Prediction of Metabolic Pathways Involvement in Prokaryotic UniProtKB Data by Association Rule Mining

Imane Boudellioua\textsuperscript{1*}, Rabie Saidi\textsuperscript{2†}, Maria Martin\textsuperscript{2}, Robert Hoehndorf\textsuperscript{1} and Victor Solovyev\textsuperscript{1}

\textsuperscript{1}Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia, 23955 and \textsuperscript{2}European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, UK

ABSTRACT

The widening gap between known proteins and their functions has encouraged the development of methods to automatically infer annotations. Functional annotation of proteins encoded in newly sequenced genomes is expected to meet the conflicting requirements of providing as much comprehensive information as possible while avoiding erroneous functional assignments and over-predictions. This trade-off imposes a great challenge in designing intelligent systems to tackle the problem of automatic protein annotation. In this work, we present a system that utilizes rule mining techniques to predict metabolic pathways in prokaryotes. The resulting knowledge represents predictive models explaining pathway involvement of UniProtKB entries. We carried out an evaluation study of our system performance using semantic similarity and cross-validation technique. We found that it achieved a very high accuracy of pathway identification with an F1-measure of 0.987 and AUC of 0.99. Then, our prediction models were successfully applied on 4.6 million UniProtKB/TrEMBL reference proteome entries of prokaryotes. As results, 551,418 entries were covered, where 371,265 of them lacked any previous pathway annotations.

INTRODUCTION

One of the central research goals of systems biology is modelling various biological processes. Elucidation of chemical reactions and pathways is one of the challenging problems in this field. A biological pathway is formed by a series of chemical reactions catalyzed by enzymes within a cell. Some of the most common biological pathways are those associated with metabolism, regulation of gene expression and transmission of signals. A metabolic pathway involves the step-by-step modification of an initial molecule to form another product. The resulting product can be stored by the cell, secreted, used immediately, or used to initiate another metabolic pathway. An example of a metabolic pathway is the cellular respiration equation where glucose is oxidized by oxygen to produce ATP, adenosine triphosphate \cite{1}. Pathways play a key role in advanced studies of functional genomics. For instance, identifying pathways involved in a disease may lead to effective strategies for diagnosing, treating and preventing diseases. Moreover, by comparing the behaviour of certain pathways between a healthy person and a diseased person, researchers can discover the roots of the disorder and use the information gained from pathways to develop new and better drugs \cite{2,3,4}. It is increasingly clear that mapping dysregulated pathways associated with various diseases is crucial to fully understand these diseases \cite{5}. In addition, pathways are often conserved, thus studying their interactions in model organisms may help elucidate cellular response mechanisms in other organisms.

The widening gap between known proteins and their functions has encouraged the development of methods to automatically infer annotations. Functional annotation of proteins encoded in newly sequenced genomes is expected to meet the conflicting requirements of providing as much comprehensive information as possible while avoiding erroneous functional assignments and over-predictions. This trade-off imposes a great challenge in designing intelligent systems to tackle the problem of automatic protein annotation. Hence, the need for automated methods is more than urgent to help with increasing the annotation coverage, detect inconsistencies and provide seeds for manual curation. There are several approaches proposed in the literature for such a task. A quite promising approach is to apply knowledge discovery and data mining techniques to predict some protein features based on a set of known data. Such rule-based methods provide rich automatic functional annotation and aid in performing integrity checks. For instance, Kretschmann et al \cite{6} applied C4.5 data mining algorithm \cite{7} to gain knowledge about the Keyword annotation from UniProtKB/Swiss-Prot \cite{8}. Rule-base \cite{9} is another semi-automatic annotation system run on UniProtKB/TrEMBL \cite{8}. It uses the annotation of UniProtKB/Swiss-Prot entries that possess a set sequence signatures to annotate UniProtKB/TrEMBL entries that contain the same signature, fundamentally with keywords and comments. Other examples of systems for automatic annotation that has annotations integrated in UniProtKB/TrEMBL are HAMAP \cite{10}, EDIT to UniProtKB/TrEMBL \cite{11}, and PIR \cite{12}.

\*To whom correspondence should be addressed. Email: imene.boudellioua@kaust.edu.sa
\†To whom correspondence should be addressed. Email: rsaidi@ebi.ac.uk

© The Author(s)
Literature-based manual annotation of pathways cannot scale to thousands of recently sequenced genomes. Therefore, computational methods are needed for the identification and the mapping of pathways. We suggested that association rule mining could be used effectively as a computational method for pathway prediction. Association rule mining is a technique originated from the analysis of data on market baskets. The objective is to locate trends by means of association relationships and correlations within a dataset. Essentially, the aim of such analysis is to discover a set of useful rules that are shared among a percentage of the dataset.

An association rule is an implication expression of the form $X \implies Y$ where $X$ and $Y$ are disjoint itemsets. Association rule mining was used in several applications of bioinformatics including mining gene expression data (13) and identifying related GO terms (14). Moreover, association rule mining was used to improve the quality of automatically generated annotations by detecting anomalies in annotation items (15). In the context of automated protein annotation, we consider association rules in the form of many-to-one implications. If an annotation satisfies a rule with accepted quality of metric values, then we hypothesize that such a rule may reflect a biological regularity. An example of an association rule in a database of annotated proteins is: “Nuclear localization $\implies$ Origin: eukaryota”, which describes that every protein which is annotated as localized in nucleus has a eukaryotic origin (15).

One of the very first pathway prediction systems was PathFinder (10) which aims to identify signaling pathways in protein-protein interaction networks. It extracts the characteristics of known signal transduction pathways and their functional annotations in the form of association rules. There are also tools that predict biodegradations pathways such as META (17), CATABOL (18) and UM-PPS (19). In addition, relative reasoning has been used in the prediction of mammalian detoxification pathways in order to limit combinatorial explosion (20). Association rule mining was used in (21) to mine the rules linking enzymes and domains of various genes. The Pathologic component of the Pathway Tools software (22) is the state of the art in pathway prediction. It performs prediction of metabolic pathways in sequenced and annotated genomes using MetaCyc as the reference metabolic pathway database. One of the limitations of this system is extendibility due to the fact that its logic is hardcoded. That is because Pathologic incorporates rules and heuristics developed using feedback from biologists to improve the accuracy of the predictions. Another limitation is becoming more apparent with the growth of MetaCyc size, resulting in PathoLogic suffering from more false positive pathway predictions. In addition, the algorithm is limited to Boolean predictions with a coarse measure of prediction confidence making it difficult to filter the predictions with a probability cutoff. A comparative analysis was conducted (23) revealing that some machine learning approaches performed better than Pathologic in pathway prediction.

In this work, we are tackling the problem of pathway prediction in the context of metabolism. We introduce a pathway prediction system that can be used to enhance the quality of automatically generated annotations as well as annotating proteins with unknown function. The pathway prediction system utilizes data from UniProtKB/Swiss-Prot (24), which is a high quality manually annotated and non-redundant protein sequence database containing experimental results, computed features and scientific conclusions. Our pathway prediction system uses InterPro (24) signatures and organism taxonomy attributes of UniProtKB/Swiss-Prot entries to predict metabolic pathways associated with each protein entry. The association algorithm, Apriori (25), is used at the learning phase to identify significant relationships between the attributes of UniProtKB/Swiss-Prot annotations. Furthermore, we use a filtering method, SkyRule (26, 27), to select the best rules based on a combination of several interestingness metrics. This approach adopts the notion of dominance between association rules to discover the most interesting ones without favouring or excluding any measure. We finally present an evaluation study on UniProtKB prokaryotic entries to demonstrate the performance, capability and robustness of our approach.

MATERIALS AND METHODS

The system is designed to solve the following problem: given a set of UniProtKB/Swiss-Prot entries, generate models for pathway prediction using rule mining techniques. As any machine learning system, the system has two major phases, the learning phase, and the applying phase. The learning phase involves the training and testing on UniProtKB/Swiss-Prot input data while the applying phase involves applying the generated pathway models on the respective UniProtKB/TrEMBL entries.

Dataset preparation

The current status in UniProtKB for prokaryotes is summarized in Table 1 which shows that InterPro covers over 70% of prokaryotic entries in UniProtKB/TrEMBL. This high coverage will aid us in the learning process by using InterPro signatures identifiers as an attribute type for the prediction models. Firstly, the system loads all prokaryotic protein entries from UniProtKB/Swiss-Prot. After that, we filter out the entries that do not contain pathway functional annotation as an attribute. Moreover, in order to maintain data quality, the system only considers entries with manual assertion evidences. An evidence is described by a code from the Evidence Codes Ontology (ECO) (28). ECO is a controlled vocabulary of terms that describe scientific evidences in the realm of biological research. ECO can be used to document both the evidence that supports a scientific conclusion and how that conclusion was recorded by a scientist. The evidence types that are used in UniProtKB for manual assertion are described in Table 2.

Then, the system extracts the necessary information from the loaded UniProtKB/Swiss-Prot entries using metabolic

| Table 1. Current status in UniProtKB for prokaryotes |
|-----------------------------------------------|
|                  | Swiss-Prot | TrEMBL  |
|------------------|------------|---------|
| Total number of entries                          | 351,482    | 30,443,697 |
| Percentage of entries with pathway annotations   | 30.44%     | 5.42%   |
| Percentage of entries with InterPro annotations  | 98.71%     | 74.46%  |

As of June 2015.
Table 2. Considered evidences for pathway annotation in UniProtKB/Swiss-Prot

| Evidence ID     | Evidence Label                              | Description                                                                 |
|-----------------|---------------------------------------------|-----------------------------------------------------------------------------|
| ECO:0000269     | Experimental evidence                       | Manually curated information for which there is published experimental evidence. |
| ECO:0000303     | Non-traceable author statement evidence     | Manually curated information that is based on statements in scientific articles for which there is no experimental support. |
| ECO:0000305     | Curator inference evidence                  | Manually curated information which has been inferred by a curator based on his/her scientific knowledge or on the scientific content of an article. |
| ECO:0000250     | Sequence similarity evidence                | Manually curated information which has been propagated from a related experimentally characterized protein. |
| ECO:0000255     | Sequence model evidence                     | Manually curated information which has been generated by the UniProtKB automatic annotation system or by various sequence analysis programs that are used during the manual curation process and which has been verified by a curator. |
| ECO:0000244     | Combinatorial evidence                      | Manually curated Information inferred from a combination of experimental and computational evidence. |

pathways as targets, and InterPro signatures and organism taxonomic lineages as attributes. We ended up with a total of 95,822 entries. The attribute types and target type representation in UniProtKB are described as follows:

- **Target Type: Metabolic Pathway Comment**
  Represented as a structured hierarchy of controlled vocabulary where each process is split up into super-pathway, pathway and/or sub-pathway. When known, the step number mediated by the protein within the pathway is also indicated. On the other hand, when the metabolic pathway is not fully known, only the super-pathway and pathway labels are indicated. Moreover, a protein can participate in different pathways or in different steps of the same pathway. An example of a fully known pathway representation in UniProtKB for the protein Anthranilate synthase component 1 is: L-tryptophan biosynthesis; L-tryptophan from chorismate: step 1/5.

- **Attribute Type: InterPro Signature ID**
  The InterPro signature IDs are cross-referenced from InterPro database, which is an integrated resource of protein families, domains and functional sites. InterPro provides functional analysis of proteins by classifying them into families and domains. Protein signatures are combined from 11 member databases into a single searchable resource. A protein entry could be associated with one or more InterPro IDs. An example of InterPro IDs associated with the protein Anthranilate synthase component 1 is: IPR005801 (a domain), IPR019999 (a family), IPR006805 (a domain), IPR005256 (a family), and IPR015890 (a domain).

- **Attribute Type: Taxonomic Lineage**
  The taxonomic lineage is considered as an attribute. UniProtKB Taxonomy is based on the NCBI taxonomy database and is organized in a tree structure that represents the taxonomic lineage. It contains the taxonomic hierarchical classification lineage of the source organism. It lists the nodes as they appear top-down in the taxonomic tree, with the more general grouping listed first. An example of taxonomic lineage representation for protein Anthranilate synthase component 1 is: Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas.

The extracted list of attributes and targets for each loaded entry will form an itemset. Table 3 describes some examples of the forms of itemset that are associated with some UniProtKB/Swiss-Prot protein identifiers.

**Generation of association rules**

The prepared itemsets form the input of Apriori algorithm proposed by Agarwal and Srikant. Apriori, a bottom up approach, is one of the well known association rule mining techniques. Apriori discovers all significant association rules that represent trends in a large database of entries or transactions. We use the Apriori implementation developed by Borgelt. This implementation uses a prefix tree to organize the support counters and a doubly recursive procedure to process the transaction to count the support of candidate item sets. Apriori could be configured to provide different evaluating measures for each generated association rule. Each evaluation measure tries to quantify the dependency between the antecedent and the consequent of the rule. Table 4 displays the threshold values we considered for Apriori. We use a combination of four measures to effectively minimize false positives and the number of rules generated out of pure randomness. These chosen metrics are:

- **Support**
  According to the support of an association rule \( R = A \text{ AND } B \implies C \) (noted \( \text{supp}(R) \)) is the support of the set \( S = A,B,C \) which is defined by the absolute or relative number of cases in which the rule is correct. In the prior example, it is the number of cases where the occurrence of item \( C \) follows from the occurrences of items \( A \) and \( B \). However, this definition may cause some problems if multiple evaluation measures are used. Hence, we will adopt the definition proposed by which describes the support of an association rule as the absolute or relative number of cases in which it is applicable, in other words, in which its antecedent part holds. Unlike the original definition,
the support in this case provides a useful statistical meaning of the support of a rule and its confidence (29).

- **Confidence**

Confidence metric is used to measure the quality of a particular association rule. More intuitively, it measures the reliability of the inference made by a rule. Introduced in (30), the confidence of an association rule \( R = X \implies Y \) (noted \( \text{conf}(R) \)), where \( X \) and \( Y \) are itemsets, is calculated as the support of the set of all items that appear in the rule divided by the support of the antecedent set. More formally,

\[
\text{conf}(R) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}
\]

In other words, the confidence of a rule is the number of cases in which the rule is correct relative to the number of cases in which it is applicable. A high confidence ratio indicates that its associated rule has a high probability of correctness and thus makes correct predictions. It is worth mentioning that rules with high confidence may occur by chance. Determining whether the antecedent and the consequent are statistically independent is used to detect such spurious rules. One of the measures that could assist with this is the lift value.

- **Lift Value**

The lift value, or confidence quotient is basically the quotient of the posterior and the prior confidence of an association rule. Mathematically speaking, the lift of a rule \( R = X \implies Y \) is:

\[
\text{lift}(R) = \frac{\text{conf}(X \implies Y)}{\text{conf}(\emptyset \implies Y)}
\]

where \( \text{supp}(\emptyset) \), is the number of transactions in the database. Lift measures how far from independence the antecedent and consequent are. A lift value equals to one implies that the antecedent and consequent are independent and that the support of a rule is expected considering the supports of its components which renders such rule not interesting. If the resulting lift value is greater than one, this implies that the presence of the antecedent items raises the confidence. Likewise, if the lift value is less than one, then the presence of the antecedent items lowers the confidence.

- **p-Value**

In statistics, p-value is used to measure the statistical significance of a result. Several statistical tests have been used to calculate p-values of association rules (33, 34). Here, we adopt the p-value computed from G-Statistic. Under independence, the G-statistic also has a \( \chi^2 \)-distribution. The chi-squared statistic can be used to calculate a p-value by comparing the value of the statistic to a \( \chi^2 \)-distribution. That is, the p-value is computed as the probability that the \( \chi^2 \)-value of an association rule can be observed by chance assuming that the antecedent and the consequent of the rule are independent (35). This measure does not assess the strength of correlation between antecedent and consequent. It only assists in deciding about the independence of the antecedent and the consequent in a rule. P-value is used to infer how likely the occurrence of the rule is due to a systematic effect instead of pure random chance. If a rule has low p-value, then this rule has a low chance to occur if its two sides are independent. Given that this rule is observed in the data, then its two sides are unlikely to be independent, and thus, the association between them is likely to be real. On the other hand, high p-value means that the rule has a high chance to occur even if there is no association between its two sides. Such rules should be discarded.

Given our selected dataset and parameters, Apriori successfully generated 1,125,911 rules in total. Some examples of rule representation along with their quality metrics are shown in Table 5.

### Selection of association rules

Apriori generates a large number of rules especially for large databases (mining irrelevant rules, etc). The expert is unable to determine the most interesting association rules and make decisions based on these rules. Hence, we need an efficient evaluation of rules to select those that are actually relevant.

**Table 4.** Apriori threshold values considered for the system

| Parameter                                      | Value |
|------------------------------------------------|-------|
| Minimum number of items per association rule   | 2     |
| Minimum support of an item set (absolute number of transactions) | 30    |
| Minimum confidence of a rule as a percentage   | 90%   |
The generated list of rules will be analyzed by SkyRule [26, 27] to select the best rules based on their respective evaluation measures. The SkyRule operator selects the rules that are supposed to be the most interesting ones accounting for several measures. In our case, the interestingness measures considered are support, confidence, lift, and p-value that where discussed in the previous subsection. SkyRule approach adopts the notion of dominance and comparability between association rules to discover interesting association rules without favoring or excluding any measure among the used ones. SkyRule also eliminates the need of the threshold value specification through the use of dominance relationship. The dominance relationship which is the corner stone of the SkyRule operator is applied on rules and can be presented as follows: a rule \( r \) is said to be dominated by another rule \( r' \), if for all used measures, \( f \) has better measures than \( r \), hence more relevant. Moreover, a rule \( x=(A \Rightarrow B) \) is said to be comparable to rule \( x'=(C \Rightarrow D) \) if \( B=D \) AND \( A \cap C \neq \emptyset \). This semantic comparability helps to decide if the considered association rules are semantically related (i.e. comparable). Comparability defines a kind of semantic relationship between rules and restrict the use of dominance. Concretely, the dominance between two rules must be applied only if a semantic relationship exists between them. SkyRule utilizes the concepts of dominance and comparability to select a family of inter-independent and statistically relevant rules, we term them representative rules.

At first, for each rule, SkyRule will compute the Euclidean distance to the normalized ideal metrics (1.0 for all four quality metrics we have). After that, SkyRule will sort the set of rules in a descending order by their associated distances. The first representative rule to be selected by SkyRule will be the rule which has metrics closest to the normalized ideal metrics. This rule is proven to be undominated by any other rule of the set of candidate rules [26, 27]. SkyRule will then discard all the rules that are comparable to this representative rule from the set of candidate rules. Essentially, SkyRule will filter out rules so that only undominated and incomparable rules are maintained. Out of all the rules generated by Apriori, SkyRule selected 1,065 rules as representative rules for the prediction models.

**Construction of prediction models**

The chosen rules by SkyRule will be aggregated to create a model for each pathway target. For example, if we have two rules of the form \( A \Rightarrow C \) and \( B \Rightarrow C \), then we aggregate them to a single rule such that \( A \cup B \Rightarrow C \). The set of the aggregated rules will build the final prediction models that are described in a human readable format. This process resulted in 331 prediction models. Those prediction models are applied on UniProtKB/TrEMBL entries to annotate them accordingly. Table 5 shows some examples of the aggregated rules presented in the form of prediction models. For each rule, the antecedent set is accompanied by its four evaluation measures and its Euclidean distance to normalized ideal metrics. The full list of prediction models obtained is available as part of the supplementary material.

**Annotation of UniProtKB/TrEMBL entries**

In order to capture the performance of our system, we considered the reference proteome set of prokaryotic entries of UniProtKB/TrEMBL for the purpose of annotation using our prediction models. Reference proteomes are a subset of proteomes that have been selected either manually or algorithmically according to a number of criteria to provide a broad coverage of the tree of life and a representative cross-section of the taxonomic diversity found within UniProtKB. It also covers the proteomes of well-studied model organisms and other species of interest for biomedical research. These reference proteomes are tagged with the keyword “Reference proteome”. As of March 2015, the reference proteome set of UniProtKB/TrEMBL entries of prokaryotes represents a fraction of around 6% over all prokaryotic UniProtKB/TrEMBL entries available in UniProtKB. In details, there are 4,631,729 prokaryotic reference proteome entries in UniProtKB/TrEMBL out of 74,189,325 total prokaryotic UniProtKB/TrEMBL entries. The coverage of our automatic annotations over the set specified is illustrated in the next section.

**Runtime analysis**

The system achieved a very high coverage of prokaryotic pathway prediction. That is, considering the parameters defined earlier in Table 4 out of 338 different pathways
presented in the learning dataset, 331 were validated and used to build pathway models for annotating UniProtKB/TrEMBL. The system took 77 minutes to generate the pathway models of all considered prokaryotic UniProtKB/Swiss-Prot entries with manual assertion evidences. The system was run on a 64-bit machine that has an Intel Core 3.00 GHz processor and a 16GB RAM.

RESULTS AND DISCUSSION

System evaluation

In order to evaluate the robustness of our system, we use the cross-validation technique with multiple runs. Cross-validation is a standard technique to give an insight on how the prediction models will generalize to an independent dataset. A single round of cross-validation involves partitioning data into complementary subsets and performing the analysis on one subset (called the training set), and validating the generated predictor on the other subset (called the validation set or testing set). For this experiment, we used the set of UniProtKB/Swiss-Prot prokaryotic entries with pathway annotations of manual assertion evidence for pathway annotations. Each rule is accompanied by its four evaluation measures and its Euclidean distance to normalized ideal metrics.

Table 6. Examples of prediction models obtained in the form or aggregated rules along with their evaluation measures for UniProt/Swiss-Prot prokaryotic entries with manual assertion evidence for pathway annotations.

| Prediction Model Form |
|-----------------------|
| [PREDICT] PATHWAY: Metabolic intermediate biosynthesis; chorismate biosynthesis; chorismate from D-erythrose 4-phosphate and phosphoenolpyruvate: step 2/7 [IF] [IPR:IPR016037] 0.00519535–1.0–0.0020161307216926784–1.0–1.4091160686761484 [END] |
| [PREDICT] PATHWAY: Amino-acid degradation; L-arginine degradation via AST pathway; L-glutamate and succinate from L-arginine: step 1/5 [IF] [IPR:IPR007041, IPR:IPR017650, IPR:IPR016181, TAXON:Enterobacteriales, TAXON:Enterobacteriaceae, TAXON:Gamma-proteobacteria, TAXON:Proteobacteria, TAXON:Bacteria] 5.23725E–4–1.0–0.02–1.0–1.3997688467343012 [END] |
| [PREDICT] PATHWAY: Amino-acid biosynthesis; S-adenosyl-L-methionine biosynthesis; S-adenosyl-L-methionine from L-methionine: step 1/1 [IF] [IPR:IPR022636, IPR:IPR022628, IPR:IPR022629, IPR:IPR022631, IPR:IPR022630, IPR:IPR002133, TAXON:Bacteria] 0.00627422–1.0–0.00155521106636306–1.0–1.408681270691111 OR [IPR:IPR027790, IPR:IPR002795, TAXON:Archaea] 4.60878E–4–1.0–0.00155521106636306–1.0–1.4127881840416994 [END] |

Table 7 presents the global evaluation metrics calculated over all target pathways. The table shows that our system achieved a very high accuracy of pathway identification with an F1-measure of 0.987, a precision of 0.991, a recall of 0.982 and an AUC of 0.99.

Distribution of annotation coverage

Here, we provide a comparison of our system annotation coverage over UniProtKB/TrEMBL with reference to all other automatic annotation systems run on UniProtKB/TrEMBL such as Rulebase (9) and HAMAP (10). Figure 1

Table 7. Evaluation metrics of cross-validation experiment over UniProtKB/Swiss-Prot prokaryotic entries with pathway annotations of manual assertion evidence

| Metric       | Value |
|--------------|-------|
| Accuracy     | 0.999 |
| Precision    | 0.986 |
| Recall       | 0.991 |
| F1-measure   | 0.982 |
| AUC          | 0.990 |
Figure 1. Annotation coverage for UniProtKB/TrEMBL reference proteome prokaryotic entries where (a) represents entries we could cover, (b) represents entries we could cover which lack pathway annotation, (c) represents entries we could cover which already have pathway annotation, and (d) represents entries we could not cover which already have pathway annotation.

Figure 2. Comparison of annotation coverage of UniProtKB/TrEMBL reference proteome prokaryotic entries with three main automatic annotation systems present in UniProtKB/TrEMBL which are SAAS, HAMAP-Rule, and RuleBase.

Moreover, 189,742 predictions where found to be identical matches to the annotations proposed by other systems. We also found 11,558 of our annotations similar to those proposed by other systems either being more specific or more general in their pathway hierarchical representation. Finally, there were 11,848 predictions distinct from those already assigned by the other systems.

Comparison of total number of prediction

In Figure 3, we take a deeper look into the various predictions made by our system in comparison to those made by RuleBase, SAAS, and HAMAP-Rule. Note that an entry in UniProtKB/TrEMBL could gain multiple predictions and hence obtain multiple pathway annotations accordingly. Here we were able to make a total of 636,052 predictions by our system where the majority of these predictions, 422,904, touched entries that have no previous pathway annotation.

Biological evaluation on Escherichia coli

The quantitative evaluation of the system using cross validation that we presented earlier in this section uses the...
Figure 3. Comparison of predictions applied on UniProtKB/TrEMBL reference proteome prokaryotic entries relative to three main automatic annotation systems present in UniProtKB/TrEMBL which are HAMAP-Rule, SAAS and RuleBase

Figure 4. Comparison of predictions corresponding to UniProtKB/TrEMBL reference proteome prokaryotic entries touched by both our system and RuleBase

Figure 5. Comparison of predictions corresponding to UniProtKB/TrEMBL reference proteome prokaryotic entries touched by both our system and HAMAP-Rule

Figure 6. Comparison of predictions corresponding to UniProtKB/TrEMBL reference proteome prokaryotic entries touched by both our system and SAAS

premise that proteins are already annotated with a pathway. Thus, the performance results of our system evaluation only hold for those proteins that are known to participate in some pathways. We try to circumvent this limitation by presenting a system biological evaluation of Escherichia coli.

Recently, the use of similarity measures (36) for comparison between various biological ontologies or, by extension, between entities annotated with these concepts had increased rapidly. We aim to study the relevance of GO ontologies (37) of the entries we annotated by our prediction models to those entries known to possess the same target pathway annotation. We considered, as a case study, the set of protein entries of Escherichia coli in UniProtKB/Swiss-Prot (taxid: 83333) since the coverage of GO annotation on UniProtKB/TrEMBL is low. We applied our prediction models on those entries that lack pathway annotations (3,714 entries in total). The prediction models provided 365 predictions touching 326 entries with 62 pathways that vary in their hierarchical representation. The set of those entries along with their pathway annotation are mapped to their corresponding entries of UniProtKB/Swiss-Prot with manual assertion evidences (171 entries) that share the same pathway annotation. This mapping is constructed in a form of pairs such that if protein P1 is known to participate in pathway P and protein P2 is predicted to participate in the same pathway P, then we form a pair in our mapping list as (P1 P2), and so on. We hypothesize that the computed semantic similarity of GO annotations of these pairs will be significant compared to the semantic similarity scores computed rest of pairs of proteins.

We computed semantic similarity for all GO ontology annotations available for the set of UniProtKB/Swiss-Prot entries of Escherichia coli (taxid: 83333) (3,884 entries in total). We used Semantic Measures Library SML (38) with Resnik measure (39), which is based on the information content of the most informative common ancestor. The
GO scores of the resulting pairwise semantic similarity computation with best matching average are recorded. The GO scores that corresponds to the computed mapping pairs will form the set of our positives for the Wilcoxon rank-sum test and the rest of the pairs are the negatives. We found that the p-values is less than 2.2e-16 which rejects the null hypothesis which states that there is not enough evidence to support the significance of the GO scores of our positive pairs.

CONCLUSION

In this paper, we introduced an automatic annotation method that utilizes rule mining techniques to predict metabolic pathways across wide range of prokaryotic genomes. Our system was successfully applied to gain knowledge on pathway identification from UniProtKB/SwissProt prokaryotic entries with manual assertion evidences. This knowledge was presented in the form of human readable prediction models to annotate UniProtKB/TrEMBL prokaryotic reference proteomes entries. Our prediction models annotated 551,418 UniProtKB/TrEMBL entries, where 371,265 of them lacked any previous pathway information. Furthermore, cross-validation testing demonstrated a very high accuracy of pathway identification with an F1-measure of 0.987 and an AUC of 0.99. Future development of this system includes studying the obtained pathway models to unveil pathways presence patterns across prokaryotic taxa and possible extension of the system to the annotation of eukaryotic proteins. Furthermore, we will investigate the pathways predictions in UniProtKB/TrEMBL that differ from those annotated by other systems to determine their validity.

FUNDING

This work was partially supported by competitive research funding from King Abdullah University of Science and Technology (KAUST) as a research internship for the first author in the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), UniProt team.

ACKNOWLEDGEMENTS

We would like to thank UniProt Consortium for their valuable support and feedback on the development of this work.

Conflict of interest statement. None declared.

REFERENCES

1. N. Campbell and J. Reece. *Biology*. No. v. 1 in Addison-Wesley student series, Benjamin Cummings, 2002.
2. X. Chen, J. Xu, B. Huang, J. Li, X. Wu, L. Ma, X. Jia, X. Bian, F. Tan, L. Liu, S. Chen, and X. Li. “A sub-pathway-based approach for identifying drug response principal network,” *Bioinformatics*, vol. 27, no. 5, pp. 649–654, 2011.
3. Y. Chen, Y. Hu, T. Zhou, K. K. Zhou, R. Mott, M. Wu, M. Boulton, T. J. Lyons, G. Gao, and J. xing Ma. “Activation of the wnt pathway plays a pathogenic role in diabetic retinopathy in humans and animal models,” *The American Journal of Pathology*, vol. 175, no. 6, pp. 2676–2685, 2009.
4. Y. Silberberg, A. Gottlieb, M. Kupiec, E. Ruppin, and R. Sharan, “Large-scale elucidation of drug response pathways in humans.” *Journal of computational biology*: a journal of computational molecular cell biology, vol. 19, pp. 163–174, Feb. 2012.
5. M. Parkes, A. Cortes, D. A. van Heel, and M. A. Brown, “Genetic insights into common pathways and complex relationships among immune-mediated diseases,” *Nat Rev Genet*, vol. 14, pp. 661–673, Sep 2013. Analysis.
6. E. Kretschmann, W. Fleischmann, and R. Apweiler, “Automatic rule generation for protein annotation with the c4.5 data mining algorithm applied on swiss-prot,” *Bioinformatics*, vol. 17, no. 10, pp. 920–926, 2001.
7. J. R. Quinlan, C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
8. The UniProt Consortium, “Uniprot: a hub for protein information,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015.
9. C. H. Wu, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z.-Z. Hu, R. S. Ledley, K. C. Lewis, H.-W. Mewes, B. C. Orcutt, B. E. Suzek, A. Tsigita, C. R. Vinayaka, L.-S. L. Yeh, J. Zhang, and W. C. Barker, “The protein information resource: an integrated public resource of functional annotation of proteins,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 35–37, 2002.
10. I. Pedrazzini, C. Rivoire, A. H. Auchincloss, E. Coudert, G. Keller, E. de Castro, D. Baratin, B. A. Cuche, L. Bougueleret, S. Poux, N. Redaschi, I. Xenarios, A. Bridge, and the UniProt Consortium, “Hamap in 2013, new developments in the protein family classification and annotation system,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D584–D589, 2013.
11. S. Muller, U. Leser, W. Fleischmann, and R. Apweiler, “Edittoreml: a distributed approach to high-quality automated protein sequence annotation,” *Bioinformatics*, vol. 15, no. 3, pp. 219–227, 1999.
12. C. H. Wu, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z.-Z. Hu, R. S. Ledley, K. C. Lewis, H.-W. Mewes, B. C. Orcutt, B. E. Suzek, A. Tsigita, C. R. Vinayaka, L.-S. L. Yeh, J. Zhang, and W. C. Barker, “The protein information resource: an integrated public resource of functional annotation of proteins,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 35–37, 2002.
13. C. Creighton and S. Hanash, “Mining gene expression databases for association rules,” *Bioinformatics*, vol. 19, no. 1, pp. 79–86, 2003.
14. O. Bodenreider, M. Aubry, and A. Burgun, “Non-lexical approaches to identifying associative relations in the gene ontology,” in *Pacific Symposium on Biocomputing* (R. B. Altman, T. A. Jung, T. E. Klein, A. K. Dunker, and L. Hunter, eds.), pp. 104–115, World Scientific, 2005.
15. I. I. Artamonova, G. Frishman, M. S. Gelfand, and D. Frishman, “Mining sequence annotation databanks for association patterns,” *Bioinformatics*, vol. 21, no. Suppl 3, pp. iii49–iii57, 2005.
16. G. Bebek and J. Yang, “Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks,” *Bioinformatics*, vol. 21, no. Suppl 2, pp. iii9–iii20, 2005.
17. G. Klopmann, M. Tu, and J. Talafous, “Meta. 3, a genetic algorithm for metabolic transform priorities optimization,” *Journal of Chemical Information and Computer Sciences*, vol. 37, no. 2, pp. 329–335, 1997. PMID: 9157102.
18. J. Jaworska, S. Dimitrov, N. Nikolova, and O. Mekenyan, “Probabilistic assessment of biodegradability based on metabolic pathways: Catabol system,” *SAR and QSAR in Environmental Research*, vol. 13, no. 2, pp. 307–323, 2002. PMID: 12071658.
19. B. Hou, L. Ellis, and L. Wackett, “Encoding microbial metabolic logic: predicting biodegradation,” *Journal of Industrial Microbiology and Biotechnology*, vol. 31, no. 6, pp. 261–272, 2004.
20. W. G. Button, P. N. Judson, A. Long, and J. D. Vessey, “Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics,” *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 5, pp. 1371–1377, 2003. PMID: 14502469.
21. S.-H. Chiu, C.-C. Chen, G.-F. Yuan, and T.-H. Lin, “Association algorithm to mine the rules that govern enzyme definition and to classify protein sequences,” *BMC Bioinformatics*, vol. 7, no. 1, p. 304, 2006.

22. P. Karp, M. Latendresse, and R. Caspi, “The pathway tools pathway prediction algorithm,” *Standards in Genomic Sciences*, vol. 5, no. 3, 2011.

23. J. Dale, L. Popescu, and P. Karp, “Machine learning methods for metabolic pathway prediction,” *BMC Bioinformatics*, vol. 11, no. 1, p. 15, 2010.

24. The InterPro Consortium, N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. Copley, E. Courcelle, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, S. Griffith-Jones, D. Haft, H. Hermjakob, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, S. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, F. Servant, and C. J. A. Sigrist, “Interpro: An integrated documentation resource for protein families, domains and functional sites,” *Briefings in Bioinformatics*, vol. 3, no. 3, pp. 225–235, 2002.

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

26. S. Bouker, R. Saidi, S. B. Yahia, and E. M. Nguifo, “Ranking and selecting association rules based on dominance relationship,” in *IEEE 24th International Conference on Tools with Artificial Intelligence, IJCAI 2012, Athens, Greece, November 7-9, 2012*, pp. 658–665, 2012.

27. S. Bouker, R. Saidi, S. B. Yahia, and E. M. Nguifo, “Mining undominated association rules through interestingness measures,” *International Journal on Artificial Intelligence Tools*, vol. 23, no. 4, 2014.

28. M. C. Chibucos, C. J. Mungall, R. Balakrishnan, K. R. Christie, R. P. Huntley, O. White, J. A. Blake, S. E. Lewis, and M. Giglio, “Standardized description of scientific evidence using the evidence ontology (eco),” *Database*, vol. 2014, 2014.

29. C. Borgelt and R. Kruse, “Induction of association rules: Apriori implementation,” in *Proc. of the 15th Conference on Computational Statistics (COMPSTAT)*, pp. 395–400, Physica Verlag, 2002.

30. R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB 94, (San Francisco, CA, USA), pp. 487–499, Morgan Kaufmann Publishers Inc., 1994.

31. C. Borgelt, “Efficient implementations of apriori and eclat,” in *Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL)*, CEUR Workshop Proceedings 90, p. 90, 2003.

32. C. Borgelt, “Recursion pruning for the apriori algorithm,” in *FIMI* (R. J. B. Jr., B. Goethals, and M. J. Zaki, eds.), vol. 126 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2004.

33. S. Brin, R. Motwani, and C. Silverstein, “Beyond market baskets: Generalizing association rules to correlations,” in *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, SIGMOD 97*, (New York, NY, USA), pp. 265–276, ACM, 1997.

34. A. Kirsch, M. Mitzenmacher, A. Pietracaprina, G. Pucci, E. Upfal, and F. Vandin, “An efficient rigorous approach for identifying statistically significant frequent itemsets,” in *Proceedings of the Twenty-eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 09*, (New York, NY, USA), pp. 117–126, ACM, 2009.

35. N. I. o. H. National Human Genome Research Institute, “Biological pathways,” [http://www.genome.gov/27530687](http://www.genome.gov/27530687).

36. C. Pesquita, D. Faria, A. O. Falco, P. Lord, and F. M. Couto, “Semantic similarity in biomedical ontologies,” *PLoS Comput Biol*, vol. 5, p. e1000443, 07 2009.

37. T. G. O. Consortium, “Gene ontology consortium: going forward,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

38. S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, “The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies,” *Bioinformatics*, vol. 30, no. 5, pp. 740–742, 2014.

39. P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI’95*, (San Francisco, CA, USA), pp. 448–453, Morgan Kaufmann Publishers Inc., 1995.