Papers on normalization, variable selection, classification or clustering of microarray data

Over the last decade or so, there have been large numbers of methods published on approaches for normalization, variable (gene) selection, classification and clustering of microarray data. As indicated in the scope document for *Bioinformatics*, this requires papers describing new methods for these problems to meet a very high standard, showing important improvement in results for real biological data, as well as novelty. In this editorial, we describe some standards that need to be met for papers in these areas to be seriously considered. We ask that prospective authors consider these points carefully before submission of their papers to *Bioinformatics*.

The role of simulation: Simulation can be useful in investigating the properties of various methods of data analysis. Yet, there are important barriers to credible use of simulation in microarray studies, largely due to what we do not know about the statistical distribution of measured gene expression levels. First, the distribution across transcripts of true expression values is dependent on the biological state of the tissue or cell, and for a given state this is unknown, even in distributional form, and may further exhibit gene- and platform-specific effects. Second, the correlation within biological replicates of true expression is unknown, and is likely unknowable in detail given that it is expressed by a correlation matrix with on the order of a billion entries. Third, the distribution of changes from one biological state to another is unknown. Fourth, the correlation in observational errors in gene expression across genes is unknown and similarly probably unknowable in detail. On the other hand, the measurement error for a given transcript has been well described by several authors (Ideker *et al.*, 2001; Rocke and Durbin, 2001). Given this gap between knowledge and simulation specification, it is likely that any new method can be shown to be superior to some other method(s) by careful choice of simulation parameters, since simulations often include biases in the distributions selected and in other assumptions of the models. Thus, while simulation may still be worthwhile, and a useful tool for exploring robustness and parameter space of a new method, it is insufficient evidence for superiority of a new method without substantial support from significant improvement in results from analysis of real data.

Normalization: Normalization necessarily involves a trade-off between its positive role in reducing variability, and its potentially negative role in increasing bias. There are a number of good image analysis, preprocessing, transformation and normalization methods extant for single- and dual-color DNA microarrays. To show that a new method is better requires comparison demonstrating that results in differential expression analysis, classification or clustering are better with the new normalization method than with previous methods. Not one but several previous methods should be chosen for comparison including the most widely used approaches. Several datasets should be used, including spike-in and dilution studies when feasible, as well as 'real' biological datasets. Showing that more genes are differentially expressed using a normalization method is not compelling evidence of superiority without a good estimate of the false-positive rate or a compelling biological analysis of the resulting differentially expressed genes.

Variable selection: Typically, new variable selection methods are proposed as part of a classification or clustering strategy, and demonstrating superiority of the variable selection method usually means demonstrating superiority of the combined methodology. It is quite important that metrics for evaluation be used that are robust to intra-array correlations and variable selection artifacts. For example, in cross-validation studies in which variable selection is followed by a classification method, selection of variables using all the data and then cross-validating the classification accuracy introduces substantial bias, making classification methods appear more accurate than they really are (Ambroise and McLachlan, 2002). It is important that any method be compared with several of the most widely used existing methods, including baseline approaches such as filtering by *t*-score or forward stepwise analysis. Such comparisons should be performed on more than one biological dataset. Further, the method must demonstrate significant improvement over existing methods; incremental improvements will not be considered of sufficient interest to warrant review.

Classification and prediction: New classification or prediction methods for microarray data enter a crowded arena. From long-standing techniques such as logistic regression and linear discriminant analysis to the more modern support vector machines and neural networks, most known classification methods have already been applied to microarray data. To show that a newly proposed classification method is a real advance, a substantial improvement in performance needs to be shown over a reasonable selection of existing datasets and methods, including commonly used or simple methods. This is because, consciously or subconsciously, the developer of a new method optimizes its characteristics against the datasets to be used for evaluation. Variable selection and parameter choice for all methods needs to be done strictly in the training set (whether there is one training set or many as in cross validation). Resampling methods like permuting the class labels on the arrays or the bootstrap can be used to provide robust estimates of the significance of differential expression, but do not in themselves give estimates of classification performance except to show that the performance is better than chance. Experience shows that there is considerable noise in classification accuracy experiments, so modest increases in achieved accuracy are usually not convincing. Experience also shows that classification performance in a microarray problem depends strongly on the dataset, and less on the variable selection and classification methods. More than modest differences are required to excite interest in a new method. Authors should keep in mind the 'No Free Lunch Theorems' of Wolpert and Macready (1997) which demonstrated that there is no optimization/classification method that outperforms all others in all circumstances (Wolpert, 1996).

Clustering: Demonstrating superiority of a clustering method is in many ways more difficult than demonstrating superiority in a

classification method. Usually, there is no ground truth against which to compare the clustering results. Defining a criterion (e.g. the Rand index) and showing that a clustering method achieves better scores on this criterion is often not compelling, since such criteria are easily optimized (again, consciously or subconsciously) to ensure superiority. For reasons discussed above, simulation is also not usually sufficient. Ideally, a new clustering method would demonstrate novel biological insights or some attractive statistical properties not available from previous methods, including several commonly used methods. Requiring new biological findings is a difficult standard, but a necessary one to insure that new published methods are useful and likely to be used.

To conclude, microarrays remain a useful technology to address a wide array of biological problems and the optimal analysis of these data to extract meaningful results still pose many bioinformatics challenges. However, with a number of successful methods already addressing the well-established microarray data analysis problems, publication of new methods in this area requires either identification of a new challenge and formulation of a new problem or development of a substantially better methodology then those existing that can be benchmarked on a variety of datasets. We hope that suggestions provided above for evaluation and validation of such new methods would increase the likelihood of them supporting biological discoveries in the future.

David M. Rocke, Trey Ideker,
Olga Troyanskaya, John Quackenbush and Joaquin Dopazo

REFERENCES

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Ideker,T. *et al*. (2001) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Rocke,D.M. and Durbin,B.P. (2001) A model for measurement errors for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

Wolpert,D. (1996) The lack of a priori distinctions between learning algorithms. *Neural Comput.*, **8**, 1341–1390.

Wolpert,D.H. and Macready,W.G. (1997) No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, **1**, 67–82.