Commentary

Extracting functional information from microarrays: A challenge for functional genomics

Michael Q. Zhang*

*E-mail: mzhang@cshl.edu

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724

The advent of the human and model organism genome project has provided an increasingly complete list of genes that code for the building blocks of life on earth. Deciphering the functions of all these genes has proven to be no easy task. The availability of mountains of transcriptional profiling data from modern large-scale gene expression technologies, such as serial analysis of gene expression (SAGE) (1), oligonucleotide arrays (2) and cDNA microarrays (3), represents a tremendous windfall for computational biologists who have largely migrated from many different fields. One paper appearing in this issue of PNAS (4) introduces a novel computational approach, Shortest Path (SP) analysis, to assign gene functions in a transitive fashion along a correlation linkage path terminated by two known genes belonging to the same functional category.

Currently the most popular way to identify interesting genes and their functions is to perform cluster analysis on the relative expression pattern changes (Figure 1a) in typical microarray experiments that survey a range of conditions (e.g. reviewed in ref. 5). The fundamental premise of clustering approach is that genes having similar expression profile across a set of conditions (cellular process, responses, phenotypes, etc) may share similar functions (6). Obviously the word “function” is too general to be precise and quantitative, and is too broad to be specific and meaningful. Genes, whose products may have same function (say, phosphorylating other proteins), do not necessarily share similar transcriptional pattern. Conversely, genes having different functions can have similar expression profile simply by chance or stochastic fluctuations. Even though many potential caveats exist, large numbers of functionally related genes do show very similar expression pattern under relevant set of conditions, especially genes that are co-regulated by common transcription factors or their products are the components of a larger complex. And this is why a simple clustering of genes with similar expression pattern has allowed to assigning a putative function to unknown genes via “guilt by association” arguments (e.g. 7-8). Several clustering techniques, such as hierarchical clustering (9), Kmeans (10) and self-organizing map (SOM) (11), have been adopted from other fields and applied widely to microarray data analyses. Successful as it is, clustering cannot reveal functional relation among genes whose expression patterns show very little correlations (they may be related by a time-delay for instance), Figure 1b.

A major goal of microarray data analyses is to identify genes that interact with each other in a particular cellular process (or pathway) where not every player has a similar transcriptional profile. The crucial aspect of Zhou et al. approach is to extend the co-expression concept to a more general “transitive co-expression” which appears to be an important characteristic of many biological processes: two genes involved in the same process may not be strongly correlated in expression directly, but can be both strongly correlated with the same set of other genes. Another widely recognized point is that functional annotations should really be incorporated early in the data analysis, not surprisingly the starting point of the Zhou et al. work is exploiting the controlled vocabulary tree in the biological process categories of Gene Ontology (GO) (12), Figure 1c.

In essence, this SP analysis method starts from a pair of genes belonging to the same biological process category and to the same major cellular compartment (mitochondrial, cytoplasmic or nuclear) according to GO, and constructs the SP through a chain of pair-wise strongly correlated genes, with a distance function that further contracts the strongly correlated genes. Unknown transitive genes on the SP are assigned with the function of the lowest common ancestor of all the process subcategories corresponding to the known genes on the same SP (Figure 1b,c,d). To define a sufficiently specific gene function, the total SP length must be very short and this lowest ancestral node must be at least four levels below the root of the GO tree. In particular, if all the known genes are in the same node, the lowest common ancestor is the starting terminal gene process category itself (level L0 assignment): if they are in different nodes but all share a direct parent with the terminal genes, this parent node will be identified as the lowest common ancestor (level L1 assignment), Figure 1c.d.

To test the validity of their SP method, Zhou et al. had applied it to the analysis of the S. cerevisia gene expression profiles of the Rosetta Compendium (13) which measured the response of 300 gene-deletion and drug treatment
experiments. Firstly, they used only the known genes (~1,300 that have GO cellular process and localization annotations). The SP method was able to successfully called 64%/84% (cytoplasm), 59%/69% (mitochondria) and 39%/51% (nuclear) transitive genes at the LOF/L1 levels, and these results are highly significant as shown by further permutation tests. Encouraged by the bench-mark tests, they extended the graphs of SPs of known genes to additional ~3,300 unknown ORFs and were able to assign functions (i.e. cellular process categories) to 146 ORFs which include 75 high confident predictions (a gene function assignment is high confident if the gene is the only unknown gene on the SP). As a gene may belong to several SPs, it can therefore get multiple-function assignments. One may choose not to make a prediction on an unknown gene if known genes on the SP fail to have a consistent annotation as the terminal genes. As often faced by many computational biologists, Zhou et al. had spent tremendous amount of efforts in trying to substantiating the biological content of their findings by extensive literature searches. Among the 75 high confidence annotations, 24 were found in the Yeast Proteome Database (YPD: http://www.proteome.com) and 16 (83%) were confirmed by YPD documented experiments. More encouragingly, their computational results appear to be able to correct some database annotation errors after a closer scrutiny.

As stated by the authors, the strength of their method is to use the SP to link “transitive co-expressed” genes even if some of the genes (especially the terminal genes) on the SP do not have correlated expression profiles directly. Further advantage is exemplified by the “active incorporation of the biological annotation into the knowledge discovery process”. But the conceptual significance actually lies in a much deeper level. For example, one could also ask if two known targets of a transcription factor (TF) are taken as the terminal nodes, could more targets along the SP be identified analogously? If not for the SP defined by the particular distance function, maybe some other SP defined by a more appropriate distance function would have to be used? In general, one could argue that, to certain extent, the goal of all microarray data analyses is to identify a functionally linked sub-network hidden in the expression profiles. Suppose we view the expression profile space consisting clouds of points (genes), if we connect all genes (assuming we know every gene function) involved in a particular part of a cellular process (say, cell cycle progression), we would trace out a sub-network path. We could do the same for a different process and would get another path. The intersection would define gene(s) that are involved in both processes. If the two processes are so linked, we could actually trace out a connected sub-network, Figure 1.d. Conversely, discovering such hidden functional linkages (paths, sub-networks, etc.) activated by response or process variables (such as time-shift in the cell cycle process) would be the central task. The expression space does not have to be limited to relative mRNA density changes at different time or conditions, it could also include proteome information, localization variables, tissue and developmental parameters. It is actually non-trivial to find the right metric function that defines relevant distance relations appropriate to the cellular processes interested and allows investigators to construct the SP links capable of tracing out the functional sub-networks. Although the particular distance function and related SPs of Zhou et al. may not be sufficient for identifying all types of processes, the general methodology does represent a significant extension on our microarray data analysis repertoire beyond cluster analysis. It is not clear how far one can take this empirical SP approach. If the two terminal genes are both multi-functional, will there be more likely a single SP with multi-functional transitive genes or multiple SPs with largely single functional transitive genes on each SP? It is more likely that the incomplete knowledge of the existing GO tree and the current resolution for most microarray data will prevent us from getting the answers to such questions. But the real key for understanding transcriptional profiles and gene regulation networks is to link expression pattern to TF binding sites (cis-regulatory elements). Recent advances in computational (14, 15, reviewed in 16) and experimental technologies (17, 18) have opened up real opportunities for annotating gene functions, not only at the phenomenological levels but also at the mechanistic levels.
a. Expression profile matrix (table). \( t \) = \((t_1, t_2, \ldots)\) is the experimental condition index, in this example it indicates a set of time points.

\[
g_{i,j}
\]

b. Expression profiles (patterns). \( g_1 \) and \( g_4 \) are not strongly correlated directly, but both are strongly correlated with the correlated set \((g_x, g_2)\). \( g_x \) and \( g_2 \) are the transitive genes interpolating the two terminal genes along SP1 (see c. and d.); Similarly, \( g_y \) is the transitive gene interpolating \( g_1 \) and \( g_5 \) along SP2.

c. GO biological process tree, P's are process annotation for genes at a particular node. A gene may belong to more than one node ("multiple-function", such as \( g_2 \)).

d. Expression profile space. \( g_3 \) is on the short path SP1 terminated by the known genes \( g_1, g_2, \) and \( g_4 \) and hence is assigned function of \( P_{1,1,1,1}\) (level L0) according to the GO tree in c.; \( g_3 \) is on SP2 terminated by \( g_1, g_5 \) and is assigned function of \( P_{1,1,1}\) (level L1). \( g_3 \) is shared by both SPs and may be involved in both processes. This means the processes represented by SP1 and SP2 actually cross-talk to each other, the linked gene network can be formed by the subgraph \( SP_1 + SP_2 \).

**Figure 1.** Relations among different concepts in the Shortest Path (SP) analysis method.