Inference of Gene Relations from Microarray Data by Abduction

Irene Papatheodorou, Antonis Kakas*, and Marek Sergot

Department of Computing, Imperial College London, SW7 2AZ, UK

Abstract. We describe an application of Abductive Logic Programming (ALP) to the analysis of an important class of DNA microarray experiments. We develop an ALP theory that provides a simple and general model of how gene interactions can cause changes in observable expression levels of genes. Input to the procedure are the observed microarray results; output are hypotheses about possible gene interactions that explain the observed effects. We apply and evaluate our approach on microarray experiments on *M.tuberculosis* and *S.cerevisiae*.

1 Introduction

The focus in bioinformatics has shifted from the analysis of genome sequences to functional genomics, which seeks to ascribe biological function to genes and understand gene interactions. An important tool in these studies is DNA microarray technology, which enables simultaneous measurement of expression levels of thousands of genes. A common form of experiment aims at identifying genes whose expression is affected by environmental conditions or by changes in expression of other genes. This information will give clues about gene interactions and unraveling pathways that define the cell's responses to various stimuli. Datasets are too large and complex for manual analysis. Raw data are analysed using statistical techniques to define significantly differentially expressed genes. Methods for further interpretation of the results, in terms of gene interactions, remain largely undeveloped, however, though Bayesian Networks have recently attracted attention (e.g. [1]).

We formulate the analysis of this type of microarray data as a problem of *abduction*, that is, inference from observable effects, i.e. the microarray data, to possible causes, hypotheses about possible gene interactions. We construct an Abductive Logic Program (ALP) theory which provides a simple, general model of how gene interactions can cause changes in observable expression levels of genes—essentially a formalisation of the (usually implicit) reasoning used by biologists designing microarray experiments. Adjustable parameters allow us to constrain the search for hypotheses and apply the methods to large data sets. A novel feature of our method is the ability to deal with observations in many separate experiments together.

* Visiting from Department of Computer Science, University of Cyprus.

The model is validated by comparing the inferred hypotheses against known gene interactions and by assessing the biological plausibility of the hypotheses where detailed information is lacking. We use microarray data sets on *M.tuberculosis* and *S.cerevisiae* (yeast). Section 3 presents an example of inferences that re-discover part of the *M. tuberculosis* heat shock response pathway.

There are many issues and other experimental methods in the search for gene regulation mechanisms. To our knowledge, inference of gene networks from microarray data has not previously been formulated as a problem of abduction, though abduction has been used in [4] to construct a genetic network from classical genetics experiments. The nature of the data, the hypotheses and the model itself, differ from what is addressed here.

2 The Model

Input to the procedure is a set of observations expressed as logic assertions of the form *increases_expression(Expt, Gene)* and *reduces_expression(Expt, Gene)*. They are obtained by statistical analysis of the raw microarray data to determine the significance of measured differences of expression levels of each gene [3].

The output is a set of abducible relations of two different types: *induces(Gene1,Gene2)* and *inhibits(Gene1,Gene2)* for the hypothesis that *Gene1* induces the expression of *Gene2*, or inhibits it, respectively. Each individual experiment provides partial clues about possible *induces*/*inhibits* relations between genes.

The ALP Framework

The modelling framework we employ is Abductive Logic Programming (ALP) [5], an extension of logic programming that allows declarative logical representations of the problem domain and supports abductive reasoning. A theory is represented by a triple (P, A, IC), where P is a logic program, A a set of abducible predicates and IC a set of classical logic formulae, the *integrity constraints*.

An abductive explanation for a query Q is a set Δ of ground abducible atoms on the predicates A such that: $P \cup \Delta \models_{LP} Q$, $P \cup \Delta$ is consistent, $P \cup \Delta \models_{LP} IC$, where $\models_{LP}$ denotes a standard entailment relation of logic programming.

The integrity constraints IC impose additional *validity* requirements on the hypotheses Δ. They are modularly stated in the theory, in addition to the basic model in P. They augment any partial information on the abducible predicates or impose other constraints on the abductively generated explanations. We form integrity constraints (IC) of three different types:

(1) self-consistency: For example, a gene cannot both inhibit and induce the same gene at the same time (under the same conditions).

(2) consistency with background information: Background knowledge, such as known inhibitor genes can also be expressed in the form of IC.

(3) experimental consistency: When analysing the results of an experiment E in which a gene G is mutated, we may want to consider as 'intermediary genes' (explained in section 2) only genes whose expression is also observed to be affected in experiment E.

Gene interactions

Top-level Rules The program P of the ALP theory represents how gene interactions can increase or reduce the expression of genes, as observed in the experiments. An assumption is that such observed variations in gene expression should be attributed directly or indirectly to the variations (gene mutations or environmental stress), carried out in the experiment(s) investigated.

For example: if an experiment E knocks out a gene G, and G inhibits gene X, then E will show an increased expression of X — subject to some possible exceptions. This rule is expressed in logic programming notation as follows:

increases_expression(E, X) ← (1)
knocks_out(E, G), inhibits(G, X),
not incr_affected_by_other_gene(E,G,X),
not incr_affected_by_EnvFact(E, X).

E is a variable that ranges over names of experiments and G, X are variables that represent genes. *increases_expression(E,X)* is observational data from the experiment E, *inhibits(G,X)* is part of the unknown information to be abduced, and *knocks_out(E,G)* provides background knowledge about the experiment E.

The last two conditions express possible exceptions that deal with the possibility that the difference in gene expression can be attributed to a factor other than the mutated gene: e.g. (a) a gene other than G, or (b) an environmental factor. Here, *not* is the logic programming construct 'negation as failure', used to express that (1) is a default general rule subject to the stated exceptions.

Similarly, there is a rule that deals with the cases of reduced expression of G in experiment E. Similar rules cover the cases of over-expressing G and further rules deal with the various combinations of gene mutation and changes in environmental conditions according to our classification of experiment types.

Rule (1) only accounts for direct relationships between the mutated gene and the differentially expressed one. These relationships could be indirect: Inference of intermediate steps of interaction is achieved by further recursive rules:

increases_expression(E, X) ← (2)
mutates(E, G), intermediary_gene(E, Gx, G),
reduces_expression(E, Gx), inhibits(Gx, X),
not incr_affected_by_other_gene(E,Gx,X),
not incr_affected_by_EnvFact(E, X).

If gene Gx inhibits gene X, and the expression of gene Gx is reduced (directly or indirectly) by the mutation of gene G in experiment E, then the expression of X is increased in the experiment E. The relation *mutates(E,G)* covers both knock-out and over-expression of gene G in the experiment E.

The Parameters are relations that control the genes taken into account when searching for hypotheses. In the general case, where every gene is possibly related to other genes, there may be an exponential number of possible hypotheses. With the parameters we constrain the problem by reducing the search space. By varying their definition, we can test different possibilities of the model. There are two parametric relations including *intermediary_gene/3* in rule (2), as well

as the integrity constraints, which provide another means of constraining the search space.

3 Application: Heat Shock Response of *M. tuberculosis*

M. tuberculosis data sets were obtained from our collaborators at the Centre for Microbiology and Infection, Imperial College London [3] and publicly available tables from the Schoolnik lab, Stanford University. Observations and inferred hypotheses are presented as directed graphs using our visualisation tool [2].

Observations from 5 experiments were selected according to a conservatively chosen significance threshold in the 1st phase statistical analysis. Each experiment knocks out or over-expresses a gene believed to be involved in heat shock response and known to function as a transcriptional regulator (regulator of expression of other genes). The two parameters of the model were defined to restrict attention to possible interactions between 16 genes of known regulatory function. Analysis of the observations in all experiments together generated a single hypothesis, shown in graphical form in figure 1, that explained all observations.

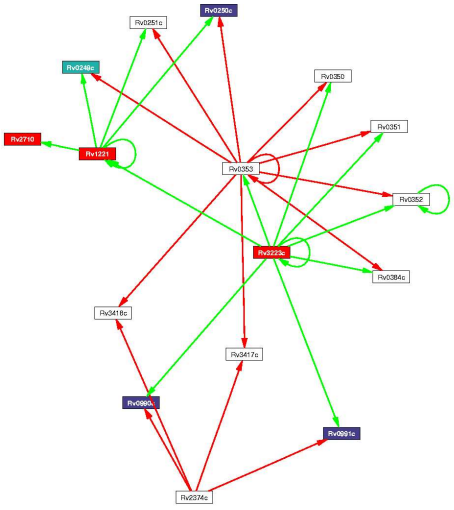

Fig. 1. The nodes represent genes, colour coded according to a standard functional classification. The edges show the inferred relations between genes, red for *inhibits* and green for *induces*. Cyclical edges represent the autoregulation relationships abduced.

The resulting hypothesis is in agreement with previous knowledge [3]. The DnaK operon[1] (genes Rv0350–353, on the right of the figure) is controlled by the positive regulator sigH (Rv3223c) and the negative regulator hspR (Rv0353). The acr2 operon (genes Rv0249c–251c, at the top left of the figure) is controlled by the positive regulator sigE (Rv1221) and negative regulator hspR (Rv0353). The groES/EL genes (Rv3417c–Rv3418c) are under dual negative control by

[1] group of genes that reside next to each other on the DNA and are expressed together

hspR (Rv0353) and hrcA (Rv2374c). Known feedback loops are also discovered: the DnaK operon (Rv0350–353) is negatively regulated via its member Rv0353.

Finally, there is a group of genes whose function in heat shock response is not clear but are linked in the explanatory hypothesis. Rv0249c and Rv0250c are both unknown genes, repressed (inhibited) by hspR, both next to Rv0251c in the chromosome. This could be a real effect, suggesting they are in an operon, or it could be some artefact due to their place on the chromosome and the way data is collected. Similarly, Rv0990c and Rv0991c could also be members of an operon, but isolated with no obvious function in heat shock. Our collaborators are planning to investigate these hypotheses in a new set of experiments. Further discussion of our methods and their applications is available in [6].

4 Conclusions

We develop a *general method* to support the analysis of an important class of microarray experiments. The novel feature is a simple, general model of how gene interactions can cause changes in observable expression levels of genes under differing conditions, and the use of abduction to infer explanatory hypotheses. This method allows us to infer regulation relations across several experiments.

The declarative and modular nature of this *gene interaction model* allows us to experiment easily with variations and new general rules suggested by our biological collaborators, and to add biological knowledge as it becomes available. The parameters in the model allow us to constrain the search space of possible hypotheses and thereby apply the methods to realistically large data sets.

Tests on *M. tuberculosis* rediscovered part of the heat shock response mechanism and suggested further experiments. We are presently engaged in a systematic exploration of the various possibilities afforded by the model and an extensive validation against known gene regulation processes in yeast.

We have been able to apply these methods in practice to the analysis of large data sets. Whatever the biological significance of this technique turns out to be in the long-term, the model provides a valuable test case for those concerned with the development of abductive reasoning technology.

References

1. Friedman N., Linial M., Nachman I., Pe'er (2000) Using Bayesian Networks to analyze expression data. *J. Comp. Bio.* 7:601–620.
2. Papatheodorou I., Sergot M., Randall M., Stewart G., Robertson B. (2004) Visualisation of Microarray results to assist interpretation *Tuberculosis* 84:275–281
3. Stewart, G. *et al* (2002) Dissection of the heat-shock response in *Mycobacterium Tuberculosis* using mutants and microarrays. *Microbiology* 10:3129–3138.
4. Zupan, B., Demsar, J., Bratko, I. *et al* (2003) GenePath: a system for automated construction of genetic networks from mutant data. *Bioinformatics* 19:383–389.
5. Kakas, A.C., Denecker, M. (2002) Abduction in Logic Programming. In *Computational Logic: Logic Programming and Beyond, Part I.* Springer-Verlag, pp402–436.
6. Technical Report No. 2005/3, Department of Computing, Imperial College London, 2005