BIOINFORMATICS

Vol. 21 Suppl. 3 2005, pages iii49–iii57
doi:10.1093/bioinformatics/bti1206

Data and text mining

Mining sequence annotation databanks for association patterns

Irena I. Artamonova[1], Goar Frishman[1], Mikhail S. Gelfand[2,3,4] and Dmitrij Frishman[1,5,*]

[1]Institute for Bioinformatics, GSF-National Research Center for Environment and Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany, [2]Institute for Information Transmission Problems RAS, Bolshoi Karetny pereulok 19, Moscow, 127994, Russia, [3]State Scientific Center GosNIIGenetika, 1st Dorozhny proezd 1, Moscow, 117545, Russia, [4]Department of Bioengineering and Bioinformatics, M.V.Lomonosov Moscow State University, Vorobievy Gory 1-73, Moscow, 119992, Russia and [5]Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftzentrum Weihenstephan, 85350 Freising, Germany

Received on July 8, 2005; accepted on August 16, 2005

ABSTRACT

Motivation: Millions of protein sequences currently being deposited to sequence databanks will never be annotated manually. Similarity-based annotation generated by automatic software pipelines unavoidably contains spurious assignments due to the imperfection of bioinformatics methods. Examples of such annotation errors include over- and underpredictions caused by the use of fixed recognition thresholds and incorrect annotations caused by transitivity based information transfer to unrelated proteins or transfer of errors already accumulated in databases. One of the most difficult and timely challenges in bioinformatics is the development of intelligent systems aimed at improving the quality of automatically generated annotation. A possible approach to this problem is to detect anomalies in annotation items based on association rule mining.

Results: We present the first large-scale analysis of association rules derived from two large protein annotation databases—Swiss-Prot and PEDANT—and reveal novel, previously unknown tendencies of rule strength distributions. Most of the rules are either very strong or very weak, with rules in the medium strength range being relatively infrequent. Based on dynamics of error correction in subsequent Swiss-Prot releases and on our own manual analysis we demonstrate that exceptions from strong rules are, indeed, significantly enriched in annotation errors and can be used to automatically flag them. We identify different strength dependencies of rules derived from different fields in Swiss-Prot. A compositional breakdown of association rules generated from PEDANT in terms of their constituent items indicates that most of the errors that can be corrected are related to gene functional roles. Swiss-Prot errors are usually caused by under-annotation owing to its conservative approach, whereas automatically generated PEDANT annotation suffers from over-annotation.

Availability: All data generated in this study are available for download and browsing at http://pedant.gsf.de/ARIA/index.htm.

Contact: d.frishman@wzw.tum.de

Supplementary information: http://pedant.gsf.de/ARIA/index.htm

1 INTRODUCTION

Amino acid sequences are meaningless unless some kind of functional annotation is attached to them. The Webster dictionary defines annotation as ‘a note added by way of comment or explanation’. In particular, annotation of proteins can be defined as the systematic collection of facts explaining their cellular role and mechanism of action. Ideally, such annotation should include only experimentally verified information, typically extracted from scientific literature. This approach is being centrally pursued by the Swiss-Prot (now UniProt; Bairoch *et al*., 2005) database which is considered to be the gold standard of protein annotation owing to its high level of manual curation and resulting good data quality. In addition, many individual genome sequencing projects as well as several bioinformatics teams also produced in-depth annotation of each gene product for numerous completely sequenced genomes (e.g. Tatusov *et al*., 1996; Bult *et al*., 1996; Mewes *et al*., 1997; Cole *et al*., 1998; Karp *et al*., 1999; and many others). Literature curation plays a crucial role in creating reference annotation datasets, but at the same time it remains a time-consuming and creative process which, of course, is also not completely faultless. Numerous discrepancies between different careful annotation efforts have been reported (Brenner, 1999).

Unfortunately, only a small fraction of known proteins has been studied experimentally, and the explosion of sequence data makes it impossible to annotate all proteins manually. Most of the information about molecular sequences in today’s databanks is inferred by similarity with previously analyzed entities. Although such annotation transfer is technically efficient and often yields quite accurate function assignments, it also has severe intrinsic limitations (Smith, 1996; Wilson *et al*., 2000; Devos and Valencia, 2000) and is in some cases notoriously error-prone (Bork and Bairoch, 1996; Galperin and Koonin, 1998), especially for multidomain eukaryotic proteins (Hegyi and Gerstein, 2001). As a result, transitive annotation errors propagate in sequence databases, leading to the gradual deterioration of the total corpus of available annotation and complicating further analysis efforts (Gilks *et al*., 2002).

The problem acquired a new dimension in the last decade with the advent of whole-genome sequencing. In response to genomic data deluge large-scale software systems, such as GeneQuiz (Andrade *et al*., 1999), MAGPIE (Gaasterland and Sensen, 1996), WIT (Overbeek *et al*., 2000), PEDANT (Frishman *et al*., 2001), CMR (Peterson *et al*., 2001) and ENSEMBL (Hubbard *et al*., 2005), to name just a few, have been developed to conduct initial first-pass annotation of complete proteomes by systematically applying a large set of bioinformatics methods to gene products and assigning functional and structural attributes to them according to fixed recognition thresholds. The UniProt database offers automatic annotation of all available sequences through its TrEMBL supplement. As a result, millions of currently available protein annotation entries in various

[*]To whom correspondence should be addressed.

databases have never been and will unlikely be verified by human experts.

Thus, a difficult and timely challenge in bioinformatics is the development of intelligent systems aimed at improving the quality of automatically generated annotation. In addition to constant efforts to improve the effectiveness of individual bioinformatics techniques, a possible approach to this problem is to check the consistency of assigned features and then to mark potentially erroneous features for manual inspection or to add those features that could have been missed. In recent years a large number of consensus-based approaches to improving gene function prediction have been proposed which usually operate by detecting inconsistencies in the annotation of related proteins forming a sequence cluster (Xie *et al*., 2002; Kaplan *et al*., 2003; Kunin and Ouzounis, 2005).

Another promising approach to intelligent filtering and improvement of biological annotation is through knowledge discovery techniques aimed at detecting common patterns, rules or anomalies. In addition to being widely used for mining biological literature (Hu *et al*., 2005) and experimental data (Michailidis and Shedden, 2003), rule-based approaches have been applied to predict protein annotation features from a set of other annotation features (Eisenhaber and Bork, 1999; Kretschmann *et al*., 2001; Yu, 2004). Major protein annotation efforts routinely use rule-based procedures for checking the integrity of information, finding minor errors and automating trivial annotation procedures which do not require human intervention (Fleischmann *et al*., 1999; Gattiker *et al*., 2003; Wu *et al*., 2003). Uninformative pieces of information (e.g. description lines containing only words such as ‘hypothetical’, ‘putative’, ‘unknown’ transferred from the best similarity hit) can be filtered out using simple lexical analyses based on specially prepared vocabularies (Andrade *et al*., 1999; Kasukawa *et al*., 2003).

A more sophisticated approach to this problem involves automatic learning of rules from a highly curated and reliable database, such as Swiss-Prot, and then using these rules either to further improve annotation in the same database, or in another automatically generated database, such as TrEMBL. Kretschmann *et al*. (2001) applied the C4.5 data mining algorithm to derive decision trees representing the knowledge on Swiss-Prot keywords. Rules obtained in this fashion combined with information on sequence groups gleaned by sequence analysis can be applied both for consistency checks within Swiss-Prot and for generating keywords for new TrEMBL entries with high accuracy. Conversely, exclusion rules for a specific protein group (e.g. sharing the same sequence motif) can be generated by the C4.5 algorithm to detect contradicting annotation items, as implemented in the Xanthippe post-processing system (Wieser *et al*., 2004).

Our purpose is to improve the automatic annotation in the PEDANT Genome Database (Riley *et al*., 2005; http://pedant.gsf.de) containing precomputed information resulting from bioinformatics analyses of publicly available genomes. Its main mission is to provide robust and up-to-date annotation of the vast majority of amino acid sequences which have not been subjected to in-depth manual curation by human experts in high-quality protein sequence databases, such as Swiss-Prot. Computational results produced by PEDANT are not verified manually and hence contain a large amount of erroneous assignments, ranging from completely false similarity hits accepted as true positives to more subtle cases where coarse function prediction is correct but functional specificity is mispredicted.

In this work we set out to find annotation errors by applying the association rule mining technique (Zhang and Zhang, 2002) to large protein annotation databases, such as PEDANT. This technique, originating from the analysis of data on market baskets involves the discovery of association relationships or correlations among a set of items. Association rule mining has been previously applied in bioinformatics to identify pairs of related GO terms (Bodenreider *et al*., 2005), interpret gene expression data (Creighton and Hanash, 2003; Becquet *et al*., 2002) and investigate relationships between different types of genomic data (Satou *et al*., 1997).

Association rules may have the form of simple implications. For example, in a database of annotated proteins, one such rule is the implication ‘Nuclear localization ⇒ Origin: eukaryota’, i.e. every protein annotated as localized in nucleus has a eukaryotic origin. The rules are not necessarily absolutely strict. For instance, the rule ‘Alternative splicing ⇒ Origin: eukaryota’ has exceptions, because viral genes also may be spliced (and alternatively spliced as well). However, this is still a valid rule, because the exceptions comprise a small fraction of the database. Thus, this rule is naturally interpreted as ‘the majority of proteins with evidence of alternative splicing originate from eukaryotic organisms’. ‘Many-to-one’ rules can also be considered. For instance, ‘Alternative splicing and Kinase ⇒ Origin: eukaryota’. In this example, alternatively spliced proteins are specific to eukaryotic organisms or viruses and kinases belong to either eukaryota or prokaryota. If a protein kinase is annotated as resulting from alternative splicing, then it is a eukaryotic protein.

Formally, we consider a database with multiple entries and each entry in the database is ascribed a finite number of features. In the particular implementation used here, an association rule is formulated in the form $(A_1\&\cdots\&A_n) \Rightarrow Z$. Here $A_1,\ldots,A_n$ (the Left-hand side or LHS) and Z (the Right-hand side or RHS) are different features, and the rule means ‘database entries that possess all features $A_1,\ldots,A_n$ are likely to possess feature Z’. Each rule is characterized by its coverage, the number of entries in the database that satisfy the LHS (possess all features $A_1,\ldots,A_n$), its support, the number of entries satisfying both the LHS and the RHS simultaneously, and its strength, the fraction of entries that satisfy LHS and RHS among the entries satisfying the LHS, i.e. strength is the probability that an entry will satisfy the RHS given that it satisfies the LHS.

In the application of association rule mining technique to improving annotation, the main assumption is that if the database annotations satisfy a rule ‘A and B imply C’ with a high support and a very high strength, then such a rule reflects some biological regularity or maybe a peculiarity of the annotation process. If the strength is very close, but not equal, to 1, then the rule has a minor number of exceptions. Although in some cases such exceptions may reflect biological reality, it is plausible that a significant fraction of them are actual errors in annotation. Hence our strategy is to find rules of high strength, e.g. in the range [0.95; 1), filter them, identify proteins that are exceptions from such rules, mark the features from the left-hand side and add the right-hand side feature of the rule to the annotation of such exception proteins.

To evaluate the validity of this approach we first analyze association rules derived from the high-quality Swiss-Prot database, focusing on the most formalized non-overlapping fields of a standard Swiss-Prot entry, such as protein length, the highest-level taxon of the protein origin, assignment of InterPro domains, keywords and features from the feature table. We then apply the technique to the full body of annotation produced by PEDANT and present both computationally and manually derived estimates of its usability. We demonstrate that exceptions from strong rules in the PEDANT database are indeed significantly enriched in errors which can be automatically flagged. We call our approach ARIA: association rules to improve annotation.

2 DATA AND METHODS

2.1 Extracting item sets from the Swiss-Prot database

We used release 44.0 of the Swiss-Prot database containing 153 871 protein sequence entries as the source of high quality manual annotation. In addition, to assess the dynamics of annotation updates in Swiss-Prot we also considered two preceding releases, 42.0 and 43.0, as well as three subsequent releases, 45.0, 46.3 and 47.0. Each Swiss-Prot entry corresponding to a single protein contains textual annotation presented in the form of records beginning with a two-letter record identifier. For example, the unique protein code identifier can be found in the ID record, description lines containing protein function begin with DE, the date of the last update is specified in the DT record and so on. According to the Swiss-Prot user manual (http://us.expasy.org/sprot/userman.html), there currently exist 22 record types, out of which the following 15 were not considered suitable for association rule mining in this study:

- records not containing protein annotation: date (DT), amino acid sequence (SQ)
- records containing unique protein attributes, such as protein code (ID), accession numbers (AC) and gene names (GN)
- records containing free-text annotation: literature reference (RN, RP, RC, RX, RG, RA, RT, RL), description line (DE) and comments (CC), since this part of Swiss-Prot annotation is not easily machine-parseable.

We also did not use the following two Swiss-Prot records—OS (Organism species) and OX (Taxonomy cross-reference). These two record types provide information that was considered redundant to the OC (Organism classification) record.

The five Swiss-Prot record types used in this work are as described below:

(1) *OC Organism classification.* In this work we only considered the top-level taxon which can take one of the four values, namely Eukaryota, Bacteria, Archaea or Viruses.

(2) *FT Feature table data.* This record contains positional features of protein sequences as described in biological literature, such as post-translational modifications or binding sites. Each record of this type consists of a record name (e.g. TRANSMEM), start and stop positions of the given feature and a brief description line. Release 44.0 of Swiss-Prot distinguishes between a total of 33 feature names (see http://us.expasy.org/sprot/userman.html# FT_keys for a full list). Many features do not reflect general properties of the entire protein chain. For example, the feature 'CONFLICT' indicates that different sources reported differing sequences for a given Swiss-Prot entry. Other local features, however, describe some characteristic of the associated protein. For example, the feature VARSPLIC not only points to the alternative part of the protein, but also serves as an indication that the protein is subject to alternative splicing. We selected for our work the following 16 features that have general discriminatory biological sense: ACT_SITE, DNA_BIND, CA_BIND, CARBOHYD, DISULFID, LIPID, METAL, MOD_RES, NP_BIND, PROPER, REPEAT, SE_CYS, SIGNAL, TRANSMEM, VARSPLIC and ZN_FING. These feature names were used by themselves as protein characteristics. Furthermore, description lines of four features, TRANSIT, CARBOHYD, LIPID and METAL, were used as additional sequence attributes since they have a restricted number of possible values. For example, the description line of the TRANSIT feature is currently limited to mitochondrion, chloroplast, thylakoid, cyanelle or microbody.

(3) *KW keyword.* There are 890 individual keywords in release 44 of Swiss-Prot. Entries contain on average 3.55 keywords with a maximum of 21 keywords. Three frequent keywords ('Direct protein sequencing', 'Complete proteome' and 'Pharmaceutical') were ignored as they have no biological meaning.

(4) *DR Database cross-references.* This record contains references to other general purpose databanks, such as the EMBL Nucleotide Database, the PDB, a repository of known three-dimensional structures and several others. These references do not provide detailed functional or structural annotation, but merely contain individual database IDs of the external databases. In addition, the DR record indicates the occurrence of protein sequence and structure motifs and domains as defined in PFAM, PRINTS, PROSITE, SCOP, ProDom, SMART and TIGRFAMs databases (reviewed by Liu and Rost, 2003), as well as in the InterPro resource which integrates all major signature databases (Mulder *et al*., 2005). Since the InterPro resource is essentially a superset of a large number of domain databases, we chose to use only InterPro domain assignments listed in the DR records to avoid redundancy. The release 44.0 of Swiss-Prot contains references to the total of 9101 distinct InterPro domains, or 2.06 per Swiss-Prot entry on average. InterPro IDs begin with the three letters IPR followed by six digits specifying the number of the corresponding InterPro signature (e.g. IPR002394).

(5) *OG Organelle.* This record indicates if the gene coding for a protein originates from the mitochondria, the chloroplast, the cyanelle, the nucleomorph or a plasmid. In this study we considered only entries mitochondrion, chloroplast, cyanelle and nucleomorph.

The five Swiss-Prot records described above contain nominal attributes that may take a fixed number of textual values. In addition, one may in principle consider numerical attributes describing protein features. Since association rule mining cannot be directly performed over numerical data, such numerical attributes are converted to nominal form by binning their values and assigning textual labels to each bin. In this work the only numerical attribute considered was amino acid chain length which was extracted from the Swiss-Prot entry ID line and binned over four intervals, namely short (denoted S, $<$120 amino acids), medium (M, 120–1000 amino acids), long (L, 1000–1500 amino acids), and very long (XL, $>$1500 amino acids).

Two types of Swiss-Prot entries were excluded from consideration—those containing protein fragments [as indicated by the word 'Fragment(s)' in the description line] and those containing the keyword 'Hypothetical protein' in the list of keywords. The remaining 125 642 entries were exported to a text file in which each protein sequence was represented by a single line containing the list of protein characteristics delimited by commas. For example, the human p53 protein (Swiss-Prot code P53_HUMAN) is represented by the following line.

> Apoptosis, VARSPLIC, Anti-oncogene, MOD_RES, Nuclear protein, length:M, Polymorphism, Phosphorylation, Zinc, Eukaryota, DNA-binding, IPR008967, Li-Fraumeni syndrome, Transcription regulation, Activator, DNA_BIND, Disease mutation, METAL, ZINC, 3D-structure, IPR002117, Alternative splicing, Metal-binding, Acetylation, Glycoprotein, IPR010991

This line contains 3 InterPro domain assignments, 17 keywords, 5 positional features, the item 'length:M' which stands for medium sequence length range (see above), and a taxonomic assignment (Eukaryota). Following the generally accepted terminology in the area of association rule mining, we will call each such feature an item, and a set of all or some features related to one protein an item set.

2.2 Extracting item sets from the PEDANT genome database

The PEDANT Genome Database (http://pedant.gsf.de) contains precomputed information resulting from bioinformatics analyses of publicly available genomes. The main vehicle for similarity searches is the PSI-BLAST algorithm (Altschul *et al*., 1997). This method is used for searches against the full non-redundant protein sequence databank as well as against a number of special datasets including the MIPS functional categories (see below) and the COG database (Tatusov *et al*., 2003). In addition, the detection of PROSITE (Sigrist *et al*., 2002), PFAM (Bateman *et al*., 2004) and BLOCKS (Henikoff *et al*., 1999) sequence motifs was performed. For those sequences that have significant matches in the Uniprot/Swiss-Prot database, the annotation of the respective entries is analyzed and keywords and enzyme classification are extracted. Structural categorization of gene products involves secondary structure prediction as well as PSI-BLAST searches against the sequences with known 3D structure as deposited in the PDB databank (Deshpande *et al*., 2005) and SCOP database of known structural domains (Andreeva *et al*., 2004). Other calculated or predicted structural features include molecular weight, pI, low complexity regions (Wootton, 1994), membrane regions (Krogh *et al*., 2001), coiled coils (Lupas, 1997) and signal peptides (Bendtsen *et al*., 2004).

Functional roles of gene products are described in terms of the manually curated hierarchical functional catalog developed by MIPS (FUNCAT; Ruepp *et al*., 2004). Each of the 16 main classes (e.g. metabolism, energy) may contain up to 6 subclasses. Correspondingly, the numeric designator of a functional class can include up to six numbers. For example, the yeast gene product YGL237c is attributed to the functional category 04.05.01.04, where the numbers, from left to right, mean transcription, mRNA transcription, mRNA synthesis and transcriptional control. An essential feature of FUNCAT is its multidimensionality, meaning that any protein can be assigned to multiple categories. Similarly, the SCOP database (Lo Conte *et al*., 2002) provides a hierarchical classification of protein structural domains. It has 11 main classes, designated by letters a through to j. The most prominent classes—a, b, c, d and e—correspond to all-α, all-β, α/β, $\alpha + \beta$, and multidomain proteins, respectively. A numeric designator of each SCOP fold always starts with a letter denoting its main class as described above and contains three numbers corresponding to three further levels of hierarchy (fold, superfamily, family). In this work both FUNCAT and SCOP designators were truncated to include only the two upper levels of hierarchy.

A typical description line extracted from PEDANT for association rule mining has the following form:

> COG0106, Alpha_Beta, PS00167, SCOP:c.1, complete proteome, COG0434, pI:L, FC:01.01, COG0159, length:M, COG0826, COG1646, COG0269, EC:4.2.1.20, FC:16.21, PF00290, hydro-lyase, complexity:low, carbon-oxygen lyase, tryptophan biosynthesis, lyase.

This protein, the *trpA* gene product from the *Acinetobacter sp.* ADP1 genome, is of medium length, has low pI and sequence complexity, and was assigned to five different COGs, SCOP fold c.1, EC number 4.2.1.20, and functional category 16.21. In addition, it contains one PROSITE (PS00167) and one PFAM (PF00290) motif, belongs to the α/β structural class and was annotated with the keywords hydro-lyase, carbon–oxygen lyase, tryptophan biosynthesis and lyase.

In total, we extracted from the PEDANT database annotation for 106 914 genes from 38 prokaryotic genomes uniformly using a rather stringent E-value threshold of 0.00001 for all similarity search methods.

2.3 Extracting rules from annotation databases

Files containing item sets generated from the Swiss-Prot and PEDANT databases served as input data to extract rules using the well established *Apriori* algorithm for association rule mining. The basic *Apriori* algorithm, described in detail by Agrawal and Srikant (1994), is designed to find frequent item sets by consecutive expansion of candidate item sets on every step based on the simple notion that all subsets of a frequent item set are also frequent. In this work we used this algorithm as implemented in a commercial software package Magnum Opus (Webb, 2000). Public domain implementations also exist (e.g. as a part of the Weka machine-learning workbench (http://www.cs.waikato.ac.nz/∼ml/weka/)), but are less efficient and not yet suitable for analyzing very large databases. If not specified otherwise, all rules with the coverage of at least 50 proteins and strength of at least 0.1 were retained for further analysis.

The application of Magnum Opus results in a file containing one rule per line. Each line lists the LHS and the RHS as well as several numerical characteristics of the rule delimited by commas. A typical rule line in the Magnum Opus output file looks as shown below:

> DNA_BIND & Activator, Transcription regulation, 0.009,1144,0.050, 6338,0.009,1144,1.000,19.89,0.0086,1086.5

Items in the LHS are joined by the '&' symbol, followed by the RHS and the list of numerical attributes of the rule, such as coverage, coverage count, RHS coverage, RHS coverage count, support, support count, strength, lift, leverage, leverage count. In this work we will only use 'coverage count', 'support count', and 'strength' (1144 and 1.000 in our example) to characterize rules generated.

The extracted rules were subjected to the following post-processing. Several items, namely all values of the highest-level taxon of the protein origin—Eukaryota, Bacteria, Archaea and Viruses, as well as the intervals of protein length—length:S, length:M, length:L and length:XL, were forbidden on the RHS.

2.4 Swiss-Prot release dynamics

In order to compare rule statistics for consecutive releases of the database we calculated association rules for the whole set of protein entries in every release separately as well as for the reduced set of entries shared by all releases. The latter sample was formed by all protein entries not annotated as hypothetical or a fragment in each release and had one descendant in all subsequent releases. If two protein entries had one common descendant, both entries were ignored. If a protein entry had more than one descendant we selected only one of the descendants at random. The final set of shared protein entries included 108 984 proteins. For every release we used annotation restricted to the vocabulary shared by all releases. All new annotation terms introduced after release 42.0 were ignored.

To reveal the corrections introduced by the Swiss-Prot staff we examined the protein entries that constituted rule exceptions and classified these entries as corrected if in a subsequent database release, either one of the items forming the LHS of the associated rule was deleted from the annotation or the item from RHS was newly introduced or altered.

2.5 Manual verification of rules

For manual verification of association rules we randomly selected a limited sample of all protein entries from Swiss-Prot and PEDANT that constituted exceptions from rules with strength in the range [0.97; 1.0) and, in the case of Swiss-Prot, were not corrected by the Swiss-Prot staff in subsequent database releases. Items of the annotation of these proteins mentioned in the LHS or the RHS of the rules were subjected to careful manual analysis by an experienced protein annotator according to the established procedures routinely used at MIPS for genome annotation (Mewes *et al*., 1997; Horn *et al*., 2004). These include assessment of similarity hits and predicted protein features as well as in-depth examination of literature describing experimental studies. The exception was classified as an error if one of the items of the LHS of the rule was assigned wrongly to a given protein entry, or the required item on the RHS of the rule was missing.

We calculated error rate for PEDANT as the fraction of exceptions classified as annotation errors among all manually verified exceptions. For Swiss-Prot the overall error rate was calculated as (the number of exceptions corrected in subsequent releases + manually verified error rate for non-synonymous rules $*$ number of uncorrected exceptions from non-synonymous rules $+$ manually verified error rate for synonymous rules $*$ number of exceptions from synonymous rules)/overall number of exceptions.

3 RESULTS AND DISCUSSION

3.1 Statistics of association rules in Swiss-Prot

Application of the *Apriori* algorithm to the item sets extracted from Swiss-Prot resulted in 302 459 rules with strength >0.1 and minimal coverage count 50. As expected, these rules vary greatly in terms of their coverage and strength. For example, the rule 'Alternative Splicing & Transmembrane ⇒ Eukaryota' extracted from Swiss-Prot has coverage count 1433, support count 1417 and strength 0.989 (≈1417/1433), indicating that there are 1433 proteins in Swiss-Prot with simultaneously assigned keywords 'Alternative splicing' and 'Transmembrane' and 1417 of them are of eukaryotic origin. The remaining 16 proteins originate from viruses. However, the rule 'Alternative splicing & Nuclear protein ⇒ Repressor' has 129 confirmations from 1288 covered proteins (support count = 129, coverage count = 1288, strength 129/1288 ≈ 0.1), implying that only a small fraction of all nuclear proteins subjected to alternative splicing are repressors, whereas these three keywords taken separately occur frequently among Swiss-Prot protein entries.

Strength distributions of Swiss-Prot rules for different values of minimal coverage count are shown in Figure 1. A prominent feature of these distributions is the presence of two distinct peaks in the regions of very weak and very strong rules, with rules in the medium strength range being relatively infrequent. A large number of weak rules (strength <0.2) originate from diverse combinations of frequent items, such as the majority of the Swiss-Prot keywords or features. These combinations are typically not wrong, but they do not represent typical associations between items. For example, the Swiss-Prot entry Q6W2J9 (BCoR protein from *Homo sapiens* functioning as transcriptional corepressor) contains the keywords 'Alternative splicing', 'Nuclear protein' and 'Repressor' and conforms to the rule 'Alternative splicing & Nuclear protein ⇒ Repressor'. It has repressor function, localizes in the nucleus, as in the case of the majority of transcription factors, and is subject to alternative splicing. But only a certain fraction of all repressors have multiple alternatively spliced isoforms, and thus this rule is not classified here as a biological regularity.

The other extreme on Figure 1 is constituted by very strong rules with strength values in the range roughly between 0.95 and 1.0. In particular, there are 63, 287, 1807, 7751 and 24 288 rules whose strength exactly equals 1.0 for the minimal coverage counts of 1000, 500, 200, 100 and 50, respectively. For example, all 1554 proteins annotated with the keyword 'G-protein coupled receptor' also have the keyword 'Transmembrane' whereas all 1904 proteins having the feature 'MITOCHONDRION' in the FT line also contain the keyword 'Transit peptide'.

In this study we specifically focus on the strong rules with the strength over a certain threshold (typically 0.95), but <1.0. These rules are nearly always fulfilled, but exceptions from them do occur in the database. For example, at the minimal coverage count of 50 there are 7396, 4956 and 4046 rules in the Swiss-Prot database that are not fulfilled exactly once, twice or three times. As argued above, such exceptions may constitute annotation errors which can be detected and corrected, or at least flagged, automatically.

Can strong rules occur by chance? To answer this question we studied the behavior of a database generated by randomly shuffling annotation items such that feature frequencies and the number of features for each protein were preserved. As seen in Figure 2a, the random distribution is characterized by complete absence of rules stronger than 0.88 and a much larger fraction of weak rules (strength below 0.2) than the real database. Notably, two unexpected peaks with strength 0.40 and 0.62 present on this curve reflect frequent occurrence of rules containing one of the three most frequent items in their RHS. More specifically, the left-hand peak of strength ∼0.40 almost completely (444 rules of 450, 98.6%) consists of rules with one of the two features—Bacteria or Eukaryota—in the RHS. The majority (221 of 259, 85.3%) of the rules corresponding to the peak at strength ∼0.62 include the feature 'length:M' (middle size range) in their RHS, whereas all remaining rules of the peak have 'Bacteria' or 'Eukaryota' in the RHS. These three features are the most frequent items in Swiss-Prot annotation among all the items considered in this study. Exclusion of rules with such RHS eliminated all rules

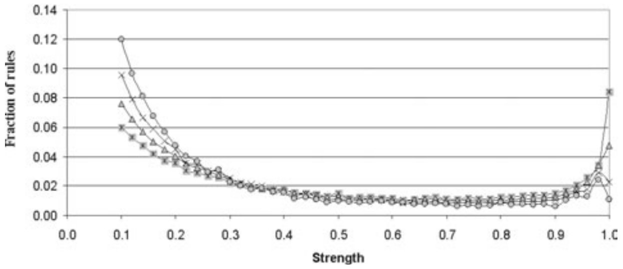

Fig. 1. Distribution of association rule strength in the Swiss-Prot database. The four curves correspond to sets of rules with minimal coverage 50 (–∗–), 100 (–△–), 200 (—×—) or 500 (–○–).

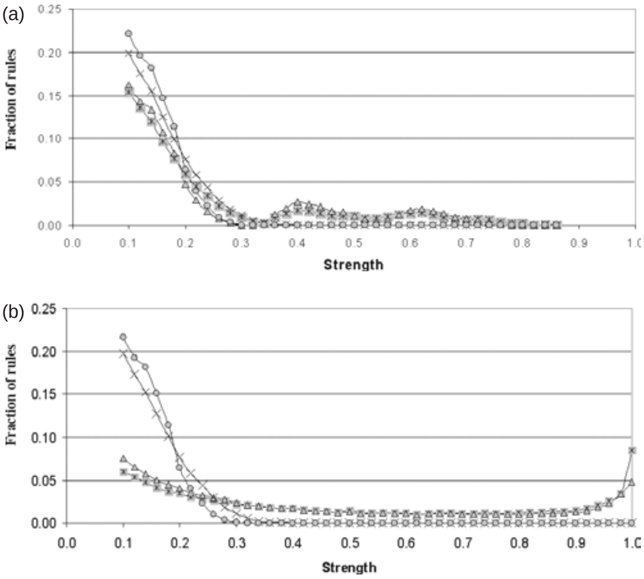

Fig. 2. (**a**) Distribution of association rule strength in the random database constructed based on Swiss-Prot (see Section 2). The first two curves correspond to the rules with minimal coverage 50 (–∗–) or 100 (–△–), the last two curves, (—×—, –○–), to the same sets after filtration (deletion of all rules with one of three most frequent items in their RHS). (**b**) Distribution of association rule strength in the real (Swiss-Prot) and random databases after filtration (see Section 2). The first two curves correspond to the rules with minimal coverage 50 (–∗–) or 100 (–△–) for the real Swiss-Prot database, the last two curves (—×—, –○–), to the corresponding rules of the random Swiss-Prot-based database.

stronger than 0.38 in the random database (Fig. 2a), but did not influence the real Swiss-Prot curve (data not shown). We thus conclude that the presence of strong, and indeed even medium-strength rules (with strength roughly over 0.4), is significantly non-random and cannot be explained by mere statistical effects caused by particular feature composition of the Swiss-Prot database. Such rules reflect meaningful associations between protein annotation attributes.

It should be noted that the rule strength distributions shown in Figure 2a were constructed without the post-processing usual for the real Swiss-Prot database (see Section 2). Application of such post-processing rules to the random database led to disappearance of the peaks at strength values 0.40 and 0.62 discussed above because the prominent items (Eukaryota, Bacteria and length:M) responsible for the frequent occurrence of these rules got eliminated from the RHS (Fig. 2b). The corrected curves for the random database should consequently be used as a comparative standard for Swiss-Prot since they are derived according to exactly the same procedure that was used for the real Swiss-Prot database.

We also studied the behavior of two non-protein databases, the Forest CoverType database and the Adult Database taken from the UCI Knowledge Discovery in Databases Archive (http://kdd.ics.uci.edu/) and found qualitatively similar tendencies in rule strength distributions (see Supplementary Text 1). We thus believe that the shape of rule strength distribution shown in Figure 1 may be typical for many (although possibly not all) sufficiently large databases containing significant amounts of non-random data. A detailed theoretical analysis of these phenomena is beyond the scope of the present work.

3.2 Types of Swiss-Prot-derived association rules

For better understanding of association rules derived from the Swiss-Prot annotation we classified them according to the composition of their LHS and RHS. Three types of LHS were considered: (1) those containing keywords only, (2) those with mixed annotation (i.e. LHS includes several items of different Swiss-Prot fields) and (3) those with LHS containing items from one and the same Swiss-Prot field different from keywords. Similarly, we distinguished the RHS types containing InterPro domains, a feature from the Swiss-Prot FT line, or a keyword. As seen in Figure 3, the most prominent rules are those inferring keywords and InterPro domains from mixed LHS. The fraction of the rules of the latter group grows significantly with rule strength, whereas the former group displays an opposite behavior. Also strongly dependent on strength are rules of the type 'mixed LHS to FT feature' which are characteristic for strong rules. All other rule types appear to be evenly distributed in terms of their strength. The total fraction of rules with RHS other than keyword, feature and InterPro is rather insignificant.

3.3 Dynamics of Swiss-Prot releases and error correction

Approximately every half year, a new release of the Swiss-Prot database is made available with novel protein entries added as well as some preexisting entries revised. At the same time, annotation vocabulary gets extended by new terms which may be introduced both to old and new entries. We analyzed the influence of these changes on the complete set of association rules. As seen in Figure 4, the total number of rules derived from Swiss-Prot steadily grows with the number of entries, from 232 556 for release 42.0 to 445 270 for release 47.0, mostly owing to introduction of new database entries. If one takes into account only the protein entries shared between all releases under study, the growth of the rule number becomes much slower, especially if only shared annotation vocabulary is considered (Fig. 4). In this case the slight increase is mostly because of the introduction of additional annotation items from the shared vocabulary to revised Swiss-Prot entries, or, in very rare cases, to elimination of certain items.

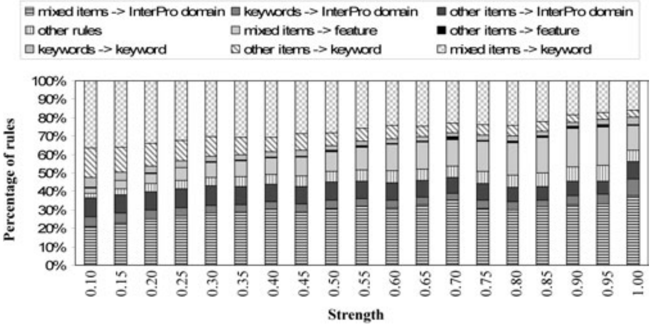

Fig. 3. Composition of rules in terms of their constituent items. For each interval of strengths the fraction of rules of a given type according to the Swiss-Prot field classification is given (see text for details).

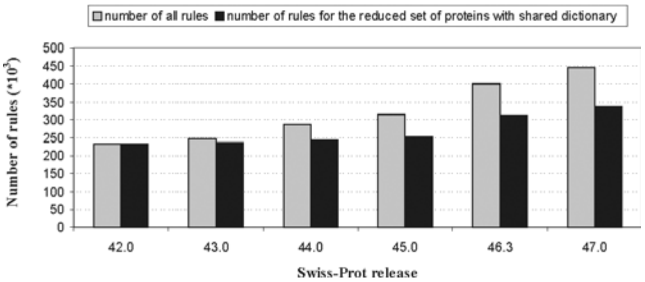

Fig. 4. Dynamics of rules the Swiss-Prot annotation. The reduced set of proteins is constituted by the 108984 Swiss-Prot entries common to all releases considered, with the annotation dictionary corresponding to Release 42.0.

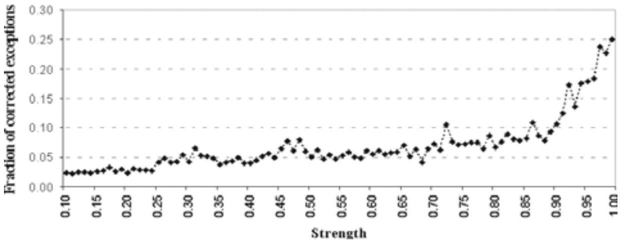

Fig. 5. Distribution of annotation corrections over rule strength ranges in Swiss-Prot releases 44.0–47.0.

As one progresses from one release to another, many exceptions to association rules get corrected. Figure 5 displays the strength distribution of such corrected rules. In line with our main assumption, the exceptions to the strongest rules get corrected more often. 23.5% of protein entries that constituted exceptions to the rules with strength in the range [0.97; 1) found in the Swiss-Prot release 44.0 were corrected until the release 47.0 (*see Section 2*), while the average percentage of corrected exceptions to the rules of any strength was 3.47. As an additional test, all 350 exceptions from randomly selected 149 rules in the strength range [0.97; 1) not corrected by Swiss-Prot staff were subjected to careful manual evaluation. We found that 24.7% of these exceptions indeed constituted annotation errors (see Supplementary materials). We thus estimate that the total percentage of exceptions associated with errors in strong rules is ∼41%.

As an additional test for errors we investigated the compatibility of annotation within protein entries by confronting the annotation synonyms from different annotation fields. All items whose names differed only in the letter cases were considered synonyms. In addition, the following trivial correspondences in feature and keyword fields were treated as synonyms: TRANSMEM and ‘Transmembrane’, VARSPLIC and ‘Alternative splicing’, LIPID and ‘Lipoprotein’, ZN_FING and ‘Zinc-finger’, CA_BIND and ‘Calcium-binding’, SE_CYS and ‘Selenocystein’, DNA_BIND and ‘DNA-binding’, METAL and ‘Metal-binding’, as well as CARBOHYD and ‘Glycoprotein’. There are also some correspondences between the KW and OG as well as between the FT, KW and OG fields. As one would expect, some of the rules containing synonyms of this kind in their LHS and RHS are perfect, meaning that, for example, the rule ‘Signal (keyword) implies SIGNAL (feature table)’ is true in all possible cases (has strength 1) as well as the reverse rule, ‘SIGNAL (feature table) implies Signal (keyword)’, which is also always true. However, quite often synonym pairs generate perfect rules only in one direction, e.g. ‘LIPID implies Lipoprotein’, whereas the reverse rule (‘Lipoprotein implies LIPID’) is satisfied only in 97% of cases. At the same time, many synonym pairs are not strongly coupled. For example, for proteins with documented alternative splicing, 4739 entries include both the keyword ‘Alternative splicing’ and the feature VARSPLIC, 564 entries include only the keyword ‘Alternative splicing’ and, finally, 5 entries include only VARSPLIC. All rules of the type ‘item implies its synonym’ were considered synonymous rules. We also identified a large number of nearly identical rule pairs differing by only one item in which synonyms are used.

Based on manual verification of the 15 strongest and 30 randomly selected synonymous rules not corrected by the Swiss-Prot staff with the total of 161 exceptions from rules within the strength range [0.97; 1) we estimate that ∼33% of these exceptions contain errors. Thus, taking into account synonymous rules, the overall fraction of exceptions associated with annotation errors is slightly over 43% for rules in the considered strength interval. Among all the manually verified Swiss-Prot exceptions classified as errors only 14% were owing to wrong assignment of one item in the LHS of the rule. All other errors result from omission of an RHS item, implying that Swiss-Prot annotation errors are typically caused by under-annotation. In most cases these errors are constituted by essentially trivial misannotations in which missing features can easily be reconstructed by similarity searches or computational predictions. For example, the TMHMM software predicts five transmembrane segments with the probability 1.0 in the entry P34641, but both the TRANSMEM item of the feature table and the keyword ‘Transmembrane’ are not specified. Sometimes correct information is present in free-text annotation but not in one of the formalized fields. For example, nuclear localization of the protein Q9U405 is indicated in ‘Comments’ based on the literature reference with the PubMed ID 10683177, but the keyword ‘Nuclear protein’ is omitted.

As a less trivial example, translation initiation factor eIF-2B α subunit (P14741) (synonym: transcriptional activator GCN3) regulates translation of GCN4 and represents the only exception from the rule ‘DNA-binding & Activator ⇒ Transcription regulation’. Detailed analysis of this protein failed to produce any evidence of its capability to bind DNA. It does not appear to possess a DNA-binding domain, and no experimental confirmation was found in the literature. Consequently, this exception was classified as an LHS error (erroneous assignment of the item from the LHS of the rule). Another example is Q47689 associated with one of the four exceptions from the rule ‘length:M & IPR002293 ⇒ Amino-acid transport’. The RHS keyword was omitted in the annotation of this entry, whereas this protein’s involvement in the transport of sulfur amino acids is documented in the literature (PMID: 9882684). This case was classified as an RHS error (or the omission of the RHS item).

Another interesting issue is the proteins associated with the largest number of exceptions from strong rules. The most ‘disobedient’ protein entry in Swiss-Prot is P00545, tyrosine-protein kinase transforming protein fms. Its annotation contains exceptions to 409 different association rules in the strength range between 0.95 and 1, with 286 of them attributed to the fact that this protein is the only non-precursor (consequently, without a signal peptide) among many tyrosine kinases with very similar annotation; thus, this case does not constitute an error. Furthermore, 104 rules contained the feature ‘ACT_SITE’ (enzymatic active site) in their RHS that was erroneously omitted in annotation of this entry, although it can easily be reconstructed by similarity-based annotation transfer (data not shown). The second and third most disobedient proteins are Q9LYN8, a precursor of leucine-rich repeat receptor protein kinase EXS, which contradicted 321 rules, and P42159 (class II receptor tyrosine kinase) which is an exception to 305 rules, respectively. Q9LYN8 contradicts 286 rules with ‘N-LINKED’, ‘CARBOHYD’ or ‘Glycoprotein’ in their RHS, features that can be reconstructed by similarity-based methods. Furthermore, 40 other rules associated with Q9LYN8 having exceptions contain the InterPro domain IPR001245 which was falsely (according to InterProScan, http://www.ebi.ac.uk/InterProScan/) assigned to this protein entry up to release 47.0. P42159 is the only protein annotated in the mature form in its group. Consequently, 214 rules with the keyword Signal or feature SIGNAL in their RHS and the exception in this protein are correct, whereas the remaining 91 rules for P42159 lead to keyword ‘Glycoprotein’ that have been omitted in its annotation up to the release 46 of Swiss-Prot.

Finally, what is the nature of the exceptions from strong rules that are not annotation errors? Quite often such exceptions stem from exotic peculiarities of one member in a protein family, exemplified by the protein entry P00545 discussed above. In some other cases they reflect real biological regularities which do in fact have a minor number of exceptions. This is the case for the rule involving the two keywords ‘Ubiquinone ⇒ NAD’, exceptions from which are constituted by only 5 of all 1014 proteins interacting with ubiquinone that use FAD as a cofactor instead of NAD.

3.4 Estimating the level of annotation errors in PEDANT

The above findings, based on the in-depth analysis of Swiss-Prot annotation, provide justification for our strategy to flag exceptions from strong rules as potential annotation errors. Application of the same procedures to the automatically generated PEDANT annotation revealed that statistical properties of the association rules gleaned from the PEDANT database are similar to that of Swiss-Prot (Fig. 6).

Table 1. Estimation of the number of errors

Strength interval	Rules (Sw-Pr)	Exceptions (Sw-Pr)	Errors (Sw-Pr), estimate	Rules (PEDANT)	Exceptions (PEDANT)	Errors (PEDANT), estimate
0.99-1.0	4032	5474	2772 (1368[a])	17 426	25 694	18 438
0.98-0.99	7391	14 597	7099 (3302[a])	38 724	72 658	47 467
0.97-0.98	4024	18 634	6814 (4431[a])	16 520	72 259	49 425

The number of errors was estimated based on the extrapolation from the limited randomly selected sample of rules: see the text for details.
Sw-Pr: Swiss-Prot.
[a] Are already corrected by Swiss-Prot staff.

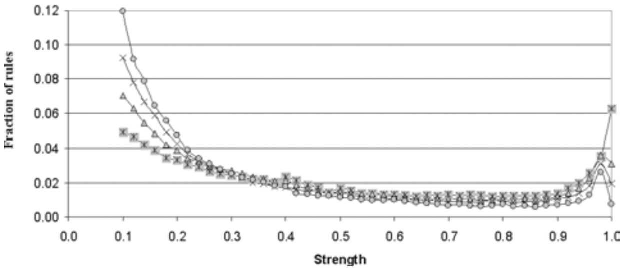

Fig. 6. Distribution of association rule strength for the PEDANT database. The four curves correspond to rules with minimal coverage 50 (-✳-), 100 (-△-), 200 (—×—) and 500 (-○-).

Here too, strong enrichment of very weak and very strong rules was observed, with rules in the intermediate strength range being relatively rare. Compositional breakdown of PEDANT rules (Fig. 7) revealed that functional category assignments play an increasingly important role as the strength of rules grows. Interestingly, most of the FUNCAT inferences in the medium range of rule strength are the result of mixed items in LHS (LHS includes several items of different Swiss-Prot fields), whereas the strongest rules typically derive a functional category from one or more other functional categories. This behavior is the direct consequence of the FUNCAT structure which, by design, is multidimensional such that each gene can be assigned to multiple categories (Ruepp *et al.*, 2004). Rules inferring BLOCKS and PFAM domains in the RHS occur only in low strength range, implying their marked independence from other types of annotation as well as from each other.

Since PEDANT annotation typically never gets corrected manually (except for occasional in-house genome annotation efforts, see Horn *et al.*, 2004), and there is thus no release dynamics as in the case of Swiss-Prot, the only way to estimate the amount of errors in the strong rules with exceptions is by manual verification. We analyzed a randomly selected sample of 144 rules in the strength range [0.97; 1). About a half of the sample was selected at random, whereas the second half was selected among rules showing at least one exception associated with a protein contained in Swiss-Prot to ensure that this protein is reasonably well documented. To reduce the manual effort, for rules with more than 10 exceptions we curated carefully only a subset of exceptions, unless they were associated with Swiss-Prot proteins. The overall number of curated exceptions was 330. The total fraction of exceptions classified as errors in PEDANT was close to 68%. The estimated number of errors for each considered strength interval is presented in Table 1.

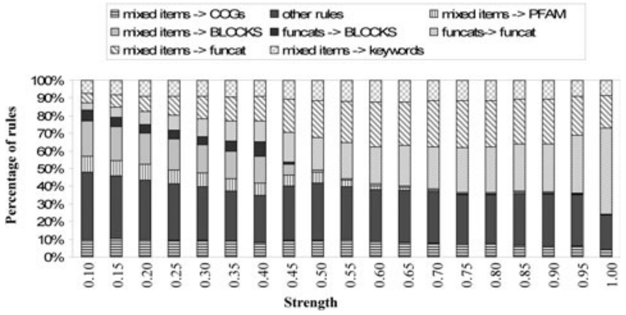

Fig. 7. Composition of rules in terms of their constituent items in the PEDANT database. For each interval of strengths the fraction of rules of a given type according to the PEDANT field classification is given (see the text for details). All rule categories involving $>5\%$ of all rules in the given strength interval are accounted for separately, with all other rules shown as ‘others’.

In contrast to Swiss-Prot, 30% of all errors revealed in PEDANT by manual curation were because of omission of an RHS item. Other errors resulted from false assignment of the items in LHS of the rules, or over-annotation. The list of all manually verified cases is available in Supplementary materials.

4 CONCLUSIONS

We have developed a general methodology called ARIA for improving the quality of biological data based on the notion that exceptions from strong association rules derived from annotation items often point to errors. In contrast to the Xanthippe system (Wieser *et al.*, 2004), such exceptions are sought in the entire database, and not within individual protein families. This new approach is primarily made possible by the much higher computational efficiency of the *Apriori* algorithm compared with the C4.5 method used by Wieser *et al.* which scales roughly as a cube of the number of examples (Cohen, 1995) and would require years of CPU time to process databases with hundreds of thousands of entries such as PEDANT. Higher efficiency of association rule mining also allowed us to consider essentially all biologically meaningful annotation items whereas the Xantippe system primarily operates with rules involving organism and domain information in their LHS to derive keywords and protein names in the RHS. The sets of rules generated by ARIA and Xanthippe are thus fundamentally different as are their application domains. Xanthippe is well suitable for a highly organized and curated database, such as Swiss-Prot, where reliable core annotation exists, whereas ARIA may be applied to large automatically generated collections of data where no gold standard of correctness is available.

ACKNOWLEDGEMENTS

We would like to thank Hans-Werner Mewes and Philip Wong for a careful reading of the manuscript and many useful comments, and Stefan Kramer for his valuable advice. This work is conducted in the framework of the BioSapiens project funded by the European Commission FP6 Programme, under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LHSG-CT-2003-503265. M.G. is partially supported by grants from the Howard Hughes Medical Institute (550003), the Ludwig Institute of Cancer Research (CRDF RB0-1268), the Russian Science Support Fund and the Russian Academy of Sciences (program 'Molecular and Cellular Biology').

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. and Srikant,R. (1994) Fast algorithms for mining association rules. *Proc. of the 20th International Conference on Very Large Data Bases*, 487–499.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–402.

Andrade,M.A. *et al*. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bateman,A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Becquet,C. *et al*. (2002) Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biol.*, **3**, RESEARCH0067.

Bendtsen,J.D. *et al*. (2004) Improved prediction of signal peptides: SignalP 3.0. *J.Mol.Biol.*, **340**, 783–795.

Bodenreider,O. *et al*. (2005) Non-lexical approaches to identifying associative relations in the gene ontology. *Pac.Symp.Biocomput.*, 91–102.

Bork,P. and Bairoch,A. (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet.*, **12**, 425–427.

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

Bult,C.J. *et al*. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science*, **273**, 1058–1073.

Cohen,W.W. (1995) Fast effective rule induction. *Proc. of the 12th International Conference on Machine Learning*, 115–123.

Cole,S.T. *et al*. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537–544.

Creighton,C. and Hanash,S. (2003) Mining gene expression databases for association rules. *Bioinformatics*, **19**, 79–86.

Deshpande,N. *et al*. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.

Devos,D. and Valencia,A. (2000) Practical limits of function prediction. *Proteins*, **41**, 98–107.

Eisenhaber,F. and Bork,P. (1999) Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, **15**, 528–535.

Fleischmann,W. *et al*. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.

Frishman,D. *et al*. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.

Gaasterland,T. and Sensen,C.W. (1996) MAGPIE: automated genome interpretation. *Trends Genet.*, **12**, 76–78.

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico.Biol.*, **1**, 55–67.

Gattiker,A. *et al*. (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput.Biol.Chem.*, **27**, 49–58.

Gilks,W.R. *et al*. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.

Heinig,M. and Frishman,D. (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.*, **32**, W500–W502.

Henikoff,S. *et al*. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.

Horn,M. *et al*. (2004) Illuminating the evolutionary history of chlamydiae. *Science*, **304**, 728–730.

Hu,Z.Z. *et al*. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.

Kaplan,N. *et al*. (2003) PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res.*, **31**, 5617–5626.

Karp,P.D. *et al*. (1999) Eco Cyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **27**, 55–58.

Kasukawa,T. *et al*. (2003) Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res.*, **13**, 1542–1551.

Kretschmann,E. *et al*. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J.Mol.Biol.*, **305**, 567–580.

Kunin,V. and Ouzounis,C.A. (2005) Clustering the annotation space of proteins. *Bioinformatics*, **6**, 24.

Liu,J. and Rost,B. (2003) Domains, motifs and clusters in the protein universe. *Curr. Opin. Chem. Biol.*, **7**, 5–11.

Lo Conte,L. *et al*. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.

Lupas,A. (1997) Predicting coiled-coil regions in proteins. *Curr.Opin.Struct.Biol.*, **7**, 388–393.

Mewes,H.W. *et al*. (1997) Overview of the yeast genome. *Nature*, **387**, 7–65.

Michailidis,G. and Shedden,K. (2003) The application of rule-based methods to class prediction problems in genomics. *J.Comput.Biol.*, **10**, 689–698.

Mulder,N.J. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **1**, D201–D205.

Overbeek,R. *et al*. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.

Peterson,J.D. *et al*. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res.*, **29**, 123–125.

Riley,M.L. *et al*. (2005) The PEDANT genome database in 2005. *Nucleic Acids Res.*, **33**, D308–D310.

Ruepp,A. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.

Satou,K. *et al*. (1997) Finding association rules on heterogeneous genome data. *Pac.Symp.Biocomput.*, 397–408.

Sigrist,C.J. *et al*. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief.Bioinform.*, **3**, 265–274.

Smith,R.F. (1996) Perspectives: sequence data base searching in the era of large-scale genomic sequencing. *Genome Res.*, **6**, 653–660.

Tatusov,R.L. *et al*. (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli. Curr.Biol.*, **6**, 279–291.

Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *Bioinformatics*, **4**, 41.

Webb,G. (2000) Efficient search for association rules. *Proc. of International Conference on Knowledge Discovery and Data Mining*, 99–107.

Wieser,D. *et al*. (2004) Filtering erroneous protein annotation. *Bioinformatics*, **20**, Suppl 1, I342–I347.

Wilson,C.A. *et al*. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J.Mol.Biol.*, **297**, 233–249.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput.Chem.*, **18**, 269–285.

Wu,C.H. *et al*. (2003) Protein family classification and functional annotation. *Comput.Biol.Chem.*, **27**, 37–47.

Xie,H. *et al*. (2002) Large-scale protein annotation through gene ontology. *Genome Res.*, **12**, 785–794.

Yu,G.X. (2004) Ruleminer: a knowledge system for supporting high-throughput protein function annotations. *J.Bioinform.Comput.Biol.*, **2**, 615–637.

Zhang,C. and Zhang,S. (2002) Association rule mining. Models and Algorithms. In Carbonell,J.G. and Siekmann,J. (eds.), *Lecture Notes in Artificial Intelligence*, 2307. Springer, Berlin.