Scalable Computational Methods for Analysis of the Low Dose Radiation Transcriptome*

Michael A. Langston[1] and Brynn H. Voy[2]
[1] Department of Computer Science, University of Tennessee, Knoxville, TN 37996
[2] Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6445

Guilt by association, the assumption that genes with similar expression patterns participate in common cellular functions, drives a growing body of effort to extract cellular pathways from microarray data. The general tenet is that genes encoding proteins participating in a common pathway will display correlated expression levels when analyzed at sufficient scale, and that the identities and known functions of these genes can be used to highlight existing and assimilate new functional pathways. A number of recent studies validate the concept of guilt by association, demonstrating that genes co-expressed across multiple conditions are more likely to represent common functions than would be expected by chance alone. To date the computational methods to extract such patterns lag far behind the general agreement about their utility.

Clustering includes a wide variety of algorithms for organizing groups of genes with similar expression patterns into a format that is easy to interpret. Virtually all clustering algorithms begin with a similarity metric (e.g., Pearson's correlation coefficient) between the expression levels of all pairs of genes in the analysis. However, there are several important limitations to the vast majority of clustering algorithms that are in contrast to the realities of biology. Specifically, the clusters produced are disjoint, requiring that a gene be assigned to only one cluster. Another major problem is that most measures of similarity used in clustering algorithms do not permit the recognition of negative correlations, which are also common and equally meaningful. As an alternative to assigning genes to clusters, the correlation matrix of expression data can be converted into a graph consisting of nodes (genes) connected by edges (gene-gene correlations). Applying a threshold to the matrix restricts the graph to only those edges likely to represent biologically meaningful relationships. The resulting graph (relevance network) contains many dense subgraphs of tightly interconnected gene sets that intuitively represent the greatest potential for identifying members of common pathways. Without a systematic means to extract the aggregate relationships between multiple genes, however, many of the most interesting relationships remain embedded within the sea of correlations.

We have developed a computational approach that exploits graph theoretical algorithms to identify comprehensively the tightly connected subsets of genes present in relevance networks. In the most extreme case, in which a subgraph contains all possible edges between vertices in the subgraph, this structure is called a *clique*. In terms of gene expression, clique represents the most trusted potential for identifying a set of interacting genes. Solving clique, however, is an *NP*-complete problem, and a classic graph-theoretic problem in its own right. We have developed novel graph algorithms that employ fixed-parameter tractability and vertex cover to solve immense instances of clique efficiently. We have applied these algorithms to identify differential gene relationships, that is, gene-gene interactions that are induced or repressed by in vivo exposure to low dose ionizing radiation (IR). The data used were derived from a study that explored the role of genetic susceptibility in the response to IR (see related abstract by Voy et al). Six strains of inbred laboratory mice were exposed to 10 cGy of X-rays, after which gene expression changes in spleen were profiled using microarrays. We describe our graph theoretical-based toolchain for identifying overlapping subsets of genes with tightly correlated expression levels, and discuss the biological insight that this method provides.

---

* Work supported by the Low Dose Radiation Research Program of the United States Department of Energy (ERKP442), and by the Office of Biological and Environmental Research, US Department of Energy, under contract DE-AC05-00OR22725 with UT-Battelle LLC, the managing organization of ORNL for the U.S. DOE.