Cancer Informatics

METHODOLOGY

Characterization of the Effectiveness of Reporting Lists of Small Feature Sets Relative to the Accuracy of the Prior Biological Knowledge

Chen Zhao[1], Michael L. Bittner[2] Robert S. Chapkin[3] and Edward R. Dougherty[1,2,4]

[1]Department of Electrical and Computer Engineering, Texas. A&M University, College Station, TX, 77843, USA. [2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, 85004, USA. [3]Center for Environmental and Rural Health, Texas A&M University, College Station, TX, 77843, USA. [4]Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston, TX, 77030, USA.
Email: edward@ece.tamu.edu

Abstract: When confronted with a small sample, feature-selection algorithms often fail to find good feature sets, a problem exacerbated for high-dimensional data and large feature sets. The problem is compounded by the fact that, if one obtains a feature set with a low error estimate, the estimate is unreliable because training-data-based error estimators typically perform poorly on small samples, exhibiting optimistic bias or high variance. One way around the problem is limit the number of features being considered, restrict features sets to sizes such that all feature sets can be examined by exhaustive search, and report a list of the best performing feature sets. If the list is short, then it greatly restricts the possible feature sets to be considered as candidates; however, one can expect the lowest error estimates obtained to be optimistically biased so that there may not be a close-to-optimal feature set on the list. This paper provides a power analysis of this methodology; in particular, it examines the kind of results one should expect to obtain relative to the length of the list and the number of discriminating features among those considered. Two measures are employed. The first is the probability that there is at least one feature set on the list whose true classification error is within some given tolerance of the best feature set and the second is the expected number of feature sets on the list whose true errors are within the given tolerance of the best feature set. These values are plotted as functions of the list length to generate power curves. The results show that, if the number of discriminating features is not too small—that is, the prior biological knowledge is not too poor—then one should expect, with high probability, to find good feature sets.
Availability: companion website at http://gsp.tamu.edu/Publications/supplementary/zhao09a/

Keywords: classification, feature ranking, ranking power

Cancer Informatics 2010:9 49–60

This article is available from http://www.la-press.com.

Introduction

One of the most challenging issues facing cancer informatics, and bioinformatics in general, is feature selection for classification. A typical microarray contains tens of thousands of genes from which a set must be chosen to form the features for whatever kind of discrimination is desired, be it for diagnosis or prognosis. Not only is feature selection problematic in high-dimensional settings; it is made more difficult by small samples, which are commonplace in high-throughput genomics and proteomics.

The supplementary table lists 20 cancer classification studies, showing the sample size, method of cross-validation error estimation, and classification problem for each. As we will discuss, finding and validating a single good feature set for these sample sizes is virtually impossible; on the other hand, reporting a list of small feature sets can, with high probability, assure one of obtaining one or more well-performing feature sets.

Numerous feature-selection algorithms have been proposed during the last few decades[1,2] and this number has increased dramatically since the advent of high-throughput genomic technology.[3,4] All feature-selection methods suffer from two problems, one inherent in the multivariate nature of classification and the other a consequence of sampling. First, even given the joint distribution of all the features and labels, if one wishes to select the best feature set of size k from among a family of n features, then all feature sets must be checked to be guaranteed that the best one is selected.[5] Nothing but an exhaustive search can assure finding the best feature set. Second, even if a feature-selection algorithm is capable of generally finding good feature sets given full knowledge of the joint distribution, in practice, feature selection must proceed from sample data and here even an exhaustive search can fail to produce a good feature set, even if one exists, the situation becoming more problematic as sample size decreases.

Perhaps the most well-known consequence of sample-based feature selection is the peaking phenomenon.[6–9] With full knowledge of the feature-label distribution, increasing the number of features cannot produce poorer classification; however, when using sample data, increasing the number of features beyond a point can degrade classification. For certain classification rules and feature-label distributions this point may consist of very few features when samples are small. Classically, peaking has been studied in the absence of feature selection, meaning that the features are added in a pre-determined manner. Peaking becomes far more complicated in the presence of feature selection.[10] Moreover, classical studies do not consider the extremely large numbers of features found in genomics, so that feature-selection performance for moderately large feature families must be re-examined in the framework of high-throughput biology.[11] In sum, the peaking phenomenon argues for limiting the number of features unless there is reason to believe that peaking will not be a problem.

When applying a feature-selection algorithm, given a classifier design rule, two basic related questions arise[12]: (a) Can one expect feature selection to yield a feature set whose error is close to that of an optimal feature set? (b) If a good feature set is not found, should one conclude that good feature sets do not exist? These questions translate quantitatively: (a) Given the error of an optimal feature set, what is the conditionally expected error of the selected feature set? (b) Given the error of the selected feature set, what is the conditionally expected error of an optimal feature set? Rather than using the conditional expectation, one can take a simpler route and look at the linear regression in both cases. For small samples it is commonplace to have very little regression in both cases and little correlation between the two errors. Thus, one cannot expect to find a close-to-optimal feature set or draw any conclusion regarding the existence of a good feature set when a good one is not found.[12]

Perhaps one might be fortunate and select a good feature set. But how would it be known that the feature set is good? Since the sample size is small, error estimation will be done on the training data and, when samples are small, error estimators such as cross-validation and bootstrap give generally poor results,[13] and the performance of these error estimators gets even worse when used in conjunction with feature selection.[14,15] The problem is that, with small samples and a large number of features (using feature selection or not), these error estimators have substantial variance, especially cross-validation, and they possess little correlation or regression with the true error.[16]

To address the dual problems of feature selection and error estimation, one can invoke two constraints. First, the number of potential features can be cut without using the data by restricting attention to features known to have some relation with the labels to be classified, say to a particular cancer of interest, and by not considering features whose measurements are suspect, say by throwing away features with missing values. Second, one can use small feature sets. Not only does this avoid the peaking phenomenon and enhance error estimation, it can also avoid feature selection altogether by facilitating an exhaustive search of feature sets. By only considering feature sets of size 1, 2, and 3, so long as the total number of features is not too large, one can test every feature set. Not only does such a restriction mitigate the statistical and computational issues, it also facilitates biological understanding. Studies have shown good classification can be achieved with 2 or 3 genes when re-examining data from studies that had originally used much larger feature sets,[17,18] with the advantage that the error estimates for the small gene sets are more credible.

While small feature sets help reduce the uncertainty introduced by feature selection and error estimation, even an exhaustive search with small feature sets does not fully overcome the problem.[19] Rather than report a single feature set when samples are small, reporting a list of the best performing feature sets increases the likelihood of finding good features sets, the idea being that some in the list of top-performing feature sets will be close to optimal. This strategy has been taken in a number of cancer classification studies that give lists of the best performing 1, 2, and 3 gene feature sets.[20–23] Given the list, one can either focus on the feature sets in the list for further sampling or take a classical wet-lab approach to determining which ones are predictive.

For illustration purposes, we briefly describe some results from a study that designed linear classifiers to distinguish four types of glioma (leaving details to the original paper): oligodendroglioma (OL), anaplastic oligodendroglioma (AO), anaplastic astrocytoma (AA), and glioblastoma multiforme (GM)—in particular, classification of OL from others, AO from others, AA from others, and GM from others.[20] The study involved 25 patients and the gene list was reduced to 597 genes prior to utilizing the data. Table 1 gives the best 5 single-gene sets and the best 10 two-gene sets based on the estimated errors. It also gives the three-gene sets among the top 50 three-gene sets for which the estimated error of the three-gene set is at least 0.03 less than the estimated error of its best two-gene subset. The purpose for placing this requirement on the marginal gain of a three-gene set over its two-gene subsets is to avoid redundancy caused by adjoining features to already strong performing feature sets. For each feature set, the table gives the error and the marginal gain.

Table 1. Errors and increments for discriminating AO from other gliomas.

Gene1	Gene2	Gene3	Error	Gain
DNaseX			0.1556	
TNFSF5			0.1658	
RAD50			0.1659	
HBEGF			0.1670	
NF45			0.1731	
DNaseX	*TNFSF5*		0.0750	0.0806
DNaseX	*PTGER4*		0.0784	0.0772
TNFSF5	*GNA13*		0.0826	0.0832
DNaseX	*HGF*		0.0892	0.0664
TNFRSF5	*PTGER4*		0.0907	0.0947
TNFSF5	*RAB5A*		0.0909	0.0749
TNFSF5	*SNF2L4*		0.0950	0.0708
erbB4	*PTGER4*		0.1012	0.0841
DNaseX	*β–PPT*		0.1013	0.0544
TNFSF5	*MERLIN*		0.1020	0.0638
DNaseX	*TNFSF5*	*RAB5A*	0.0441	0.0309
DNaseX	*TNFSRF5*	*PTGER4*	0.0454	0.0330
TNFSF5	*RAB5A*	*GNA13*	0.0464	0.0362
DNaseX	*PTGER4*	*SAP97*	0.0476	0.0308
TNFSF5	*GNA13*	*HGF*	0.0526	0.0300
TNFSF5	*β–PPT*	*PKAC–α*	0.0529	0.0549
DNaseX	*β–PPT*	*RkB*	0.0534	0.0479
TNFSF5	*PKAC–α*	*LIG4*	0.0591	0.0488
TNFSF5	*LIG4*	*HBGF–1*	0.0616	0.0474
TNFSF5	*β–PPT*	*SMARCA4*	0.0625	0.0325

The risk with this strategy is that, just as selecting a single feature set may be biased by a low error estimate, and thereby in actuality provide a poor feature set, an exhaustive list will be affected at the low end by optimistic error estimates and at the high end by pessimistic error estimates. The problem is illustrated in Figure 1, in which the x-axis shows the ranks of the selected feature sets and the y-axis shows their errors.

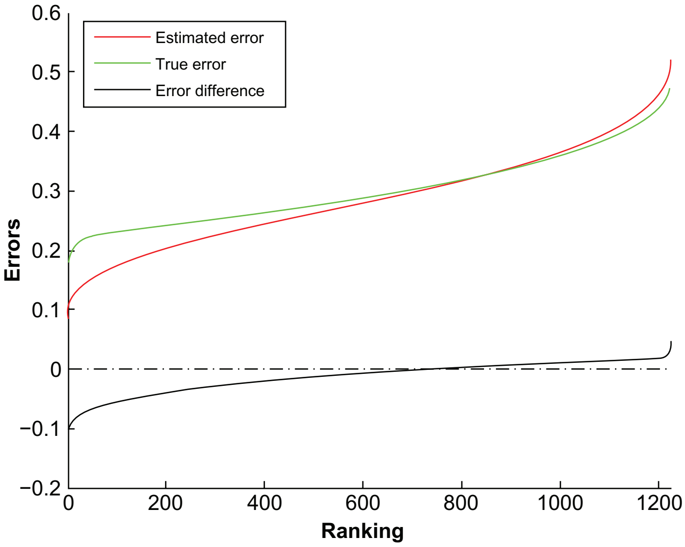

Figure 1. Error curves for a gaussian model.

The red, green, and black curves show the estimated error, true error and difference between the true and estimated errors, respectively. The curves are selecting 2 features out of 50 for linear discriminant analysis (LDA) classification in a Gaussian model. At the top end of the ranking the estimated errors are optimistic; at the low end, they are pessimistic. We only care about the top end. If the list is too short, then, there may not be any good features sets in the list; if the list is too long; then there will likely be good feature sets but the list will be impractically long.

This paper uses a model-based approach to investigate the kind of results that can be expected from generating feature-set lists. This is accomplished via a characterization of the goodness of the list relative to the number of potential features and the sample size. Since our purpose is to quantify the effects of forming a feature list and since quantification depends on the number of potential features selected by the biologist for consideration and the number of those features that are significant contributors to discrimination, only a model-based approach can provide meaningful results, since the contributive and non-contributive features are not known *a priori* for real data and could only be determined with certainty if we knew the distribution from which the real data arise, which we do not.

Systems and Methods

Ranking power

We define a measure of goodness for the list based on the closeness of the estimate-based feature sets to optimality. It depends on the feature-label distribution, the classification rule, the total number D of features, the number d of features to be selected, the sample size n, and the list length m. Let A_{best} be the best feature set relative to the feature-label distribution, $A_{(1)}, A_{(2)}, \ldots, A_{(m)}$ be a list of feature sets, ε_0 be the true error of the classifier for A_{best} designed on the sample, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ be the true errors, listed from lowest to highest, for $A_{(1)}, A_{(2)}, \ldots, A_{(m)}$ in the estimated-error list, and $r > 0$. We define the *ranking power* of the list by

$$\Delta^{n,r}_{D,d}(m) = P(\varepsilon_1 - \varepsilon_0 < r). \qquad (1)$$

The ranking power gives the probability that at least one feature set in the list has error within r of the best feature set. We would like $\Delta^{n,r}_{D,d}(m)$ to be close to 1 when m is relatively small. In practice, $A_{(1)}, A_{(2)}, \ldots,$ $\mathrm{A}_{(\mathrm{m})}$ are obtained by ranking the feature sets according to their estimated errors, $\hat{\varepsilon}_{(1)}, \hat{\varepsilon}_{(2)}, \ldots, \hat{\varepsilon}_{(m)}$, among all feature sets considered. Notice that the ith lowest estimated error $\hat{\varepsilon}_{(i)}$ corresponds to the feature set $A_{(i)}$ however, the ith lowest true error ε_i may not necessarily correspond to the feature set $A_{(i)}$. It could arise anywhere from the top m list. Our interest is to see if the top m list can produce at least a close to optimal feature set.

We consider *ranking power curves* $\Delta^{n,r}_{D,d}(m)$ as a function of m. For fixed D, d, and r, we are interested in the minimum m for which $\Delta^{n,r}_{D,d}(m) \geqslant 0:95$ (or some other threshold). For small r, the minimum m gives the length of the list required to get within r of the best feature set with probability 0.95.

In a sense $\Delta^{n,r}_{D,d}(m)$ represents a minimal measure of goodness because it requires at least one feature set satisfying the requirement. We can also consider the expected number of feature sets in the estimated-error list satisfying the requirement; to wit, we define

$$\bar{\Delta}^{n,r}_{D,d}(m) = E\left[\, card\{\varepsilon_i : \varepsilon_i - \varepsilon_0 < r\}\right], \qquad (2)$$

where *card* denotes cardinality (number of elements in the set).

In the remainder of this paper we will investigate properties of these power curves, in particular, how they are affected by D, d, and the number of marker features among the total number.

Model

We consider a hybrid Gaussian model M containing marker and noise features. The marker features come from a two-class Gaussian model M_μ with equally likely classes and class-conditional densities having common covariance matrix Σ_μ. One class mean is located at the origin $\vec{0}$ and the other at $\vec{\mu} = (a_1, a_2, \ldots, a_{D_\mu})^T$, D_μ is the total number of markers, these are divided among B blocks, and a_1, a_2, …, $a_{D\mu}$ are evenly spaced between 1 and 0.8, with $a_1 = 1$ and $a_{D\mu} = 0.8$. In this setting, every marker performs well, but not exactly the same.

The covariance matrix, Σ_μ, for M_μ is blocked:

$$\Sigma_\mu = \underbrace{\begin{bmatrix} \Sigma_\rho & 0 & \ldots & 0 \\ 0 & \Sigma_\rho & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Sigma_\rho \end{bmatrix}}_{B\ blocks}$$

where Σ_ρ has variance σ^2_μ along the diagonal and correlation coefficient ρ off the diagonal. Features from the same block have the same correlation, ρ, while features from different blocks are uncorrelated, thereby simulating the situation where genes in the same pathway are strongly correlated but those in different pathways are uncorrelated.

The noise features are modeled as zero-mean random Gaussian noise, with a total of D_n noise features, with $D_n >> D_\mu$. For the hybrid model M, there is a total of $D = D_0 + D_1$ features. D_0 reflects biologists' input to the classification problem and the larger D_0, the better the classification. In this vein, we will compare the ranking power with respect to different D and D_0. In the simulations, we will assume a given D_0, the size of D_0 reflecting the extent of the prior knowledge, and then D_0 and D_1 features will be randomly chosen from the useful and noise features, respectively.

Given the adopted model, for any feature set of size d, the Bayes classifier and corresponding Bayes error can be found. Assuming equal covariance matrices and equal prior probabilities for the classes, the Bayes error for feature set A is given by $\Phi(-\Delta/2)$, where Φ is the standard normal cumulative distribution function and Δ is the Mahalanobis distance between the class-conditional distributions corresponding to A. A_{best} is the feature set possessing minimum Bayes error, equivalently, the largest Mahalanobis distance. Note that we do not have to consider noise features when searching for the best feature set.

Two points regarding A_{best} should be recognized. First, although A_{best} provides the minimum error relative to the feature-label distribution, we design classifiers from samples and the error of the designed classifier for A_{best} may not be minimal for a given sample; nevertheless, A_{best} represents the gold standard for the feature-label distribution and therefore we use the error of its designed classifier as the benchmark. Owing to the direction of the inequality in Eqs. 1 and 2, no problem arises should the error of the classifier designed for A_{best} not be minimal. Second, A_{best} is designed relative to D_μ, not D_0. Once the classification problem is given, A_{best} should not change because it indicates the best we can do for the problem at hand. D_0 depends on the extent of the prior knowledge and the poorer that knowledge, the poorer we should expect to do compared to A_{best}.

To find the true error of a designed classifier we generate a very large test set of independent data from the feature-label distribution and compute its error rate on the test set. Estimated errors are computed by bolstered resubstitution, which has been shown to perform well in comparison with other training-sample-based error estimators when it comes to ranking feature sets.[19]

To illustrate the advantage of obtaining a list of feature sets by exhaustive search, we compare the list to ordinary feature selection by computing

$$\Omega^{n,r}_{D,d}(m) = P(\varepsilon_{FS} - \varepsilon_0 < r), \tag{3}$$

where ε_{FS} is the true error of the feature set found by feature selection. We use a two-stage approach for feature selection. The t-test is used in the first stage to reduce the number of features and sequential forward search (SFS) is use in the second stage to arrive at a feature set.

Implementation

We focus on the LDA classification rule, the power curve method being applicable to any classification rule. We utilize the following simulation procedures:

Compute $\Delta^{n,r}_{D,d}(m)$:

1. Set up M_μ and M_n, and determine A_{best} from M_μ.
2. Randomly select D_0 features from D_μ and D_1 features from D_n, and set $D = D_0 + D_1$, the number of features in the hybrid model M.
3. Generate n-point sample sets for M_μ and D_n ($n/2$ samples per class) to obtain an n-point sample T for M.
4. Compute the true error, ε_0, for A_{best} using the samples from M_μ.
5. For every feature set of size d, design a classifier from T.
6. Compute the true and estimated errors for the classifiers from step (5).
7. Rank all the feature sets by their estimated errors to get the top m estimated-error list.
8. Select the feature set in the list with the lowest true error, ε_1.
9. If $\varepsilon_1 - \varepsilon_0 < r$, set $count = count + 1$.
10. Repeat steps (2) through (9) a total of N times to get $\Delta_{D,d}^{n,r}(m) = count/N$.

Compute $\Omega_{D,d}^{n,r}(m)$:

1–4. Repeat the steps of the procedure Compute $\Delta_{D,d}^{n,r}(m)$.
5. Use the t-test to select the first d_{stage1} features and then use SFS to select the final d features.
6. Find the true error, ε_{FS}, of the designed classifier for the feature set found in step (5).
7. If $\varepsilon_{FS} - \varepsilon_0 < r$, set $count = count + 1$.
8. Repeat steps (3) through (7) a total of N times to get $\Omega_{D,d}^{n,r}(m) = count/N$.

A summary of the experimental parameters is provided in Table 2.

Experimental Results

For a given model we are mainly interested in the effects of D, D_0, and d. As should be expected, larger sample sizes will produce better results. In the paper we restrict ourselves to $n = 40$, results for $n = 60$ being given on the companion web-site. Except when otherwise specified, $\sigma_\mu^2 = 1$ and $\rho = 0.8$. From an experimental perspective, D is the size of the feature list provided by the biologist, D_0 is the number of marker features in the provided list, and d is the feature set size. Each graph of Figure 2 shows power curves $\Delta_{D,d}^{n,r}(m)$ for $r = 0.03, 0.05, 0.07$ for a value of $D = 50, 100, 150$, a value of $D_0 = 20, 10, 5$, and a value of $d = 2$ (part a) or $d = 3$ (part b). The x-axis gives the number m of feature sets in the list, the y-axis gives the probability $\Delta_{D,d}^{n,r}(m)$, and the horizontal dotted lines mark probability 0.95. Figure 3 shows corresponding curves for $\bar{\Delta}_{D,d}^{n,r}(m)$. For fixed D, as D_0 decreases, the power decreases, which reflects the fact that the list provided by the biologist contains fewer markers. Matters only get bad when $D_0 = 5$, which means the prior information is very poor. For instance, when $D = 150$ and $D_0 = 5$, only 3% of the suggested features are markers.

A related effect concerning prior knowledge has been discussed.[24] In current molecular epidemiology studies, it is common to claim an association between a genetic variant and a disease when the corresponding P value is below a certain level, say 0.05. However, the false positive report probability (FPRP) can be close to 1 if the prior probability of the true association is below 0.01. In our paper, this corresponds to the situation where a low percentage of suggested features are markers and, consequently, the power curves will be significantly lowered (Fig. 2).

The effect of increasing D and d is seen in Figure 4, in which the top and bottom rows correspond to $d = 2, 3$, respectively, the columns, left to right, correspond to $D = 50, 100, 150$, and $D_0 = 20$. The graphs show the error curves for ranking up to rank $m = 400$, the horizontal axis being the estimated-error rank and the red, green and black curves showing the estimated error, the true error and the difference between the errors, respectively. The detrimental aspect of

Table 2. Summary of experiments. ES stands for exhaustive search methods and FS stands for feature selection methods.

Exp	n	D_μ	D_n	ρ	B	D	D_0	d	r
ES 1	40, 60	50	1000	0.8	5	50	50, 20, 10, 5	2, 3	0.01 – 0.1
ES 2	40, 60	50	1000	0.8	5	100	50, 20, 10, 5	2, 3	0.01 – 0.1
ES 3	40, 60	50	1000	0.8	5	150	50, 20, 10, 5	2, 3	0.01 – 0.1
FS 1	40, 60	50	1000	0.8	5	1050	50	2, 3, 4, 5	0.01 – 0.1

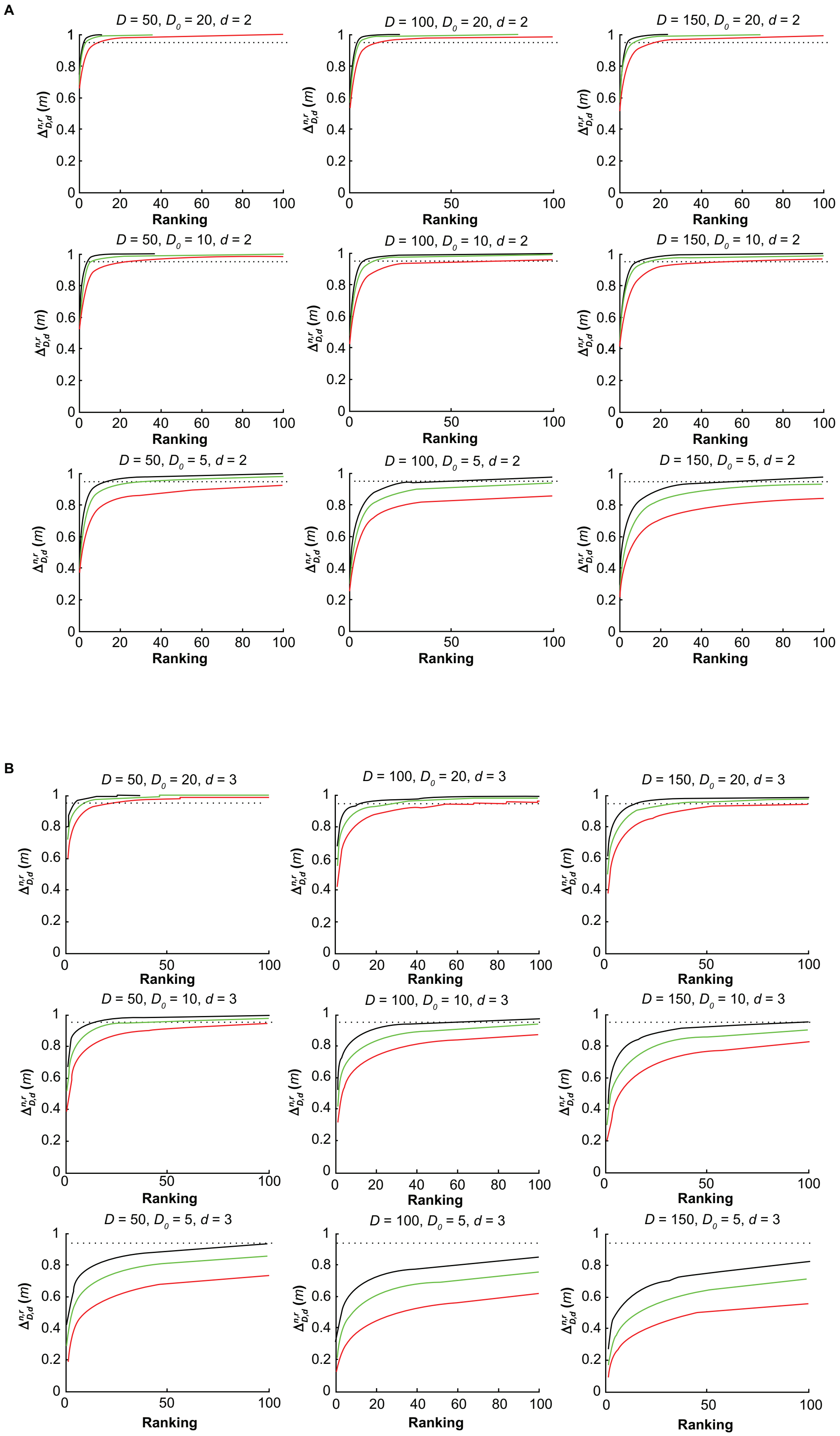

Figure 2. Power curves for different model parameters. Red: $r = 0.03$, green: $r = 0.05$, black: $r = 0.07$.

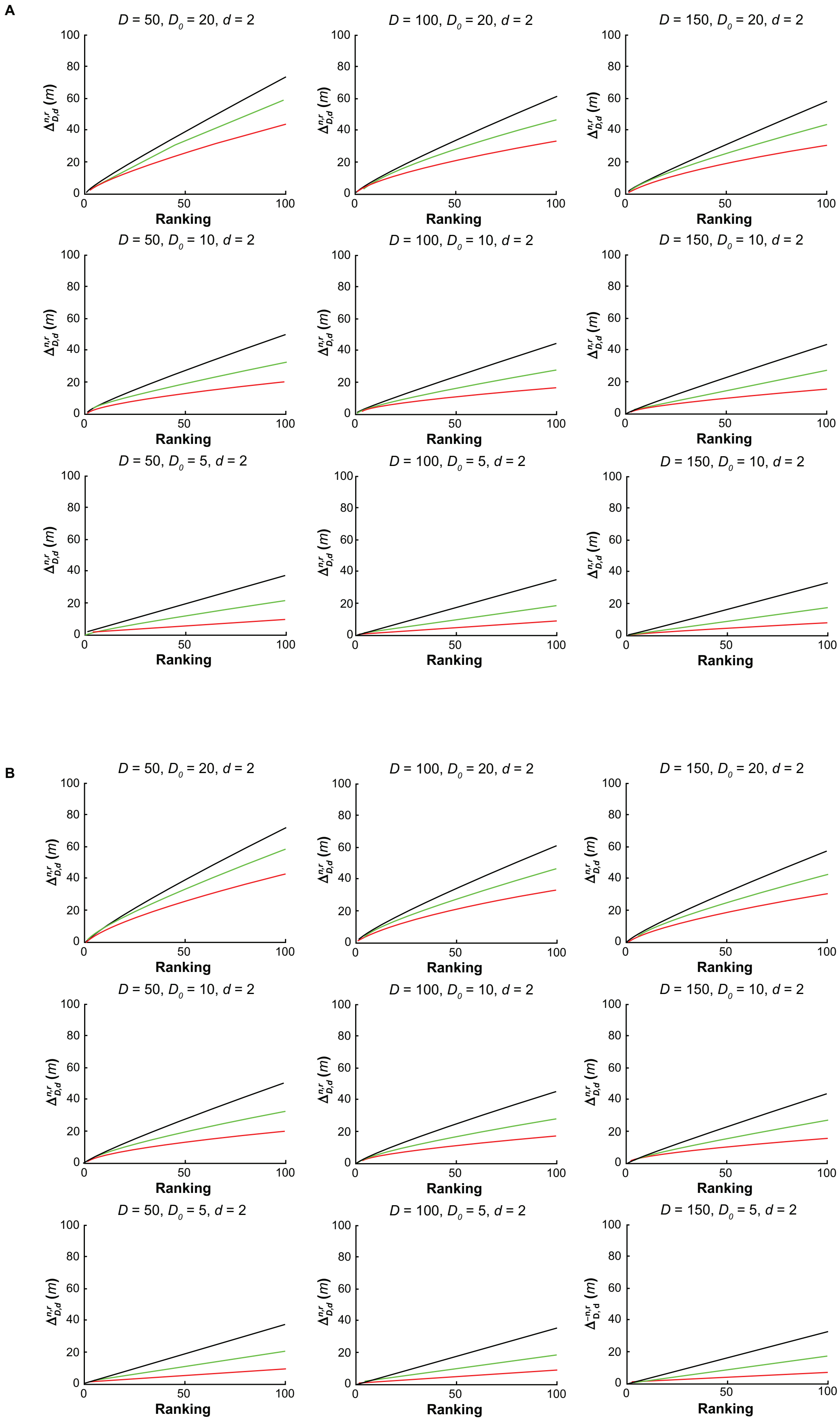

Figure 3. Average good features for different model parameters. Red: $r = 0.03$, green: $r = 0.05$, black: $r = 0.07$.

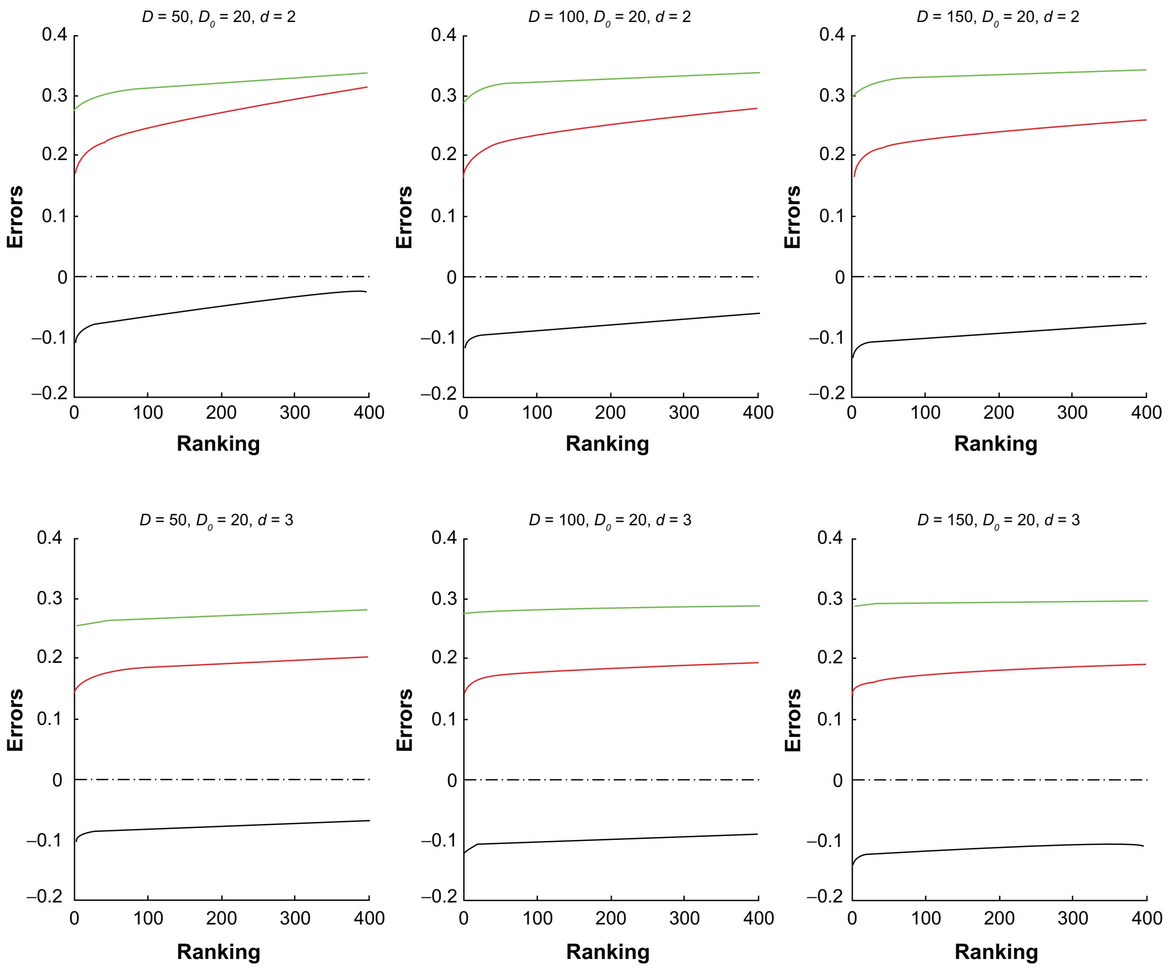

Figure 4. Error difference curves. Green: true error, red: estimated error, black: error difference.

larger D and d is seen in the difference curves, which show the increased optimistic bias for larger D and d resulting from ranking according to estimated error. Better discovery may result from using more features but one must expect greater optimistic bias in the results.

Figure 5 illustrates the effect of variance in the marker model M_μ. We set $D = 150$, $D_0 = 10$, $d = 2$ $\rho = 0.8$, and $r = 0.05$, and show power curves for $\sigma_\mu^2 = 0.5$, 1, and 2. As the variance grows, the power decreases, reflecting greater difficulty in finding the top feature sets.

Figure 6 shows the effect of a different number of blocks, a larger number of blocks representing features that are spread among more pathways. We set $D = 150$, $D_0 = 10$, $d = 2$, $\rho = 0.8$, $\sigma_\mu^2 = 1$, and $r = 0.05$, and show power curves for $B = 2$, 5, and 10. It is easier to find good features with more blocks, since the features are then less correlated.

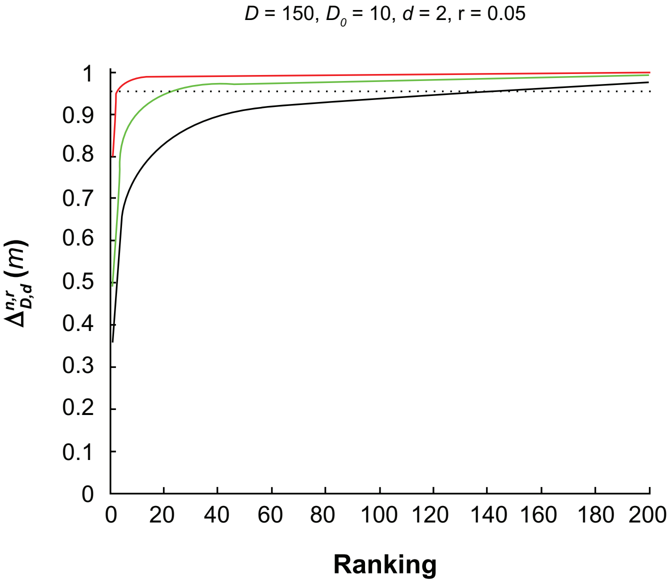

Figure 5. Power Curves for $\sigma_\mu^2 = 0.5$, 1.0, 2.0, respectively. Red: $\sigma_\mu^2 = 0.5$, green: $\sigma_\mu^2 = 1.0$, black: $\sigma_\mu^2 = 2.0$.

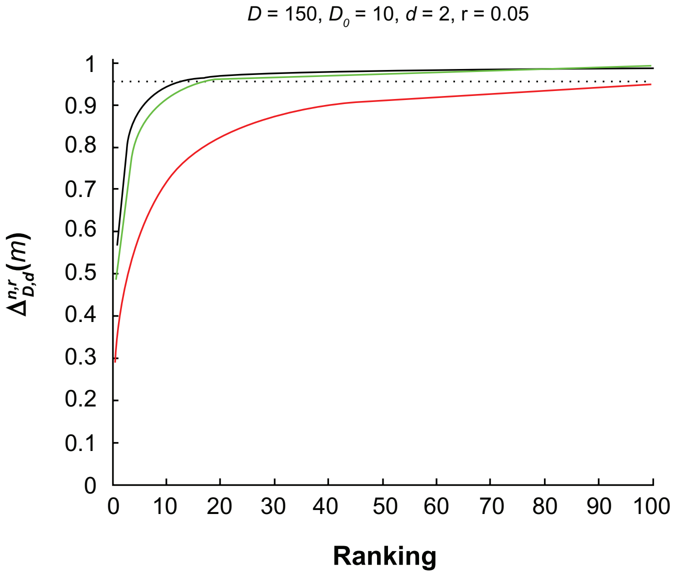

Figure 6. Power curves for $B = 2, 5, 10$, respectively. Red: $B = 2$, green: $B = 5$, black: $B = 10$.

Figure 7 shows the effect of correlation. We set $D = 150$, $D_0 = 10$, $d = 2$, $\sigma^2_\mu = 1$, and $r = 0.05$, and show power curves for $\rho = 0.1$, 0.5, and 0.8. Increasing the correlation makes it slightly harder to find good features.

It is informative to compare the ability of feature-set lists created by exhaustive search to what one obtains using feature selection. Table 3 shows the probability of finding a feature set whose error is within r of the error for A_{best} when using the t-test to reduce the original 1050 features, 50 markers and 1000 noise features, to 100 features, followed by SFS to reduce to a final feature set of 2, 3, 4, or 5 features, for sample size $n = 40$. Even at $r = 0.05$, the probability of finding a satisfactory feature set of size $d = 3$ is only 0.309. The difficulty is exemplified in Figure 8, which shows the regression of the error, ε_{FS}, for the selected feature set on the error, ε_{best}, for A_{best} and the regression of ε_{best} on ε_{FS}. Not only is there little regression in either case, the dispersion of the scatter plots is substantial.

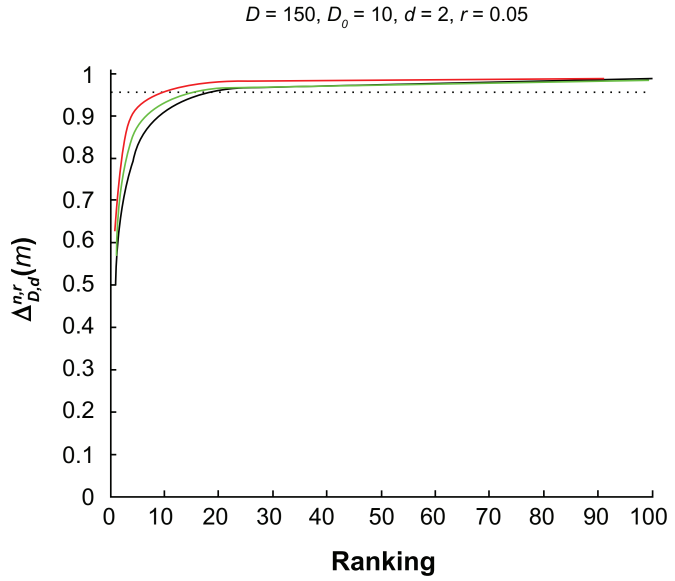

Figure 7. Power curves for $\rho = 0.1$; 0.5; 0.8, respectively. Red: $\rho = 0.1$, green: $\rho = 0.5$, black: $\rho = 0.8$.

Conclusion

The problem of developing a classifier based on biological features such as SNPs, patterns of differential gene expression, or differential protein abundances, with sufficient sensitivity and specificity for medical use has proven much harder than originally hoped. Indeed, a current question frequently raised in the field is whether it is possible to find features that are even of the same reliability as the clinical features developed by physicians in the course of treating particular

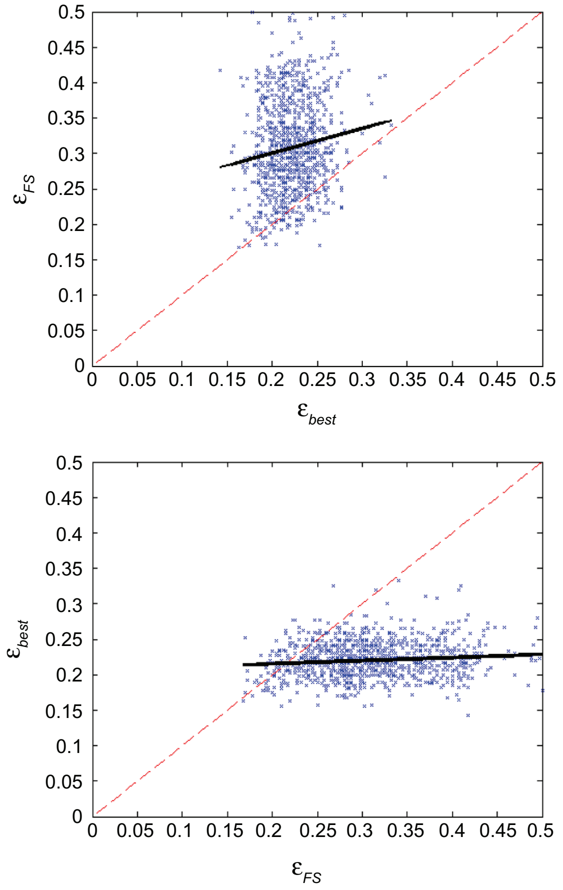

Figure 8. Scatter plot and robust linear regression line (bold line) for feature selection, $d = 3$.

diseases. One reason for the current low levels of production of these molecularly-based classifiers is that a researcher attempting to discover such features typically starts by using a "non-biased" approach where as many features are tested for candidacy as is possible with the current high-throughput methodologies appropriate for the kind of marker being studied. Such an approach has the virtue of not missing the opportunity to evaluate a candidate feature that might have been missed if a smaller set of features were examined. The approach carries the liability of not being able to recognize predictive features because the availability of sample data and the cost of the analysis has restricted the study to examining too few individuals from the study populations to provide a reliable estimate of the discriminatory power of the features. A validation study of a large number of SNP associations with acute cardiovascular syndromes based on sample sets containing no more than hundreds of sample points has produced the most explicit example of how devastating the inability to accurately estimate predictive error in such sample sets can be.[25]

This study further illustrates the degree of difficulty of producing classifiers by showing that even under highly idealized, mathematically favorable conditions—two Gaussian classes sharing the same uncorrelated-block covariance matrix and zero-mean random Gaussian noise—it is still difficult to find a close-to-optimal feature set with small samples (Table 3). These idealized conditions are far from most biological situations, since the array of cellular processes that play a role in a disease are highly integrated with other processes, have varying degrees of correlations with other cellular processes, and must be searched for not in a milieu of random operations, but in a setting of other processes that do not display a great deal of randomness.

The findings in this study suggest an alternative to the current sequence of operations employed in "non-biased" discovery methods that start with a small sample set and many thousands of candidate features. The suggestion is to apply as much biological prior knowledge and insight about the disease as possible to limit the list of features to be examined in the study. In "non-biased" discovery, prior knowledge and insight are applied after the classifier design process has been run in order to further prioritize the candidates for validation studies. If the process has failed to nominate predictive features that can be recognized on the basis of the possible fit of their functions with the disease, then they will be lost to the analyst. Alternatively, if the biologist had correctly proposed as few as five valid features as candidates for analysis in a list of fifty such candidates, then even in an examination of as small as fifty sample points, then the analyst would have an excellent chance of recognizing the predictive power of several of those features (Fig. 2).

Table 3. Probability of finding a good feature set using feature selection: $\Omega_{D,d}^{n,r}(m)$.

	d **= 2**	***d*** **= 3**	***d*** **= 4**	***d*** **= 5**
r = 0.02	0.3490	0.1580	0.0740	0.0390
r = 0.03	0.4140	0.2090	0.1100	0.0560
r = 0.04	0.4630	0.2510	0.1360	0.0840
r = 0.05	0.5020	0.3090	0.1770	0.1260

This and previous studies[9,13,16] suggest that experimental campaigns to establish classifiers with the predictive strength required for medical use can be defensibly mounted in two ways. For those indications where the existing biological knowledge is insufficiently certain that well-informed experts feel that they are unlikely to be able to produce a candidate list of one to two hundred genes with five to ten percent accuracy and where a predictive classifier would produce sufficient benefit to justify a well-powered study (thousands of sample points), a large study examining many features could be carried out. If, on the other hand, the indication is well enough studied that a candidate list with a hit rate of five to ten percent could be expected to be developed, a modest sample set is available, and the classifier would justify a smaller-scale study, then this type of study could be run. Studies between these bounds, which lack either sufficient sample size to successfully recognize predictive features or sufficient prior knowledge to reduce the set of candidate features to a level where they can be successfully identified with smaller sample sets, are unlikely to provide classifiers that could be validated and should be avoided.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors report no conflicts of interest.

References

1. Jain A, Zonger D. Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Trans Pattern Analysis and Machine Intelligence*. 1997;19:153–8.
2. Kudo M, Sklansky J. Comparison of Algorithms that Select Features for Pattern Classifiers. *Pattern Recognition*. 2000;33:25–41.
3. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23:2507–17.
4. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinformatics*. 2008;9:392–403.
5. Cover T, Campenhout JV. On the Possible Orderings in the Measurement Selection Problem. *IEEE Trans on Systems Man and Cybernetics*. 1977;7:657–61.
6. Hughes GF. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Trans*. *Information Theory*. 1968;14:55–63.
7. Jain AK, Waller WG. On the Optimal Number of Features in the Classification of Multivariate Gaussian Data. *Pattern Recognition*. 1978;10:365–74.
8. Hua J, Xiong Z, Dougherty ER. Determination of the Optimal Number of Features for Quadratic Discriminant Analysis Via the Normal Approximation to the Discriminant Distribution. *Pattern Recognition*. 2005a;38:403–21.
9. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*. 2005b;21:1509–15.
10. Sima C, Dougherty ER. The Peaking Phenomenon in the Presence of Feature Selection. *Pattern Recognition Letters*. 2008;29(11):1667–74.
11. Hua J, Waibhav T, Dougherty ER. Performance of Feature Selection Methods in the Classification of High-Dimensional Data. *Pattern Recognition*. 2009;42(3):409–424.
12. Sima C, Dougherty ER. What should be expected from feature selection in small-sample settings. *Bioinformatics*. 2006;22:2430–6.
13. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*. 2004;20:374–80.
14. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21:3301–7.
15. Xiao Y, Hua J, Dougherty ER. Quantification of the impact of feature selection on the variance of cross-validation error estimation. *EURASIP J Bioinform Syst Biol*. 2007;16354.
16. Hanczar B, Hua J, Dougherty ER. Decorrelation of the true and estimated classifier errors in high-dimensional settings. *EURASIP J Bioinform Syst Biol*. 2007;38473.
17. Grate LR. Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. *BMC Bioinformatics*. 2005;6:97.
18. Braga-Neto UM. Fads and Fallacies in the Name of Small-Sample Microarray Classification. *IEEE Signal Processing Magazine*. 2007;24(1):91–9.
19. Sima C, Braga-Neto U, Dougherty ER. Superior feature-set ranking for small samples using bolstered error estimation. *Bioinformatics*. 2005; 21:1046–54.
20. Kim S, Dougherty ER, Shmulevich I, et al. Identification of combination gene sets for glioma classification. *Mol Cancer Ther*. 2002;1:1229–36.
21. Morikawa J, Li H, Kim S, et al. Identification of signature genes by microarray for acute myeloid leukemia without maturation and acute promyelocytic leukemia with t(15;17)(q22;q12)(PML/RARalpha). *Int J Oncol*. 2003;23:617–25.
22. Kobayashi T, Yamaguchi M, Kim S, et al. Gene Expression Profiling Identifies Strong Feature Genes that Classify *de novo* $CD5^+$ and $CD5^-$ Diffuse Large B-cell Lymphoma and Mantle Cell Lymphoma. *Cancer Research*. 2003;63:60–6.
23. Zhao C, Ivanov I, Dougherty ER, et al. Non-invasive Detection of Candidate Molecular Biomarkers in Subjects with a History of Insulin Resistance and Colorectal Adenomas. *Cancer Prevention Research*. 2009;2(6):590–7.
24. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004;96:434–42.
25. Morgan TM, Krumholz HM, Lifton RP, Spertus JA. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA*. 2007;297:1551–61.