The Genomics of a Signaling Pathway: A KDD Cup Challenge Task

Mark Craven
Department of Biostatistics and Medical Informatics
University of Wisconsin
1300 University Avenue
Madison, Wisconsin
craven@biostat.wisc.edu

ABSTRACT

This article describes Task 2 of the KDD Cup 2002 data-mining competition. The task involved predicting whether deletions of individual genes in a yeast genome would affect a particular signaling pathway in the cell. With its rich data sources and abundance of missing information, the task is representative of data mining problems in genomics and proved to be quite challenging.

Keywords

data mining competition, genomics, gene regulation

1. INTRODUCTION

Molecular biology increasingly has become a data-driven science. The complete genomes of more than a hundred species have been determined and several recent technologies enable certain aspects of cellular "activity" to be measured for thousands of genes or proteins at a time [1; 2]. This new era of high-throughput, data-driven biology offers many challenges and opportunities for data mining researchers and practitioners. This article provides an overview of Task 2 of KDD Cup 2002, which investigated the use of data mining methods to help characterize the results of a recent high-throughput experiment in a molecular biology lab.

2. THE PROBLEM DOMAIN

The key source of data for Task 2 of KDD Cup came from a set of experiments performed by Guang Yao and Chris Bradfield of the McArdle Laboratory for Cancer Research at the University of Wisconsin. Yao and Bradfield were investigating the Aryl Hydrocarbon Receptor (AHR) signaling pathway, which plays a key role in how cells respond to certain toxic chemicals, among other things. AHR is a protein that can act as a *transcription factor*. When a cell is exposed to certain toxic chemicals, such as dioxin, the AHR system acts to turn on the expression of certain genes. The complete inventory of proteins (the products of most genes) that are involved in the signaling pathway is not known. The goal of Yao and Bradfield's experiment was to identify the set of genes that play a role in the pathway. Figure 1 shows a simple model of how the AHR pathway works; a detailed description can be found elsewhere [3].

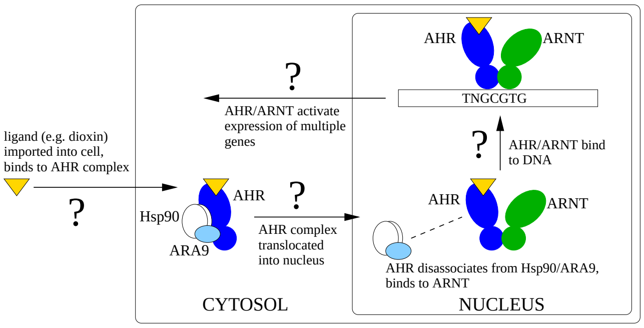

Figure 1: A simple model of the AHR signaling system. The question marks in the figure indicate steps at which there may be additional proteins involved in the pathway.

Although the gene for the AH Receptor itself is not native to yeast, Yao and Bradfield constructed a yeast model system for AHR in order to take advantage of the Yeast Deletion Library [2] which consists of a set of strains of the yeast *Saccharomyces cerevisiae*, in which each strain is characterized by a single gene being knocked out (i.e. rendered inactive). Yao and Bradfield transformed more than 4500 strains from the Yeast Deletion Library by inserting into each the AHR gene along with a reporter system that enabled them to measure how active AHR signaling was in any given strain. The result of this experiment was the identification of 134 genes that, when knocked out, cause a significant change in the level of activity of the AHR system.

For this set of 134 gene deletions, Yao and Bradfield also tested to see if they had a similar effect on a control pathway. These two experiments thus grouped the genes into three categories. The no-change category contains genes whose deletion did not significantly affect the activity of the AHR system, the control category contains genes whose deletion significantly changed the the activity of *both* the AHR system and the control system, and the change category contains genes whose deletion significantly affected the AHR system, but not the control system.

3. THE KDD CUP TASK

The challenge faced by Yao and Bradfield was to understand the biology underlying the genes that appear to play a role in the AHR system. Although it would have been interesting to have a KDD Cup task focused on *annotating* the results of this experiment, such a task would not be amenable to objective evaluation. Therefore, as a surrogate for this annotation task, the KDD Cup task involved learning models for classification.

Each instance in the data set represents a single gene, and its target value is a label indicating the combined outcome of the AHR-signaling experiment and the control experiment. The goal of the task is to learn a model that can accurately predict these class labels. For training, the groups competing were given class labels for 3,018 yeast genes and data characterizing all of the genes under consideration as well as additional yeast genes that were not assayed by Yao and Bradfield. The data describing the genes consists of the following: (i) hierarchical features characterizing the known functions of proteins encoded by various genes and the subcellular locations of these proteins. (ii) a table listing pairs of proteins that physically interact with one another. (iii) a set of 15,235 abstracts from scientific articles, along with an index listing which genes are referenced in which abstracts. The goal of the data mining task here is to identify the regularities and relationships that characterize the genes involved in regulating the AHR system. This could be done with a narrow focus: characterizing the genes that affect the AHR system but not the control system. Or this could be done with a broad focus: characterizing the genes that affect the AHR system without regard to what happens in the control system. Both of these perspectives are valuable, and both were considered in KDD Cup.

The teams competing provided predictions based on two different binary partitionings of the class labels. In the first case, the positive class consisted of those genes with the change label only and the negative class consisted of those genes with either the no-change or the control labels. This partitioning corresponds to the narrow characterization of genes affecting the AHR system described above. In the second case, the positive class consisted of those genes labeled with either the change or the control label, and the negative class consists of those genes labeled with the no-change label. This partitioning corresponds to the broad characterization of the genes affecting the AHR system.

4. EVALUATION

The test set consisted of 1489 instances. For the two partitions described above, teams provided their predictions sorted by the confidence that each gene belongs to the positive class. By varying a threshold on these sorted predictions, we constructed Receiver Operating Characteristic (ROC) curves for each team for each partition. An ROC curve is a plot of the true positive rate (fraction of the positive instances for which the system predicts positive) against the false positive rate (fraction of the negative instances for which the system erroneously predicts positive) for the different possible thresholds. Teams were evaluated by considering the area under the ROC curves (AROC) for the two partitions. The expected area for random predictions is 0.5 and the area for perfect predictions is 1. The winner was the team with the greatest summed AROC for the two curves.

Figure 2 shows a scatterplot of the AROC values for the 54 teams that competed in Task 2 of KDD Cup. Nearly all of the teams had AROC values greater than 0.5 for both partitions. This result indicates that there is predictability in the given task and representation. However, all of the AROC values are less than 0.7, indicating that the task is quite hard. The winning team was Adam Kowalczyk and Bhavani Raskutti from Telstra Research Laboratories. Four teams were accorded honorable mention for either having summed AROC values that were very close to the winning team, or for having the best AROC value on one of the partitions. Articles describing the systems of the winning and honorable-mention teams appear elsewhere in this issue.

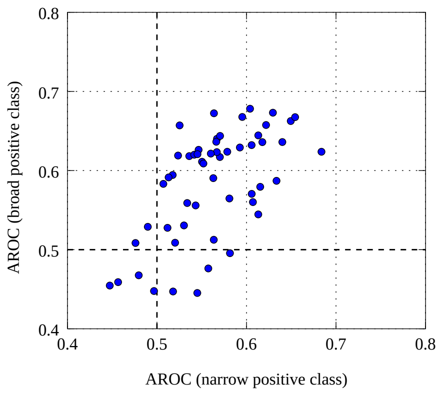

Figure 2: AROC scores for all teams for both partitions of the data set. The x-axis represents AROC for the narrow definition of the positive class, and the y-axis represents AROC for the broad definition of the positive class. Each point corresponds to the two scores for a particular team.

5. CONCLUSIONS

There are many aspects of this task that make it challenging for data mining methods: (i) the data sources are rich, including scientific text and hierarchical and relational features, (ii) there are few positive instances, (iii) there is a large amount of missing information, (iv) the target concept is most likely disjunctive (different genes affect the AHR pathway for different reasons). However, all of these aspects are representative of data mining problems in genomics. Therefore, tasks such as the one considered here should prove to be fertile testbeds for data mining research. The data for KDD Cup Task 2 and additional information is available at `http://www.biostat.wisc.edu/~craven/kddcup/`.

6. ACKNOWLEDGEMENTS

Guang Yao and Chris Bradfield of the McArdle Laboratory at the Univ. of Wisconsin kindly made available the data from their AHR experiments. The author is supported by NSF grant IIS-0093016 and NIH grant 1R01-LM07050-01.

7. REFERENCES

[1] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics Supplement*, 21:33–37, 1999.

[2] E. Winzeler and et al. Functional characterization of the Saccharomyces cerevisiae genome by gene deletion and parallel analysis. *Science*, 285:901–906, 1999.

[3] G. Yao. *Understanding the Ah Receptor Regulatory Network*. PhD thesis, McArdle Laboratory for Cancer Research, University of Wisconsin, Madison, WI, 2002.