Preprint
UCRL-JC-154431

A rule-based framework for gene regulation pathways discovery

B.Wilczynski, T.Hvidsten, A.Kryshtafovych, L.Stubbs, J.Komorowski, K. Fidelis

This article was submitted to
the IEEE Computer Society Bioinformatics Conference (CSB2003)
Stanford, CA, USA, August 11-14, 2003

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

July 21, 2003

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

A rule-based framework for gene regulation pathways discovery

Bartosz Wilczyński
Lawrence Livermore National Laboratory
Livermore, CA, USA
Warsaw University,
Warsaw, Poland

Torgeir Hvidsten
Lawrence Livermore National Laboratory
Livermore, CA, USA
Linnaeus Centre for Bioinformatics
Uppsala, Sweden

Andriy Kryshtafovych
Lawrence Livermore National Laboratory
Livermore, CA, USA

Lisa Stubbs
Lawrence Livermore National Laboratory
Livermore, CA, USA

Jan Komorowski
Linnaeus Centre for Bioinformatics
Uppsala, Sweden

Krzysztof Fidelis
Lawrence Livermore National Laboratory
Livermore, CA, USA

Abstract

We present novel approach to discover the rules that govern gene regulation mechanisms. The method is based on supervised machine learning and is designed to reveal relationships between transcription factors and gene promoters. As the representation of the gene regulatory circuit we have chosen a special form of IF-THEN rules associating certain features (a generalized idea of a Transcription Factor Binding Site) in gene promoters with specific gene expression profiles.

1 Introduction

Understanding of the gene regulation mechanisms is currently one of the most important tasks for molecular biology. New genome sequences become available every month, but we are still far from being able to reliably and consistently unravel the corresponding gene regulatory pathways. As the number of known genomes is growing and the average size of a genome sequence dataset is large, we expect the method of pathway discovery to be general, automated, and efficient. We describe our approach, based on the assumption that genome sequences, with known gene positions, together with the gene expression data, are sufficient to find all the interactions between genes and their transcription factors.

2 Algorithm outline

A given set of genes, with expression levels measured under specific conditions, is used as input to our procedure. We start with gene expression profile clustering and then iteratively perform the following steps:

- Searching for significant features in the gene regulatory regions
- Inducing rules from genes clustered by expression and a feature set
- Evaluating the rules
- Refining the feature set until no further progress can be achieved.

3 Local similarity of gene expression profiles

We will assume that the genes that are co-regulated by a common transcription factor show significant correlation in their expression profiles. Many authors (like [6, 5]) search for the transcription factor binding sites in the regulatory regions of genes with similar expression profiles. Some transcription factors, however, may bind only under certain conditions. Indeed, for example genes from baker's yeast demonstrate expression correlation only during a part of the yeast cell cycle. This leads us to a modified approach to gene expression clustering. We consider groups of genes that show very high correlation only in a part of the expression profile.

4 Features definition

We are using a term “feature” as a generalization of a transcription factor binding site. Transcription factors often interact with one another synergistically, suggesting that features more sophisticated than just the presence of a single transcription factor binding site are needed. We define a feature as a presence of one or more transcription factor binding sites and use constraints on their position and the number of occurrences in the regulatory region of the gene. We are interested in features present in regulatory regions of a sufficiently large set of genes with correlated expression.

5 Inducing rules

The core part of the algorithm is the induction of the rules from expression clustering and feature set. We are using a special form of IF-THEN rules as the representation of a simplest regulatory scheme. The example rule

$$f_1 \wedge f_2 \wedge \ldots \wedge f_n \rightarrow C_1 \vee C2 \vee \ldots \vee C_m,$$

where f_i denotes a feature and C_j denotes a expression class (a profile on a subset of datapoints), expresses our observation that if we will find all the features $f_1, \ldots f_n$ in the promoter of any gene, we should expect it’s expression profile to be similar to the classes $C_1, \ldots C_m$. To generate rules of this kind we create a binary matrix with rows representing genes associated with the expression profile class and columns representing features, and then apply the Rosetta approach [1, 2]. Rosetta is a rough-set([3, 4]) based library for general data mining and knowledge discovery. It allows us to find the smallest subsets of features sufficient to discern between different expression classes. Then it creates the rules combining the sets of features found in the regulatory regions of the genes with the observed expression profile classes.

6 Features refinement

The Rosetta library also allows us to evaluate the importance of a feature. If the feature is used in the rule with big support, we can assume it is important. By the same reasoning, we would consider a feature not used in any of the rules as useless. In each iteration of our algorithm, after inducing the rules, we refine the feature set. In the process of refinement, we remove the features that are not used and create new features. We create new features as combinations of features occurring in the left-hand side of the rules. If some certain combination of features is appearing in more than one rule, we use it to create a new feature. In the next iteration the rule inference engine can take advantage of the new feature.

7 Conclusions

We show that the IF-THEN rule representation of regulation mechanisms is sensible and can be successfully used to mine the gene expression data. As an example, we will show the rules obtained for the yeast cell cycle data.

This work was performed in part under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory (contract W-7405-Eng-48).

References

[1] J. Komorowski, A. Skowron, and A. Ohrn. The rosetta system.

[2] A. Ohrn. The rosetta homepage. http://www.idi.ntnu.no/ aleks/rosetta.

[3] Z. Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 11:341–356, 1982.

[4] Z. Pawlak. Rough sets: Theoretical aspects of reasoning about data. volume 9. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[5] Y. Pilpel, P. Sudarsanam, and G. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29:153–159, 2001.

[6] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15:776–784, 1999.

[7] J. Zhu and M. Zhang. Scpd: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15:607–611, 1999.