Bringing Out the Best Features of Expression Data

Frederick P. Roth

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115 USA

Classification and Feature Selection

Scientists are constantly classifying objects based on observation: A *Drosophila* geneticist sexes flies; a taxonomist sorts butterflies according to genus and species; a physician interviews patients, observing symptoms and rapidly classifying patients according to their disease, and predicting how they will respond to various therapies. The most successful patient interview is accompanied by prior knowledge of what questions to ask, and which observable variables ("features") are most salient to the classification problem at hand. Although the process of learning salient features in biology and medicine has traditionally been based on a combination of experience, intuition, and anecdotal evidence, it has increasingly been approached from a statistical perspective. However, choosing patient features salient to diagnosis from among tens of thousands of potential features, after drawing on experience from only tens of patients, is a whole new ball game. This is the challenge facing those who seek to derive diagnostic markers from gene expression array data.

Cancer Diagnostics from Array Data

Diagnostic and prognostic marker development is perhaps the most profound promise of mRNA and protein expression data. Markers discovered via expression arrays should, in the not too distant future, assist disease detection, disease classification, and possibly choice of therapy for specific patients. The feasibility of discriminating types of cancer with array data was shown previously by Golub and coworkers (Golub et al. 1999), who were able to distinguish acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL). "Feature selection" (marker gene selection) was accomplished using a score (similar to a Student's *t* statistic) measuring, for each gene, the difference and separation in expression levels between ALL and AML. To perform a prediction on a tissue sample of unknown type, the selected features (genes) each cast a weighted "vote" for one of the possible tissue types. The weights are determined by similarity of the measured expression in the unknown sample to measured expression in the previously observed ALL or AML training samples. A classification method such as this, that is trained on an example data set, is called supervised learning. The feasibility of prognosis using expression data was shown by Alizadeh and colleagues (Alizadeh et al. 2000). They predicted response by patients with B-cell lymphoma to chemotherapy, using a combination of tissue sample clustering on the basis of all features (an example of unsupervised learning) and intuition about the biology of B-cell lymphomas to do feature (gene) selection. Another round of tissue sample clustering using only the selected features was shown to separate patients by chemotherapy response.

E-MAIL froth@hms.harvard.edu; FAX (617) 432-3557.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.215501.

Feature Wrappers

In this issue, Xiong and colleagues (Xiong et al. 2001) attack the problem of feature selection and marker discovery from array data using a "feature wrapper" approach. The guiding principle of this approach is that the features *which can best be used* for classification of the tissue sample should be chosen. A consequence of this principle is that one must know exactly how tissue samples will be classified before feature selection can be done. The process of feature selection wraps around the classifier in the following procedure (sometimes called a "jackknife"):

(1) A candidate set of features is considered.
  a. Tissue samples are divided into training and test sets.
    i. The classifier is trained on the training set of tissue samples.
    ii. The classifier is used on the test set of tissue samples.
  b. Step 1(a) is repeated with alternative divisions into training and test sets.
  c. The candidate feature set is evaluated using all classifications from 1(a)(ii).
(2) Step 1 is repeated with another candidate feature set.

In this way, many candidate feature sets are evaluated using the training set of tissue samples, and the feature set that performs best is chosen. Xiong et al. have utilized this feature wrapper approach and compared the performance on three popular classification methods: Fisher's linear discriminant analysis (LDA), logistic regression (LR), and support vector machines (SVM).

A major advantage of the feature wrapper approach is accuracy, because the feature selection is "tuned" for the classification method. Another advantage is that the approach provides some protection against overfitting because of the internal crossvalidation employed by the jackknife approach. Yet another advantage will become apparent when classifiers are employed to distinguish between more than two tissue types, because most feature selection methods used to date have been specific to binary classification. One drawback of feature wrapper methods is that the methods can be computationally intensive.

Searching the Space of Feature Sets

Ideally, all possible candidate feature sets are considered, but this is difficult for even modest numbers of candidate features. For example, exhaustively considering all feature sets of size 10 from among 100 candidate features would require evaluation of 2×10^{13} feature sets, and from among 10,000 would require 3×10^{33} evaluations. At one evaluation/per nanosecond, evaluating this many feature sets would take more than a million times the current age of the universe (Cayrel et al. 2001). Clearly, clever strategies are required to search through the space of feature sets. Xiong et al. evaluate two relatively simple search procedures, sequential forward search (SFS) and sequential forward floating search (SFFS). The sequential forward search procedure is:

(1) Choose the single best feature
(2) Choose the best feature set of size two that includes the feature from (1)
(3) Choose the best feature set of size three that includes the feature set from (2)
(4) And so on.

The sequential forward floating search procedure is similar, but allows for removal of the worst feature in the evolving feature set when this improves classifier performance.

Why is Feature Selection Important?

Feature selection is important because some methods of supervised learning perform inaccurately and/or slowly when asked to con-

sider a large number of features. But even when using classification algorithms that are good at handling many features (such as the SVM method used by Xiong et al.), there is a more practical concern. The method of typing cancers must be reliable, inexpensive, rapid, and easily performed for it to be employed by medical diagnostic laboratories. At the moment, these criteria exclude expression arrays with thousands of genes, and feature selection is required to reduce the feature set to a manageable number of genes.

Future Challenges

The major challenge in marker discovery is obtaining sufficient numbers of tissue samples, all collected in a uniform fashion and in such a way that mRNA is preserved. The importance of having enough patient samples is paramount. Imagine that you are posing 5000 yes-or-no questions to each of 10 patients and looking for those questions that distinguish ALL from AML. Further, imagine a worst-case scenario: The answers have nothing at all to do with AML and ALL, but were determined by patients flipping coins randomly. On average, there will be eight questions (features) whose answer correlates perfectly with ALL and AML in the training set of patients. Classifiers based on these eight features will fail utterly in predicting cancer type in new tissue samples.

Assuming that sufficient patient tissue samples are available, there may be many ways to improve the analysis methodology. Because expression arrays are notoriously error-prone, analysis will inevitably improve if knowledge of measurement error is incorporated. Different strategies for searching the space of candidate feature sets might improve feature selection. The space of candidate feature sets might be explored more exhaustively using massively distributed computing. Classifiers might be tuned to avoid prediction errors that are most costly to the welfare of patients. Classifiers that discriminate many cancer types, rather than just two at a time, have only recently emerged (Yeang et al. 2001), so we may expect future improvements here as well.

Future Prospects

Extensive efforts are underway in academia and the pharmaceutical industry to find markers for disease detection, typing, and choice of therapy for individual patients. The value of early disease detection is clear, especially in cancer. The reality of personalized medicine based on gene expression levels has hardly been proven, but may soon save patients from grueling chemotherapy regimens that are unlikely to succeed. Improved biomarkers may soon resurrect drugs that might otherwise fail in clinical trials because of toxic side effects, using biomarkers to predict which patients will have adverse reactions. These potential features have everyone excited.

REFERENCES

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. 2000. *Nature* **403:** 503–511.

Cayrel, R., Hill, V., Beers, T. C., Barbuy, B., Spite, M., Spite, F., Plez, B., Andersen, J., Bonifacio, P., Francois, P., et al. 2001. *Nature* **409:** 691–692.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. 1 999. *Science* **286:** 531–537.

Xiong, M., Fang, X., and Zhao, J. 2001. *Genome Res.*. **11:** 1878–1887.

Yeang, C.H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M., Lander, E., Mesirov, J., and Golub, T. 2001. *Bioinformatics* **17: Suppl 1** S316–S322.