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

Imene Boudelioua1, Rabie Saidi2, Robert Hoehndorf1, Maria Martin2, Victor Solovyev4
1 Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia 23955
2 UniProt, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, UK

Current Status for prokaryotes & Problem
- UniProtKB/Swiss-Prot: 0.35M entries.
  - Entries with manual pathway annotations: 30.44%.
- UniProtKB/TrEMBL: 34 M entries.
  - Entries with automatic pathway annotations: 5.22%.
- Main systems on UniProtKB/TrEMBL:
  - HAMAP, Rule, RuleBase, SAAS
- Most of the sequences in UniProtKB do not have experimental pathway annotation. To tackle this problem, computational approaches are needed.

Method
- Our method will use a given set of UniProtKB/Swiss-Prot entries associated with prokaryotic taxonomies and InterPro IDs, and then generate pathway prediction models using rule mining techniques.

Hypothesis
For a given target C, If there exists a rule: 
A & B as C
with accepted quality of metric values, then we hypothesize that such a rule may reflect a biological regularity or peculiarity for C.

Applications
- Understand pathway presence patterns.
- Compare organisms based on their taxa.
- Enhance the quality of existing annotations for UniProtKB/TrEMBL entries.
- Annotate UniProtKB/TrEMBL entries with unknown function.

Dataset & Attributes
- UniProtKB/Swiss-Prot prokaryotic entries with manual assertion evidence and pathway annotations. 96,280 entries.
- UniProtKB/TrEMBL prokaryotic entries tagged as reference proteome. 6,193,540 entries.

System Pipeline
- Load Data
- Filter Data
- Extract Attributes
- Generate Rules
- Select Rules
- Choose Models
- Annotate TrEMBL
- Predict CC Pathway using Models
- Apply

System Evaluation
AUC 0.987
F-measure 0.982
Precision 0.999
Recall 0.966

Table 1: Two-run five-fold cross-validation for prokaryotic UniProtKB/Swiss-Prot entries with manual assertions.

System Settings
- Minimum support = 20.
- Minimum confidence = 100%.

Runtime Analysis
- Machine Specs:
  - 64-bit with Intel Core 3.00 GHz processor.
  - 16GB RAM.
- Total runtime for the system: 77 minutes and 19 seconds.

Prediction Results
- Number of pathways covered: 356
- Number of entries annotated: 663,724

Figure 1: Coverage for UniProtKB/TrEMBL "reference proteome" prokaryotic entries where (A) represents entries we could cover which lack pathway annotation, (B) represents entries we could cover which already have pathway annotation, and (C) represents entries we couldn’t cover which already have pathway annotation.

Figure 2: Comparison of predictions applied on UniProtKB/TrEMBL relative to other existing systems.

Figure 3: Comparison of UniProtKB/TrEMBL entries touched relative to other existing systems.

Conclusion & Future steps
- Our system produces a large number of pathway predictions with promising statistical evaluation results.
- Consult with biologists to check the correctness of NEW predictions proposed by our system.
- Further investigate conserved vs unique pathways among prokaryotic phyla.
- Infer consequent annotations.

References
1. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proc. 20th Int. Conf. on Very Large Databases (VLDB 1994: Santiago de Chile). 487-499 Morgan Kaufmann, San Mateos, CA, USA 1994.
2. Christian Borgelt and Rudolf Kruse. Induction of Association Rules: Apriori Implementation. 15th Conference on Computational Statistics (Compstat 2002, Berlin, Germany). 395-400 Physica Verlag, Heidelberg, Germany 2002.
3. Slim Bouker, Rabie Saidi, Sadok Ben Yahia, and Engelbert Mephu Ngupu. Ranking and selecting association rules based on dominance relationship. In IEEE 24th International Conference on Tools with Artificial Intelligence. ICTAI 2012, Athens, Greece, November 7-9, 2012, pages 659-663, 2012.
4. Slim Bouker, Rabie Saidi, Sadok Ben Yahia, and Engelbert Mephu Ngupu. Mining undominated association rules through interference measures. International Journal on Artificial Intelligence Tools, 23(4), 2014.

Learning Data
- UniProtKB/Swiss-Prot prokaryotic entries with manual assertion evidence and pathway annotations. 96,280 entries.

Applying Data
- UniProtKB/TrEMBL prokaryotic entries tagged as reference proteome. 6,193,540 entries.

Target and Attributes
- Target: CC metabolic pathway (Structured hierarchy of controlled vocabulary)
- Attributes:
  - Organism Taxonomic Lineage
  - InterPro IDs

Availability
http://www.ebi.ac.uk/~rsaidi/arba/