Functional discrimination of gene expression patterns using ILP

Liviu Badea
AI Lab, National Institute for Research and Development in Informatics
8-10 Averescu Blvd., Bucharest, Romania, email: badea@ici.ro

Motivation and outlook. The ever-growing amount of experimental data in molecular biology and genetics requires more and more sophisticated knowledge discovery tools. We argue that the complexity of the domain makes ILP an ideal candidate, due to its ability to deal with expressive background theories. We illustrate our claim by using an ILP learner to induce functional discrimination rules between genes studied using microarrays and found to by differentially expressed in three recently discovered subtypes of adenocarcinoma (AC) of the lung [Garber et al. 2001]. The discrimination rules involve functional annotations in terms of the Gene Ontology (GO). While most of the lower levels of gene expression data (pre)processing have been automated, our work can be seen as a step toward automating the higher level functional analysis of the data. In fact, ILP could play an even more important role in gene expression analysis, as the existing functional annotations will become more sophisticated.

Setting up the learning problem. We have used the dataset from [Garber et al. 2001] restricted to the samples classified in AC. The selection of positive examples (genes correlated with the AC3-AC1,2 distinction) was performed using a nonparametric t-test with a (very strict) P value cutoff of 0.0005, while all genes with $P \geq P_{neg} = 0.3$ were considered negative examples. Finally, only the examples with a counterpart in the HumanPSD database (thus having a GO annotation) were kept, namely 39 positive and 87 negative examples.

The GO annotations of the examples selected above were used as background knowledge. As many of the annotations are quite specific, they may be unable to cover more than one example. To allow non-trivial generalisations, we added as background knowledge an encoding of the GO hierarchy in the form of Prolog rules.
Using the ILP learner Progol [Muggleton'95], we searched for hypotheses of the form:

```
target(Gene) :- function(Gene, 'calcium binding'), function(Gene, 'protein binding').        where
function(Gene,GOterm) :- ( 'Molecular Function'(Gene,X) ; 'Biological Process'(Gene,X) ;
                           'Cellular Component'(Gene,X) ),  go(GOterm, X).
```

For each positive example, Progol's heuristic search selects *only one* "best" clause discriminating this positive example from the negative examples, although there might be several clauses with the same covering and the same heuristic estimate. Returning just one, as Progol does, may not be enough in our application, where the number of negative examples may be too small to invalidate the wrong alternatives (candidate hypotheses), and/or alternative viewpoints on the distinction between classes are highly desirable. But since Progol conducts a *complete* admissible search, it was relatively straightforward to extract *all* alternative hypotheses (having the same covering and the same heuristic estimate as the hypothesis selected by Progol).

Results. We found functional discrimination rules for virtually all the genes discussed in Garber et al. [2001] in relation to AC. The most significant ones are:

- the *'cell cycle control'* genes, especially the cyclin-dependent kinase inhibitor CDKN2A and the polo-like serine/threonine kinase PLK (not mentioned in [Garber et al. 2001]), which are both up-regulated in AC3.
- the *'cystein-type endopeptidase'* cathepsin L (CTSL), involved in (extracellular) proteolysis, which is also up-regulated in AC3 w.r.t. AC1,2.
- the *'signal transduction'* and *'growth factor'* Dickkopf (DKK1), which is known to contribute to neoplastic processes (it may play a key role in the transition from an epithelial to a mesenchymal phenotype - which is significant since AC involve epithelial cells).
- the *'transporter', 'membrane'* solute carrier protein SLC7A5, closely linked to cellular activation and division.
- in the *'tumor antigen'* category, the carcinoembryonic antigen-related cell adhesion molecule CEACAM1 is involved in tumor angiogenesis. (CEACAM1 is down-regulated in AC3 w.r.t. AC1,2.)

The need for a relational learner. The GO annotations involve only hierarchies of concepts, where the only type of relational information is in the form of *inheritance* relationships. But a relational learner could use much more sophisticated relational annotations, such as: *'part-of'* relationships (already present in the Gene Ontology), and causal relationships between genes/proteins such as *activates(G1,G2).* For example, it could learn a hypothesis of the form

```
target(Gene) :- 'Biological Process'(Gene, 'cell cycle control'), 'Cellular Component'(Gene, X),
                active(G1), 'Cellular Component'(G1, X1), part_of(X, X1).
```

References

Garber et al. Diversity of gene expression in adenocarcinoma of the lung. Proc.Natl.Acad.Sci. 98, 13784-13789, Nov. 20, 2001.
Muggleton S. Inverse entailment and Progol. New Generation Computing Journal, 13:245-286, 1995.

Acknowledgments. This work was partially supported by the EC Network of Excellence ILPnet2.