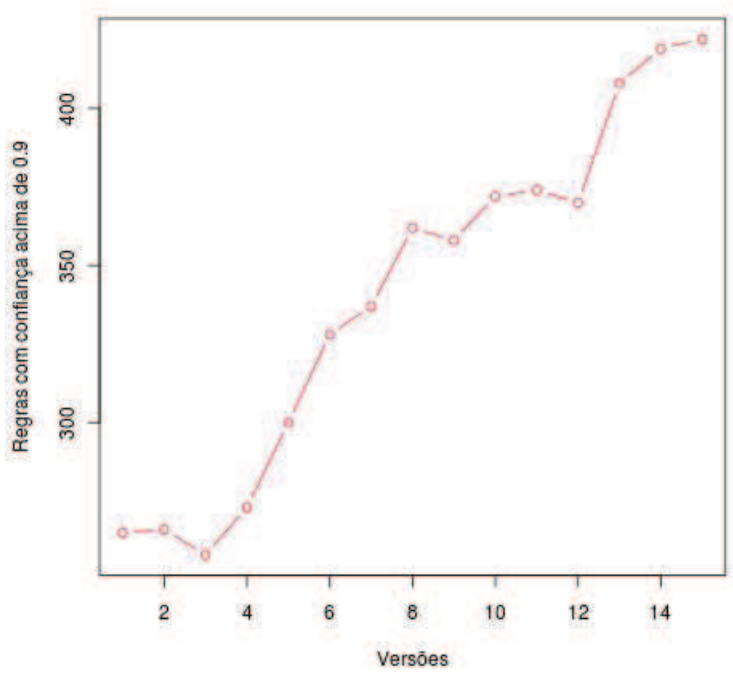

Figura 6.2: Crescimento do número de regras com confiança superior a 0.9 ao longo da versões estudadas do UniprotKB/Swiss-prot

Uma vez observada a modificação, as perguntas subsequentes são qualitativas quanto ao perfil da mudança. Ou seja, quando, quais e como as regras modificam sua importância na base. Esta análise será feita sobre o conjunto de regras com grande confiança em algum momento. Elas serão tipificadas como emergentes, decadentes e consistentes. Vide exemplo da tabela 4.6 e da figura 4.4. Para tanto são ajustados modelos ocultos de markov (HMM) para cada uma das séries, variando entre 1 e 3 estados. O objetivo da análise

é reconhecer mudanças nas força das regras, que podem ser evidenciadas por modelos de 2 estados. Os modelos com 3 estados vão capturar séries de características de 3 ou mais estados, que possuem mudanças de padrão contundente. É importante ressaltar que as séries em estudo possuem 15 observações, o que torna mais restritiva a utilização de modelos com mais do que 3 estados.

Dentre as regras mineradas presentes em todas as versões estudadas do banco UniprotKB/Swiss-prot, foram filtradas as que possuíam, em algum momento, valores de confiança superior ou igual a 0.9. O total resultante de 513 regras foram identificadas como potencialmente válidas em termos de significado semântico (valores de confiança superiores a 0.9). A partir dos diferentes ajustes das cadeias de Markov ocultas sob as séries, foram escolhidos os modelos mais adequados segundo o critério BIC (Bayesian Information Criterion). Dos modelos resultantes, 36 possuem 1 estado e 404 possuem 2 estados. As cadeias ajustadas de 2 estados, e somente com um ponto de mudança entre estados, foram divididas segundo o padrão inicial de confiança. São elas: emergentes com valores iniciais mais baixos (279); decadentes com valores iniciais mais altos (85). As séries consistentes, emergentes e decadentes são apresentadas nas figuras 6.4, 6.5 e 6.6.

Considerando a modificação no padrão das séries de dois estados, fica evidente uma mudança generalizada em torno da versão 5. Isto pode ser verificado tanto nas regras emergentes como nas decadentes. Vale ressaltar que grande parte destas séries possuem apenas um ponto de mudança entre estados, ou seja, 2 estados no modelo e a série de estados mais verossímil apresenta apenas uma mudança entre estados. As distribuições do número de mudanças de padrão de perfil (queda e salto) são apresentadas respectivamente nas figuras 6.7 e 6.8.

Além da incidência de mudança na versão 5 (versão 4 para 5), a maior intensidade da mudança (altura do salto-queda) também ocorre nesta versão. A avaliação da distribuição de quedas e saltos nas confianças das regras decadentes e emergentes indica isto (figuras 6.9 e 6.10).

Os resultados anteriores mostram uma mudança generalizada (localização e intensidade) em torno da versão 5 do UniprotKB/Swiss-prot datada de maio de 2005. A incorporação de anotação às entidades proteícas também se estabiliza a partir desta versão (resultado do capítulo 5). Entre as associações filtradas como emergentes, algumas apresentaram, no lado direito da regra o termo "Complete Proteome". O mesmo termo também é encontrado em relações decadentes. Exemplos são:

- "$\{Thiaminebiosynthesis, Transferase\} \rightarrow \{Completeproteome\}$" (emergente),
- "$\{Transcriptiontermination\} \rightarrow \{Completeproteome\}$" (emergente),
- "$\{Transferase, rRNAprocessing\} \rightarrow \{Completeproteome\}$" (emergente),
- "$\{Transferase, tRNAprocessing\} \rightarrow \{Completeproteome\}$" (emergente),
- "$\{rRNA - binding, tRNA - binding\} \rightarrow \{Completeproteome\}$" (emergente).

- "$\{Completeproteome, DNAdamage\} \rightarrow \{DNA - binding\}$" (decadente),
- "$\{Completeproteome, DNAexcision, DNArepair\} \rightarrow \{Excisionnuclease\}$" (decadente),
- "$\{Completeproteome, FAD, Oxidoreductase\} \rightarrow \{Flavoprotein\}$" (decadente).

Uma vez que este termo identifica proteínas que podem ser derivadas por tradução de genes de um genoma completamente sequenciado, tem-se um resultado trivial, já que vários projetos de sequenciamento estão sendo finalizados. Ou seja, a taxa de crescimento da frequência de "completeproteome" conjunta aos termos da regra sobe, se comparada à mesma taxa dos termos envolvidos na regra.

As regras consistentes apresentam relações óbvias sob aspecto semântico. Um exemplo é a regra "$\{Metalloprotease\} \rightarrow \{Zinc\}$". "Metalloprotease" diz respeito a enzimas catalíticas que usam metais para a reação de quebra. Em sua maioria, estes metais são Zinco, representado pelo termo "Zinc". Outro caso similar é a relação "$\{GPI - anchor\} \rightarrow \{Signal\}$". O Glycosylphosphatidylinisotol (GPI) é um grupamento químico que estabelece a ligação entre proteínas e membranas lipídicas celulares. Elas desempenham papel importante também na propagação e mediação de sinais. Esta mediação é refletida pelo termo "Signal" presente no lado direito da regra. A relação de sinalização da GPI ("$\{GPI - anchor\} \rightarrow \{Signal\}$") se mostrou estável ao longo do período de análise.

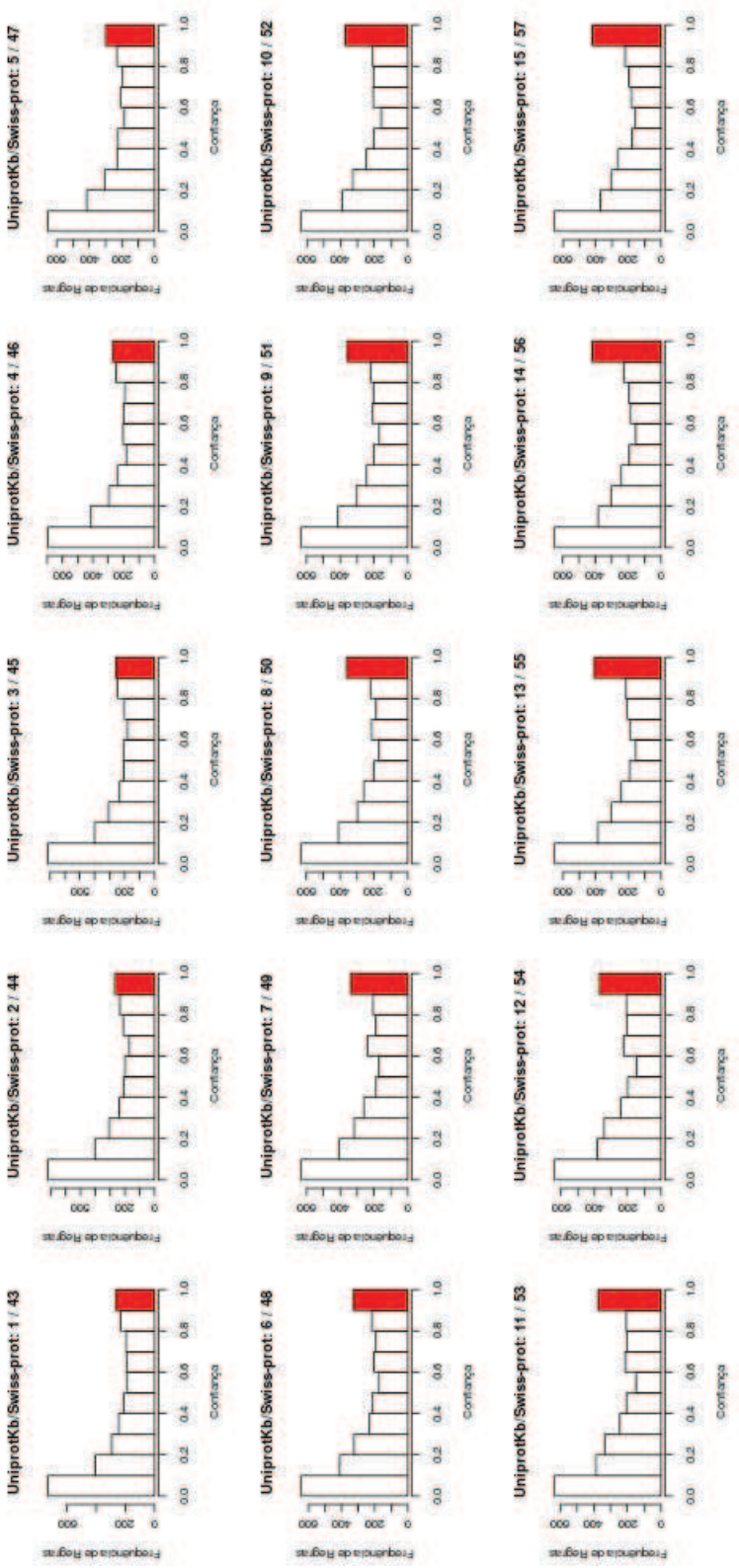

Figura 6.3: Distribuição de valores de confiança para minerações sob suporte 0.001 de 15 versões do banco de dados UniprotKB/Swiss-prot. Em vermelho são destacadas a frequência das regras com confiança acima de 0.9. Estas aumentam o número em 159%

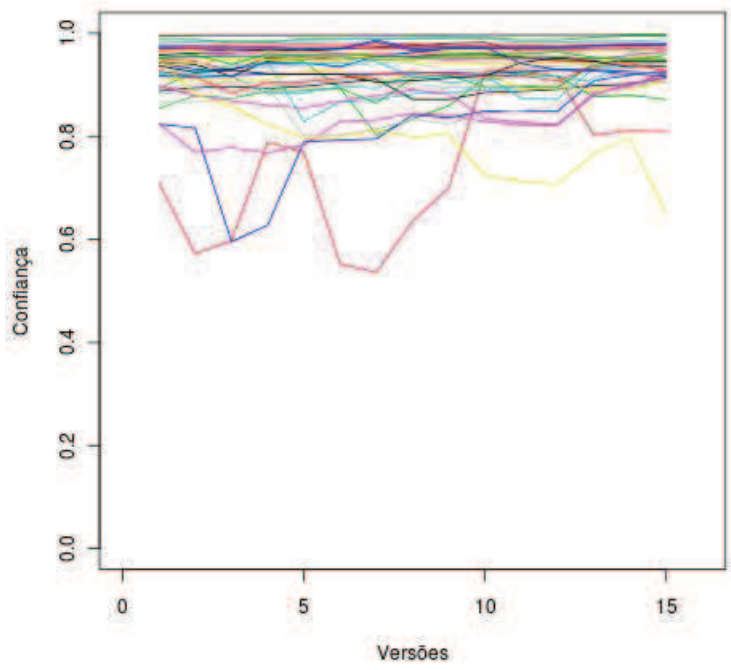

Figura 6.4: Séries de confiança referentes as 15 versões do UniprotKB/Swiss-prot, as quais os modelos identificam padrão único de variação

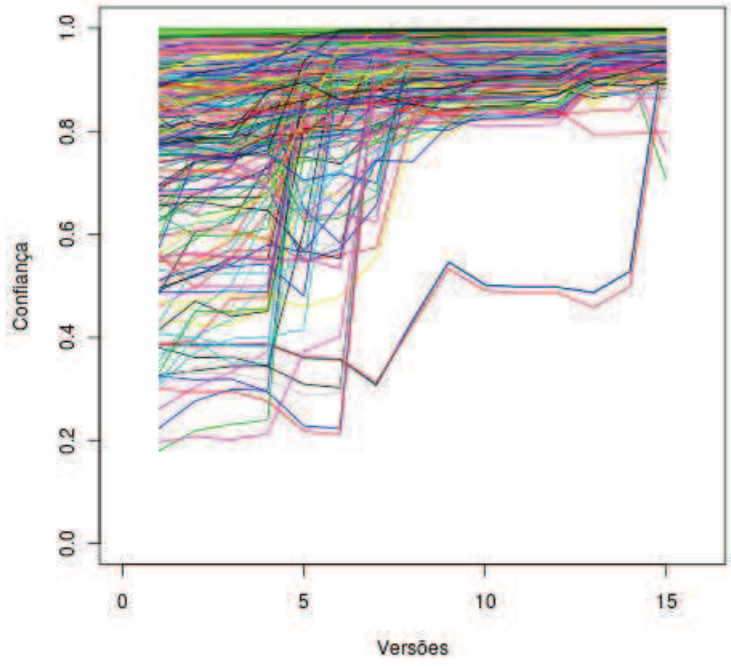

Figura 6.5: Séries de confiança referentes as 15 versões do UniprotKB/Swiss-prot, as quais os modelos identificam padrão emergente de variação

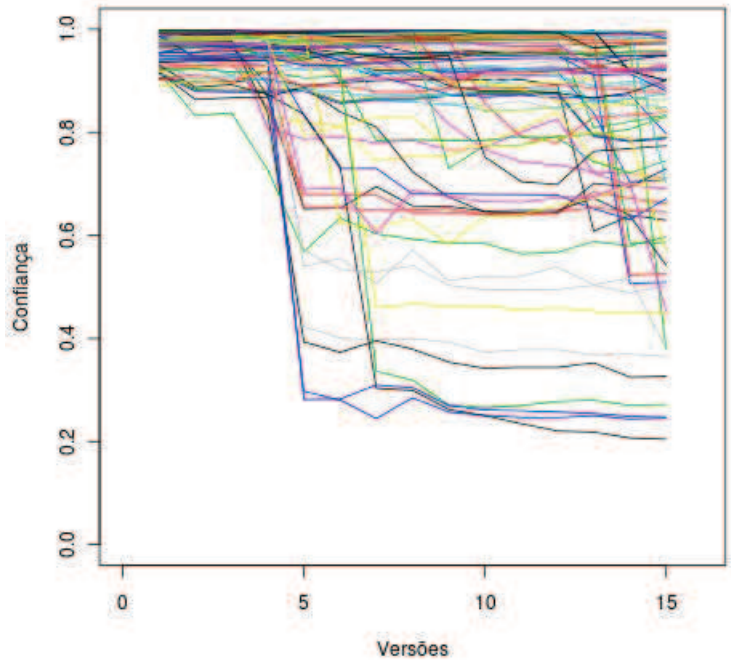

Figura 6.6: Séries de confiança referentes as 15 versões do UniprotKB/Swiss-prot, as quais os modelos identificam padrão decadente de variação

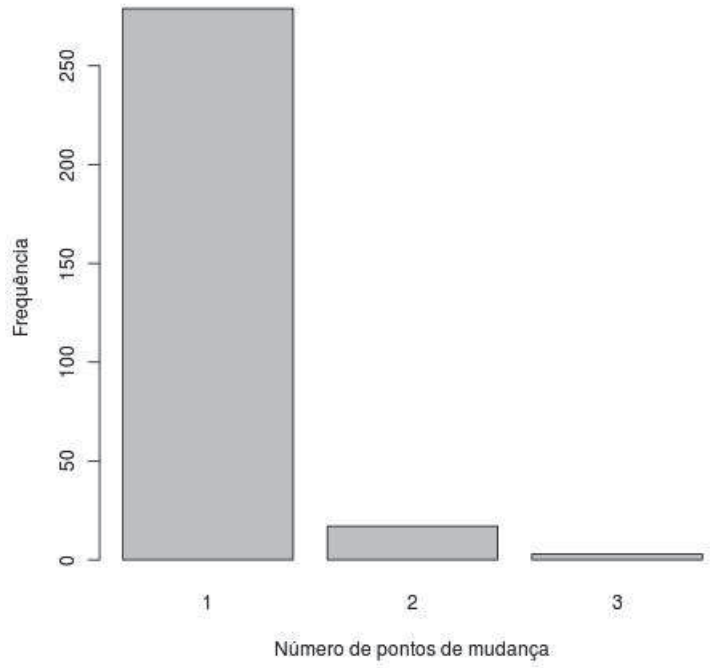

Figura 6.7: Número de pontos de mudança nas séries de 2 estados (padrão emergente)

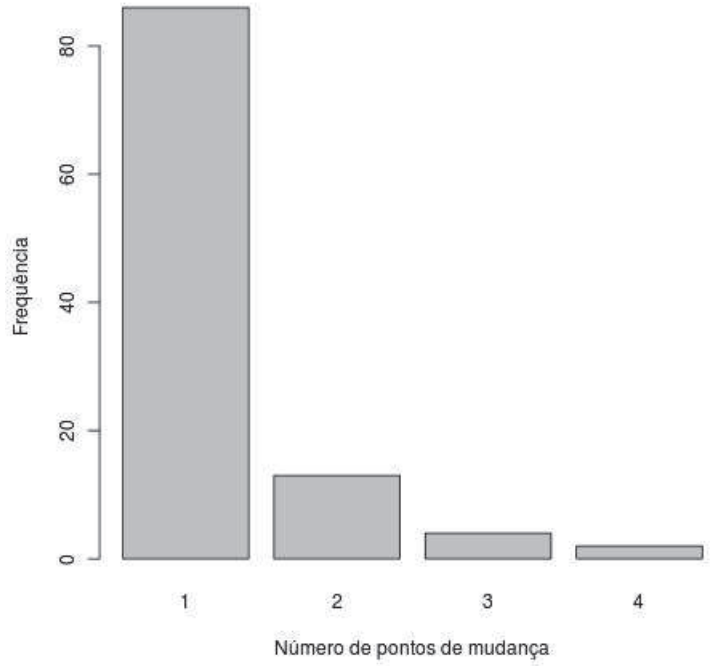

Figura 6.8: Número de pontos de mudança nas séries de 2 estados (padrão decadente)

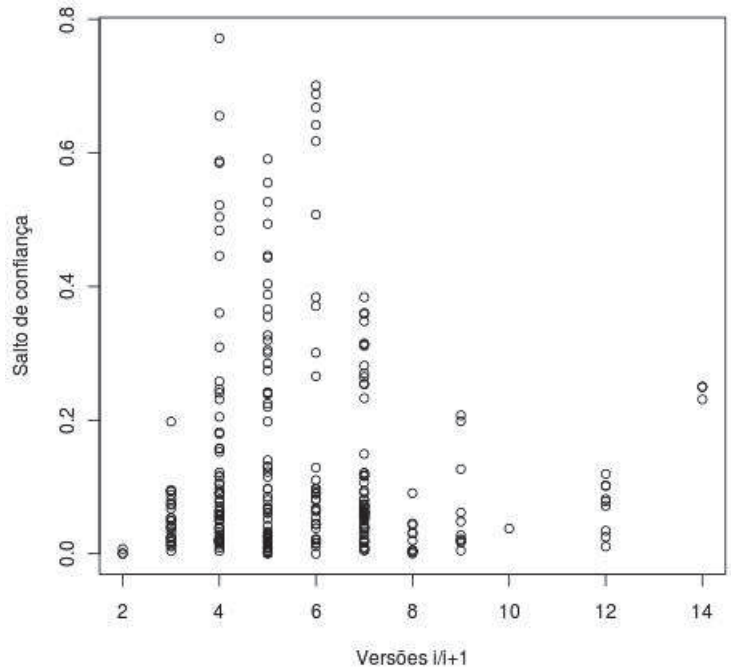

Figura 6.9: Diferença entre o padrão de confiança das regras emergentes (intensidade do salto)

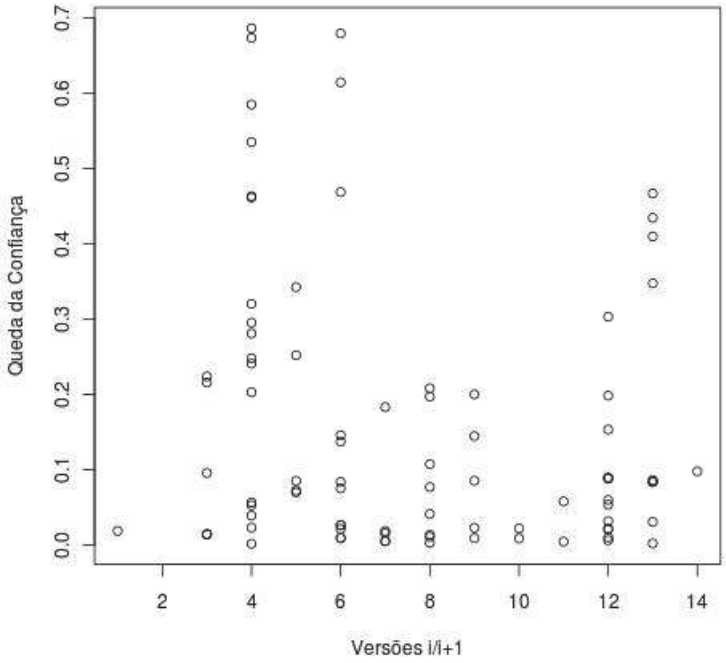

Figura 6.10: Diferença entre o padrão de confiança das regras decadentes (intensidade da queda)

Referências Bibliográficas

Aggarwal, C. C. e Yu, P. S. (2009). A Survey of Uncertain Data Algorithms and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):609–623.

Agrawal, R. e Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In Bocca, J. B.; Jarke, M. e Zaniolo, C., editores, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pp. 487–499. Morgan Kaufmann.

Artamonova, I. I.; Frishman, G. e Frishman, D. (2007). Applying negative rule mining to improve genome annotation. *BMC Bioinformatics*, 8(1):261.

Artamonova, I. I.; Frishman, G.; Gelfand, M. S. e Frishman, D. (2005). Mining sequence annotation databanks for association patterns. *BMC Bioinformatics*, 21(3):49–57.

Barbara, D.; Couto, J.; Jajodia, S. e Wu, N. (2001). Adam: a testbed for exploring the use of data mining in intrusion detection. *SIGMOD Rec.*, 30:15–24.

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.

Berg, J. M.; Tynoczko, J. L. e Stryer, L. (2008). *Biochemistry*. Guanabara Koogan.

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T. e Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem*, 80(2):319–324.

Brause, R.; Langsdorf, T. e Hepp, M. (1999). Neural data mining for credit card fraud detection. In *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '99, pp. 103–, Washington, DC, USA. IEEE Computer Society.

Brenner, S. E. (1999). Errors in genome annotation. *Trends in Genetics*, 15(132-133).

Casella, G. e Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center.

Ch, V.; Banerjee, A.; Kumar, V. e Chandola, V. (2007). Outlier detection: A survey. Technical report, University of Minnesota.

Chandola, V.; Banerjee, A. e Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41:15:1–15:58.

Cogger, K. O. (2010). Nonlinear multiple regression methods: a survey and extensions. *Int. J. Intell. Syst. Account. Financ. Manage.*, 17:19–39.

Consortium, T. U. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, 39(suppl 1):D214–D219.

Cordeiro, G. M. e Demétrio, C. G. B. (2002). Modelos lineares generalizados. Technical report, Universidade de São Paulo - ESALQ.

de Oliveira, D. C.; Queiroz, C. e Chaves, L. M. (2006). Cadeias de markov com estados latentes com aplicações em análises de sequências de dna. *Revista de Matemática e Estatística*, 24(2):51–66.

Dempster, A. P.; Laird, N. M. e Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

EBI (2011). Current trembl statistics http://www.ebi.ac.uk/uniprot/TrEMBLstats/.

Faraway, J. J. (2004). *Linear Models with R*. Chapman and Hall/CRC. ISBN 1-584-88425-8.

Genomenet (2011). Kioto university bioinformatics center http://www.genome.jp/en/db_growth.

Gilks, W. R.; Audit, B.; De Angelis, D.; Tsoka, S. e Ouzounis, C. A. (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18(12):1641–1649.

Green, T. J. e Tannen, V. (2006). Models for incomplete and probabilistic information. *IEEE Data Engineering Bulletin*, 29.

Guillet, F. e Hamilton, H. E. (2007). *Quality Measures in Data Mining*. Springer.

Harismendy, O.; Ng, P. C.; Strausberg, R. L.; Wang, X.; Stockwell, T. B.; Beeson, K. Y.; Schork, N. J.; Murray, S. S.; Topol, E. J.; Levy, S. e et al. et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10(3):R32.

Hartung, M.; Kirsten, T.; Gross, A. e Rahm, E. (2009). Onex: Exploring changes in life science ontologies. *BMC Bioinformatics*, 10.

Hartung, M.; Kirsten, T. e Rahm, E. (2008). Analyzing the evolution of life science ontologies and mappings. In *Proceedings of the 5th international workshop on Data Integration in the Life Sciences*, DILS '08, pp. 11–27, Berlin, Heidelberg. Springer-Verlag.

He, Z.; Deng, S. e Xu, X. (2005). An optimization model for outlier detection in categorical data. In *ICIC (1)*, pp. 400–409.

He, Z.; Xu, X.; Huang, J. Z. e Deng, S. (2002). Fp-outlier: frequent pattern based outlier detection. Technical report, Dalian University of Technology.

Hennessy, J. e Patterson, D. (2003). *Computer Architecture - A Quantitative Approach*. Morgan Kaufmann.

Hipp, J.; Güntzer, U. e Grimmer, U. (2001). Data quality mining – making a virtue of necessity. In *In proceedings of the 6th acm sigmod workshop on research issues in Data Mining and Knowledge Discovery (dmkd 2001)*, pp. 52–57.

Hodge, V. e Austin, J. (2004). A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22:85–126.

Hubbard, T. J. P.; Murzin, A. G.; Brenner, S. E. e Chothia, C. (1997). Scop: a structural classification of proteins database. *J. Mol. Biol*, 247:536–540.

Jones, C. E.; Brown, A. L. e Baumann, U. (2007). Estimating the annotation error rate of curated go database sequence annotations. *BMC Bioinformatics*, pp. –1–1.

Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M. e Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(Database issue):D355–D360.

Kato, K. (2009). Impact of the next generation DNA sequencers. *International journal of clinical and experimental medicine*, 2(2):193–202.

Khreisat, L. (2007). Quicksort: A historical perspective and empirical study. *International Journal of Computer Science and Network Security*, 7(12).

Kulikova, T.; Akhtar, R.; Aldebert, P.; Althorpe, N.; Andersson, M.; Baldwin, A.; Bates, K.; Bhattacharyya, S.; Bower, L.; Browne, P.; Castro, M.; Cochrane, G.; Duggan, K.; Eberhardt, R.; Faruque, N.; Hoad, G.; Kanz, C.; Lee, C.; Leinonen, R.; Lin, Q.; Lombard, V.; Lopez, R.; Lorenc, D.; McWilliam, H.; Mukherjee, G.; Nardone, F.; Pastor, M. P.; Plaister, S.; Sobhany, S.; Stoehr, P.; Vaughan, R.; Wu, D.; Zhu, W. e Apweiler, R. (2007). Embl nucleotide sequence database in 2006. *Nucleic Acids Research*, 35(Database issue):16–20.

Lee, Y. W.; Pipino, L. L.; Funk, J. D. e Wang, R. Y. (2009). *Journey to Data Quality*. The MIT Press.

Leinonen, R.; Nardone, F.; Zhu, W. e Apweiler, R. (2006). Unisave: the uniprotkb sequence/annotation version database. *Bioinformatics*, 22(10).

Mahoney, M. V. e Chan, P. K. (2003). Learning rules for anomaly detection of hostile network traffic. In *In: Proc. of International Conference on Data Mining (ICDM)*, pp. 601–604.

McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.

Mooney, S. D. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics*, pp. 44–56.

Narita, K. e Kitagawa, H. (2008). Outlier detection for transaction databases using association rules. In *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management*, WAIM '08, pp. 373–380, Washington, DC, USA. IEEE Computer Society.

Ngai, W. K.; Kao, B.; Chui, C. K.; Cheng, R.; Chau, M. e Yip, K. Y. (2006). Efficient clustering of uncertain data. In *Proceedings of the IEEE International Conference on Data Mining*.

Otey, M.; Parthasarathy, S.; Ghoting, A.; Li, G. e Narravula, S. (2003). Towards nic-based intrusion detection. In *In proceedings of the ninth ACM SIGKDD International conference o knowledge discovery and data mining*, pp. 723–728. ACM Press.

Pevsner, J. (2009). *Bioinformatics and functional genomics*. Wiley-Blackwell.

R development core team (2011). *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pp. 257–286.

Ross, S. M. (2006). *Introduction to Probability Models, Ninth Edition*. Academic Press, Inc., Orlando, FL, USA.

Schneier, B. (1995). *Applied cryptography (2nd ed.): protocols, algorithms, and source code in C*. John Wiley & Sons, Inc., New York, NY, USA.

Schnoes, A. M.; Brown, S. D.; Dodevski, I. e Babbitt, P. C. (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput Biol*, 5(12).

Tan, P.-N.; Steinbach, M. e Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

UniprotKB (2011a). Ftp site of uniprotkb ftp://ftp.uniprot.org/pub/databases/uniprot/.

UniprotKB (2011b). Uniprotkb/swiss-prot user manual http://expasy.org/sprot/relnotes/relstat.html.

Witten, I. H. e Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, San Francisco, CA, 2nd edição.

Yairi, T.; Kato, Y. e Hori, K. (2001). Fault detection by mining association rules from house-keeping data. In *In Proc. of International Symposium on Artificial Intelligence, Robotics and Automation in Space*.

Zaki, M. J. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery: An International Journal*, 9(3):223–248.