CS 229 Project: A Machine Learning Framework for Biochemical Reaction Matching

Tomer Altman[1,2], Eric Burkhart[1], Irene M. Kaplow[1], and Ryan Thompson[1]
[1]Stanford University, [2]SRI International

Biochemical reaction databases capture the sum of human knowledge of biochemical reactions and chemical compounds. As the amount of data available on metabolic reactions and chemical substrates increases, the necessity of central repositories increases as well. Unfortunately, there is no established algorithm for being able to calculate the degree of similarity between reactions in different databases. A set of features have been defined and were calculated for all pairs of reactions between the Kegg and MetaCyc reaction databases. Features include reaction name match, Tanimoto coefficient of the reactions in stoichiometric vector form, enzyme identifier matching, and Enzyme Commission classification differences. Logistic regression, non-linear SVM, naïve Bayes, and decision tree learning methods were implemented, and feature selection, cross-validation, k-means clustering and other debugging methods were applied to determine how to improve the algorithms and data. We conclude that decision trees and logistic regression provide the most accurate methods, and the Tanimoto coefficient is a key feature for the performance of both learning methods.

1. Introduction

Biochemical reaction databases capture the sum of human knowledge of biochemical reactions and chemical compounds as found across a multitude of organisms, and not one particular organism. As the amount of data available on metabolic reactions and chemical compounds increases, the necessity of central repositories increases as well. Several encyclopedic repositories of metabolic network data have been established over the years, including Rhea[1], Kegg[2], and MetaCyc[3].

The problem is that, unlike with chemical compounds and polymer sequences, there is no established algorithm for computing the degree of similarity between reactions in different databases. Without a means of comparing the contents of different reaction datasets, there is no easy way to combine or interlink these new bioinformatic data sources. We have defined a set of features for comparing two reactions from different databases and have evaluated them in the context of logistic regression, SVM, naïve Bayes, and decision tree classifiers.

In prior work[4], a set of features were defined and calculated for all pairs of reactions between the Kegg and MetaCyc reaction databases. Features include reaction name/synonym matching, cosine similarity of the reactions in stoichiometric vector form, enzyme identifier matching, and Enzyme Commission number match.

2. Features

2.1. Feature Details

Here we briefly describe the set of features:

2.1.1. *Reaction Name Matching*

Reactions either have a primary name and synonyms, or names and synonyms can be derived by the associated enzymes that are known to catalyze the reaction. Both an exact case-insensitive string match and the natural-language processing capabilities of Pathway Tools[5] for understanding enzymatic nomenclature have been utilized to find matching reaction names.

2.1.2. *Stoichiometry Vector Features*

A chemical reaction can be represented as a vector by having columns represent compounds, and by having rows represent reactions, using the stoichiometric coefficient for the numerical values. If a compound is not present in a reaction, it has a value of zero. One measure of similarity of two stoichiometry vectors is taken as the absolute value of the cosine similarity: $cos(\theta) = \frac{\vec{R}_{Kegg} \cdot \vec{R}_{Meta}}{||\vec{R}_{Kegg}|| \cdot ||\vec{R}_{Meta}||}$.

Biochemical reaction networks exhibit a "small-world" network property[6]. In other words, a small fraction of the chemicals participate in a large number of reactions. This means that trivial participants in reactions, such as NAD, ATP, water, and protons are weighted equally with more rare or signif-

icant reactants. In analogy to information retrieval from documents, a inverse-frequency scaling coefficient was applied to the values of the stoichiometry vectors when calculating the cosine similarity.

In addition to the cosine similarity, we implemented features for the Tanimoto coefficient (no inverse-frequency scaling applied), the number of compounds in common between the two reactions, the number of compounds in the MetaCyc reaction that had links to Kegg compounds, but not ones in the given Kegg reaction, the number of compounds in the MetaCyc reaction that had no known link to a Kegg compound, the number of compounds in the Kegg reaction that had links to MetaCyc compounds, but not ones in the given MetaCyc reaction, and the number of compounds in the Kegg reaction that had no known link to a MetaCyc compound.

2.1.3. *Enzyme Commission Hierarchy*

The Enzyme Commission of the International Union of Biochemistry and Molecular Biology[7] curates an ontology of enzymatic activities. Classes or instances in this ontology are assigned unique "EC numbers". This forms the back-bone of many reaction databases. Unfortunately, there are often duplicates of any given EC number, misannotations to the close, but incorrect, EC number, or partial EC numbers. We have used the ontology to compare reactions on the degree of similarity of their EC numbers, when present. The similarity was represented both as a categorical feature (with values from zero to four, representing how many levels of the hierarchy are shared), and as individual binary features for each category. Only one of the two forms (binary versus categorical) was included for any given learning algorithm.

Reactions that represent a reaction exactly as the Enzyme Commission depicts them are termed "official EC reactions". Variants of the official EC reaction will have the same EC number, but will have slight differences in the reaction equation. EC number matches for all four levels between "official' EC reactions" were provided as a separate feature.

2.1.4. *UniProt Identifier Match*

UniProt[8] is the preeminent protein database. Reactions either link directly to the UniProt entry representing the enzyme that catalyzes the reaction, or the reaction is linked to a protein object that in turn links to UniProt. Two reactions that can be mapped to the same identifier in UniProt are then catalyzed by the same enzyme, and therefore are essentially the same reaction.

2.1.5. *UniRef50 protein cluster match*

UniProt provides the UniRef50[9] dataset, which includes proteins from UniProt clustered at 50% sequence identity or higher. Proteins with this degree of sequence similarity are typically functionally related. Two reactions with enzymes within the same UniRef50 cluster are thus likely to share the same enzymatic function.

2.1.6. *Biochemical pathway match*

A network of biochemical reactions form a biochemical pathway. Kegg pathways tend to be ten times as large as MetaCyc pathways, and thus a one-to-one mapping between the two databases' pathways is not possible. Related biochemical pathways between the two datasets were determined, and pairs of reactions were evaluated to determine if they are present in related pathways.

2.1.7. *Data completeness features*

One of the characteristics of bioinformatic databases is that the attributes of entries are inconsistently populated with data. For example, even though a reaction may have a 'pathway' attribute, some entries might have zero, one, or more pathways populating that attribute. When comparing two reactions to determine if they are a match, it is possible that both, one, the other, or neither of the two reactions have pathway information specified. If we report which reaction pairs are within the same pathway, without representing the situation where there was no match possible due to one or both of the reactions missing pathway links, then we are introducing a form of bias into our training dataset.

We used four categories to represent missing

data: both, 0; Kegg only, 1; MetaCyc only, 2; neither, 3. We have created such status features for sequence data (covering exact UniProt identifier matches and UniRef50 features), stoichiometry vector features, and 'same pathway' features. We did not create a status feature for the name match feature, as the full set of attributes that the natural language processing program utilizes is unknown.

2.2. Development of a Gold Standard Training Set

In continuing work at SRI International, Ph.D. biologist curators have reviewed matches made by a conservative *ad hoc* rule over the feature set, and have identified true and false predictions. These manual assessments formed the gold standard training dataset.

Kegg version 50 and MetaCyc version 14.5 were used in this project. MetaCyc version 14.5 contains 7818 small molecule reactions, and Kegg version 50 contains 7827 small-molecule reactions. Data from MetaCyc was extracted using the Pathway Tools software, and data from Kegg was extracted using the BioWarehouse Kegg Loader[10].

This leads to a potential dataset of over 61 million pairs of reactions. Extracting features and analyzing a training dataset of millions of examples proved to be too inefficient for generating results for this project. Additionally, the majority of the full dataset contains millions of examples where there are no non-zero features. Thus, we included a sampling of the full space of reaction pairs, based on data content and curated match information. The final training dataset contained 9162 examples and twenty-one features.

3. Utilized Learning Algorithms

3.1. Naïve Bayes with Laplace Smoothing

We implemented naïve Bayes with Laplace smoothing to classify MetaCyc and Kegg reaction matches. Since the subset of EC number features are not mutually independent, we used only the "official" EC number level four feature. Similarly, the subset of stoichiometry vector-related features are also not mutually independent, so we used only the number of compounds in the Kegg reaction that did not have a match in the MetaCyc reaction but did have a matching compound in the MetaCyc database. Since 95.2% reaction pairs have a value of four or or less for this feature, we binned the integer number of compounds by setting any value greater than four to four.

3.2. Logistic Regression

We implemented logistic regression using the Newton-Raphson method and applied it to the training data.

3.3. SVM with Radial Basis Function

We have used libSVM[11] to train a non-linear SVM using a radial basis function kernel with an optimized γ value of 0.5: $e^{-\gamma \cdot |u-v|^2}$.

3.4. Decision Trees with Gini coefficient

We used a decision tree program[12] to create a decision tree based on our features.

3.5. *k*-means Clustering

A debugging technique based on unsupervised learning was utilized to locate sets of training examples that had poor separation between positive and negative labels. The training dataset was stripped of its labels, and several rounds of k-means clustering of the examples were performed for values of k ranging from 2 to 100. Manual inspection of the resulting clusters revealed five predominant clusters with more than 1000 examples. Clusters with approximately 50% positively-labeled members were examined to search for potential missing features that could be added to better separate the clusters.

4. Results

We ran our learning algorithms with 30/70 hold-out cross-validation to assess performance. The results for the utilized learning algorithms are summarized in Table 1.

We ran naïve Bayes with 30% hold-out cross-validation for ten randomly partitioned sets. In order to determine the most significant features, we ran naïve Bayes several times with different features removed.

Table 1. Learning method performance

Method	Mean Training Set Error	Mean Test Set Error	Most Important Feature	2^{nd} Most Important Feature
Naïve Bayes	9.86%	9.67%	# Mismatched Compounds in Kegg	EC Match 4 Official
Decision Tree	2.46%	5.88%	Tanimoto Coefficient	EC Match 4 Official
Logistic Regression	4.88%	4.84%	Tanimoto Coefficient	EC Match level 2
SVM	4.81%	5.13%	Tanimoto Coefficient	EC Match 4 Official

The most useful feature for each of the ten decision trees tested was the Tanimoto coefficient (the top node of the decision tree always split on a threshold of the Tanimoto coefficient). Cosine similarity and "official" EC number match level four features were secondary in importance, always splitting nodes at the second level of the decision tree.

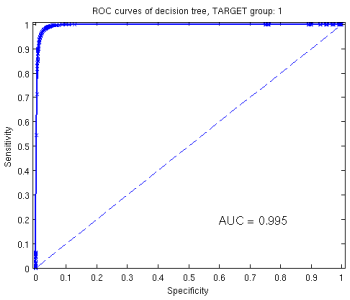

Fig. 1. Receiver-Operator Characteristic curve for decision tree classifier.

Points on an ROC curve were plotted for each leaf corresponding to a true match based on a decision tree trained on the entire dataset (Figure 1). Many of the points being located in the upper-left corner, along with an AUC value of 0.995 suggests that our classifier performs well.

The logistic regression cross-validation training error mean was 4.88% with a standard deviation of 0.24% and the test error mean was 4.84% with a standard deviation of 0.49%. The optimal θ found the had the following components with the largest magnitudes: 5.59, Tanimoto Score; -2.47, EC Match 2; 1.40, EC Match 4; 1.29, Name Match.

All of the classifiers had indications of high bias in our learning problem. The logistic regression classifier was used to construct a learning curve (Figure 2), which illustrates the bias.

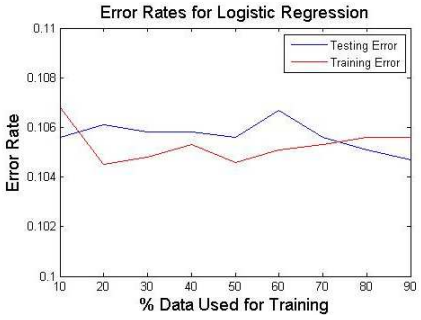

Fig. 2. Learning curve for Logistic Regression.

The decision tree additionally shows evidence of a variance problem, with the test error being twice the size of the train error.

5. Conclusion

5.1. Performance of Learning Algorithms

Naïve Bayes did not perform as well as the other classifiers. This is possibly due to a violation of the mutual independence assumption. Another possible reason for the inferior performance could be due to a suboptimal categorization of the integer and real-valued features, such as the Tanimoto coefficient and cosine similarity. Alternative categorization approaches could be explored to attempt to improve performance.

The decision tree classifier has the best performance (Table 1) in terms of train error. In contrast to other classifiers, its performance on the training set was consistently better than its performance on the test set. Its performance may be explained in part because the majority of our features are binary or categorical.

The logistic regression classifier has the best performance in terms of test error. It also lacks the variance problem observed with the decision tree, indicating that it might be a more reliable classifier.

The superior significance of the Tanimoto coefficient might indicate that the inverse frequency scaling applied to the cosine similarity was detrimental, and may be analyzed in future work. The k-means clustering analysis also indicated problems with cosine similarity not differentiating properly between positive and negative examples.

The machine learning framework approach to the problem of matching biochemical reactions demonstrates an improvement over the *ad hoc* rule previously implemented. It can be summarized as: `Official EC Match 4 OR (cosine similarity > 0.75 AND UniProt Link) OR (cosine similarity > 0.75 AND Name Match)`.

Applied to the training set, the *ad hoc* rule has an error rate of 11.7%. Every learning method attempted succeeded in surpassed the performance of the *ad hoc* rule, with the best approaches reducing the error rate by more than half.

5.2. High Bias Problem

The small difference between training and test error for logistic regression in Figure 2 indicates the chosen hypothesis has not over-fit the data. The magnitude of the errors indicate that the available features are not sufficient and that we can benefit from finding additional useful features.

A fundamental question is whether we have a systematic bias in our training dataset due to the nature of evaluating pairs of entities. For two databases with N reactions each, there are potentially order $O(N^2)$ reaction pairs, with only $O(N)$ pairs being true matches. This means that training on the full set of pairs, or a random sampling of the pairs might have one or more orders of magnitude in difference between the number of positively and negatively labeled examples.

In our sampling of the $O(N^2)$ reaction pair space, we specifically filtered out reaction pairs that had no non-zero feature values. This introduces another layer of bias.

As we discovered useful features, we found ways to break up the feature into a number of more refined features. This has incrementally improved the performance of our learning algorithms on our dataset. Future work will involve evaluation of classification results to identify additional hidden features and refinement of well-performing features.

We would like to acknowledge the work of Dr. Ron Caspi and Ms. Anamika Kothari, who reviewed predicted matches and curated our gold-standard training dataset with true matches and mismatches, and helpful discussions on the work with Dr. Douglas Brutlag, Dr. Peter D. Karp, Joseph M. Dale, and Dr. Luciana Ferrer.

References

1. Rhea database. `http://www.ebi.ac.uk/rhea/`
2. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M.; "KEGG for representation and analysis of molecular networks involving diseases and drugs". *Nucleic Acids Res.* 38, D355-D360 (2010).
3. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." *Nucleic Acids Res.* 2010. 38(Database issue):D473-9.
4. Altman T.."BioChem 218 Final Paper: Large-Scale Alignment of Encyclopedic Metabolic Networks." 2009. `http://tinyurl.com/29cells`
5. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R. "Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology." *Brief Bioinform.* 2010 Jan;11(1):40-79.
6. Barabasi AL, Oltvai ZN. "Network biology: understanding the cell's functional organization." *Nature Reviews Genetics* 5, 101-113. 2004.
7. The Enzyme Commission - IUBMB. "Enzyme Nomenclature", Supplement 5. *Eur. J. Biochem.* 1999, 264, 610-650.
8. The UniProt Consortium. "The Universal Protein Resource (UniProt) in 2010." *Nucleic Acids Res.* 38:D142-D148 (2010).
9. Suzek B.E., Huang H., McGarvey P., Mazumder R., Wu C.H. "UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters." *Bioinformatics* 23:1282-1288(2007).
10. Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, Karp PD. "BioWarehouse: a bioinformatics database warehouse toolkit." *BMC Bioinformatics.* 2006 Mar 23;7:170.
11. Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : a library for support vector machines", 2001. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
12. Padoan, Andrea. "Decision Trees and Predictive Models with cross-validation and ROC analysis plot." `http://tinyurl.com/22padwn`. 2010.