Managing the Data Deluge

J. Sunil Rao, Meredith Bond

Over the last 10 to 20 years, the search for mechanisms responsible for cardiac remodeling during cardiac hypertrophy and failure has been hampered by the experimental tools available (primarily Western blot analysis and polymerase chain reaction). This is because these approaches only permit measurement of the expression levels of a few preselected genes at one time. However, there is increasing evidence that at the molecular level the changes that occur during development of heart failure represent a complex series of interrelated events. Thus, to identify the full scope and complexity of the subcellular changes that take place and thus make more rapid progress in identifying complete, the good news is that the number of genes or gene fragments whose expression can be assessed in a single pass by high-throughput analysis will steadily increase, and analysis and display programs that can handle and display the enormous amount of data should become more readily available. As the choices of microarrays increase and (hopefully) become cheaper, more and more investigators will be accessing this technology. However, the study by Liu et al in this issue of Circulation Research reminds us that in our rush to embrace these new technologies, we need to take pause and consider several important issues pertinent to data analysis and interpretation (Figure). Stage 1 of this gene exploration is relatively straightforward. It depends on the size of an investigator’s supply budget for purchase of cDNA arrays or GeneChip® and, secondly, on his or her technical skill to consistently produce high-quality labeled RNA or cDNA. Stage 2, data analysis, is more problematic. There are several issues to consider. First is simply the size of the data sets; gigabytes of computer storage and very fast computers are now routinely required for storage and manipulation of gene expression data. In theory, this problem can be overcome by use of faster computers, large disk storage arrays, fast network interconnects, and modern data backup and archiving systems (albeit at great expense). The second issue is a thornier one, and this is addressed by Liu et al. How can we determine which changes in gene expression are statistically significant? How do we set the sensitivity and specificity of the analysis, and, in view of the very large number of genes analyzed, how do we avoid false positives?

Gene screening involves statistical hypothesis testing and as such has built in type I and type II errors. There are two issues at play here, one of which is addressed by the study by Liu et al. The other is indirectly addressed. The first issue deals with how replicates increase the accuracy of database estimates and hence statistical hypothesis testing. To investigate this question, Liu et al have chosen as their test system the changes in gene expression in isolated cardiac myocytes stimulated by insulin-derived growth factor-1 (IGF-1). IGF-1 is one of several factors known to trigger changes leading to cardiac hypertrophy, resulting in increased cell size, assembly of sarcomeres, and reexpression of fetal genes. Liu et al cap a rigorous statistical analysis of their cDNA expression data with a report of identification of several novel genes. Recently, this same issue was formally addressed for microarray data. However, the study by Liu et al uses a more heuristic
approach to demonstrate how increasing the number of replicates reduces false detection rates (FDRs) of gene expression changes during cardiac remodeling.

The second issue deals with the multiplicity of statistical tests conducted. In this case, the usual error rates ($P<0.05$) applied to each test are no longer valid. Instead, family-wise error rates (cumulative error rates over the total number of hypotheses tested) need to be considered and procedures need to be developed to ensure that the overall error rate over all tests conducted is below some threshold. However, when thousands of tests are conducted, as in the case of gene screening, this becomes impractical. Therefore, the notion of FDRs$^{16,17}$ has been developed to answer the following question: out of all of the hypothesis tests rejected (ie, significant results and consideration of FDRs), what proportion are rejected incorrectly? FDRs can be estimated from data using permutation or bootstrap methodologies (simulation techniques used when traditional assumptions, such as normality, do not hold) and have been successfully applied to gene screening for microarrays.$^{18}$ A minor point is that the authors equate low FDRs with high specificity, whereas low FDRs actually indicate high sensitivity. To achieve high specificity, one would have to have some knowledge of false negatives. Whereas some theoretical work along these lines has been done, nothing has yet been extended to the microarray problem.

Acknowledgments

The authors wish to acknowledge Carley Gwin, Director, Gene Expression Core, and Eldon Walker, Director, Research Computing Services, at the Lerner Research Institute, The Cleveland Clinic Foundation.

References


Key Words: microarrays | statistical analysis | cDNA | cardiac remodeling
Microarrays: Managing the Data Deluge
J. Sunil Rao and Meredith Bond

Circ Res. 2001;88:1226-1227
doi: 10.1161/hh1201.093165

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circres.ahajournals.org/content/88/12/1226

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation Research can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation Research is online at:
http://circres.ahajournals.org/subscriptions/