Poster L-10

Query-driven Search for Analysis of Large Microarray Collections

Authors:
Matthew A. Hibbs (*Lewis-Sigler Institute for Integrative Genoimcs, Department of Computer Science, Princeton Universit*)
Kai Li (*Department of Computer Science, Princeton University*)
Olga G. Troyanskaya (*Lewis-Sigler Institute for Integrative Genoimcs, Department of Computer Science, Princeton Universit*)

Short Abstract: In order to explore vast compendiums of microarray data, new methodologies are required to provide fast, intuitive access to the collection. We developed a methodology for fast search and analysis of large collections of microarray data taking into account the functional context of each query.

Long Abstract:
The availability of a large and growing collection of microarray datasets from diverse experimental contexts provides hope of broadly characterizing gene expression and regulation in a variety of conditions. However, such vast amounts of data can be unwieldy to analyze and misleading conclusions can be drawn due to biases of functional coverage. Searches and analyses of these data must be targeted and interactive in order to allow expert biologists to leverage their own knowledge to quickly formulate and test new hypotheses and conclusions.

We have built a database of S. cerevisiae microarray datasets from over 80 publications, totaling roughly 2400 microarray conditions. Traditional analysis methods, including many forms of clustering, can lead to misleading results given such a large and diverse compendium of data. Most clustering methods can be driven by “global” signals common to many of the conditions in the collection, while missing more “local” signals that may be present in only a small subset of the conditions. For example, in our compendium more than 1000 of the 2400 conditions relate to stressing cells in some way (heat shock, exposure to chemicals, deprivation of nutrients, etc.) causing all of these conditions share a general stress response signal, while less than 30 of the 2400 conditions allow cells to grow into the sporulation phase where a meiosis-specific signal can be found. Given this type of functional disparity, traditional clustering methods miss the specific sporulation signal as it is washed out by the general stress response signal.

One way to address this problem is by clustering together not only groups of genes, but by identifying bi-clusters of genes that are coherent under a particular subset of conditions. The general problem of bi-clustering is very computationally difficult (NP-complete) which makes the application of straightforward algorithms infeasible on large data collections. Many simplifications, heuristics, and approximations have been applied to make the problem tractable, but most solutions still require days or more of computation to perform a complete bi-clustering of data of such size, which eliminates these algorithms from use in an interactive manner.

We propose a technique that leverages query-based analysis and pre-processing of the compendium to allow for fast, targeted searches through a large amount of data. In our approach a domain expert supplies a query of a small number of genes known to functionally interact with a reasonable expectation of having related expression in at least some conditions. Given this query, our methodology selects subsets of data most relevant to the biological process where query genes display the strongest relationship to each other. Then within this subset, we can identify additional genes with strong relationships to the query set. This targeted approach allows us to generate accurate results in real time. Furthermore, our methodology allows iterative improvements of the search, which facilitates more thorough exploration of the data compendium by combining expert knowledge with fast data analysis to investigate the function of interest.

Using this approach we can quickly recapitulate known pathways and functions given a small seed set of related genes, as well as predict novel players in many specific contexts. Further, examining under which subsets of data these predictions show strong relationships to the query set can suggest what laboratory conditions may be appropriate for further experimentation to confirm these functional predictions.