A Stochastic Model for Optimizing Composite Predictors Based on Gene Expression Profiles

Murali Ramanathan[1,2]

Received December 6, 2002; accepted April 2, 2003

Purpose. This project was done to develop a mathematical model for optimizing composite predictors based on gene expression profiles from DNA arrays and proteomics.

Methods. The problem was amenable to a formulation and solution analogous to the portfolio optimization problem in mathematical finance: it requires the optimization of a quadratic function subject to linear constraints. The performance of the approach was compared to that of neighborhood analysis using a data set containing cDNA array-derived gene expression profiles from 14 multiple sclerosis patients receiving intramuscular inteferon-β1a.

Results. The Markowitz portfolio model predicts that the covariance between genes can be exploited to construct an efficient composite. The model predicts that a composite is not needed for maximizing the mean value of a treatment effect: only a single gene is needed, but the usefulness of the effect measure may be compromised by high variability. The model optimized the composite to yield the highest mean for a given level of variability or the least variability for a given mean level. The choices that meet this optimization criteria lie on a curve of composite mean vs. composite variability plot referred to as the "efficient frontier." When a composite is constructed using the model, it outperforms the composite constructed using the neighborhood analysis method.

Conclusions. The Markowitz portfolio model may find potential applications in constructing composite biomarkers and in the pharmacogenomic modeling of treatment effects derived from gene expression endpoints.

KEY WORDS: biomarkers; composite; Markowitz portfolio; pharmacogenomic modeling.

INTRODUCTION

Genomic technologies such as DNA arrays and proteomics are now capable of simultaneously measuring the expression of thousands of genes from single samples. These techniques are being increasingly applied to problems in disease diagnosis, drug response, and pharmacodynamics.

Such massively multiplexed measurement systems provide a broad snapshot of the state of the biologic system that is particularly useful for hypothesis generation. The genes and expression patterns that specifically characterize the disease or treatment—informative genes and expression patterns—must then be gleaned from this snapshot. However, the mining and analysis of gene expression profiling data sets can be formidable because they will contain measurements on both informative and uninformative genes. The initial analysis challenge often is to identify the subset of informative genes; in principle, this can be accomplished by selecting genes that meet statistical significance criteria in an appropriate test. The statistical issues are complicated by the varied sources of random error and bias and the large number of multiple comparisons involved in array experiments. However, these issues have been addressed in the literature (1–3).

Here, an approach for the important downstream problem of identifying an optimal subset of genes for the response prediction problem is formulated and developed. The solution will yield the weights in which the expression of individual genes should be combined to yield an optimal composite treatment effect measure that is based on multiple genomic endpoints.

The most commonly used approach for constructing such composite pharmacodynamic measures is neighborhood analysis (4). In this approach, the subset of genes with strongest (positive or negative) associations with the class distinction of interest is selected for the composite measure. The normalized expression values of these genes are linearly combined in proportion to the strength of the association to provide a composite measure. The results will show that the model proposed yields composite measures that are distinct from and superior to those from neighborhood analysis.

MATERIALS AND METHODS

Study Population

With informed consent, peripheral blood anticoagulated with heparin was obtained by venipuncture from 14 consecutive patients (10 female, 4 male; mean age 42.2 years, standard deviation 9.23 years) with active relapsing remitting MS [Expanded Disability Status Scale (EDSS) range 1–4] who were to receive IFN-β therapy. Patients had not previously received IFN-β and were clinically stable for the preceding 4 weeks. Additional blood samples were obtained from these patients 24 h after the first IM administration of 30 μg interferon-β1a (IFN-β1a, Avonex®). A total of 14 pairs of samples were analyzed using the method.

Total RNA Preparation

Peripheral blood mononuclear cells were isolated from the anticoagulated blood within 4 h of collection using the Hypaque-Ficoll method (Histopaque reagent, Sigma Chemical, St. Louis, MO). Monocytes were depleted from the peripheral blood mononuclear cells by the plastic adhesion method. Total RNA was prepared from the monocyte-depleted peripheral blood mononuclear cells using the TRI reagent method (Molecular Research Center Inc., Cincinnati, OH) (5).

DNA Array Protocol

The DNA arrays, specifically GeneFilters GF211, from Research Genetics Inc. (Huntsville, AL), which contain approximately 4,000 named human genes, were used. GeneFilters contain 5,188 spots, each with 5 ng of an approximately 1,000 base long, 5′-end-derived PCR fragments. Briefly, 5 μg of total RNA was radioactively labeled with ^{33}P CTP using reverse transcriptase. The labeled cDNA was used to probe the GF211 membrane. The membranes were washed, and the

[1] Department of Pharmaceutical Sciences, State University of New York at Buffalo, Buffalo, New York 14260-1200.

[2] To whom correspondence should be addressed. (e-mail murali@acsu.buffalo.edu)

0724-8741/03/0700-0996/0

bound radioactivity was visualized using a Cyclone phosphorimager (Packard Instrument Company, Meriden, CT).

Data Analysis

The images from the phosphorimager were imported directly into Pathways software program (Research Genetics Inc., Huntsville, AL). The images were aligned, gridded, and quantified according to recommended procedures. Filters were normalized using the intensity from all spots, and the software also normalized for intensity ranges in two-filter comparisons. The data were exported to an Excel spreadsheet (Microsoft Corp., Bellevue, WA) for further analysis. Statistical analysis was carried out using Excel. An α-value of 0.05 was used to determine statistical significance.

Before modeling, the pre- and posttreatment expression changes of three genes, β_2 microglobulin, signal transducer, and activator of transcription-1 and 1-8D, which are known to be IFN-β inducible, were compared using the paired *t* test. The paired *t* test demonstrated that the expression levels of all three IFN-β-inducible genes were significantly higher in the posttreatment samples, confirming that the treatment had evoked a pharmacologic response.

Data Transformations and Terminology

All calculations for neighborhood analysis and in our model used expression data that were natural logarithm transformed.

In this analysis, change in logarithm-transformed gene expression was treated analogously to "assets" in mathematical finance. The fractional change of gene expression E_i for the *i*th gene on treatment is defined in terms of the pre- and posttreatment expression levels X_i, as:

$$E_i = ln\left(\frac{X_{i,after\ treatment}}{X_{i,before\ treatment}}\right) \textit{ for up-regulated genes}$$

$$E_i = -ln\left(\frac{X_{i,after\ treatment}}{X_{i,before\ treatment}}\right) \textit{ for down-regulated genes}$$

Neighborhood Analysis

The method of Golub *et al.* was used (4). The strength of association, P_i, between the expression of the *i*th gene and the treatment effect was measured using the formula:

$$P_i = \frac{\mu_{i,after\ treatment} - \mu_{i,before\ treatment}}{\sigma_{i,after\ treatment} + \sigma_{i,before\ treatment}} \textit{ for up-regulated genes}$$

$$P_i = -\frac{\mu_{i,after\ treatment} - \mu_{i,before\ treatment}}{\sigma_{i,after\ treatment} + \sigma_{i,before\ treatment}} \textit{ for down-regulated genes}$$

The $\mu_{i,before\ treatment}$ and $\mu_{i,\ after\ treatment}$ are mean values, whereas $\sigma_{i,before\ treatment}$ and $\sigma_{i,after\ treatment}$ are the standard deviations of the natural logarithm–transformed expression levels for the *i*th gene before and after treatment, respectively.

The genes in the 0.5th and 99.5th percentiles were used to construct the neighborhood analysis predictor. Because the array contained approximately 4,000 known genes, the composite predictor consisted of the 20 IFN-β1a up-regulated genes and 20 down-regulated genes with largest magnitudes for P_i. There were several known IFN-regulated genes in the predictor, and further follow-up validation studies are under way for many of the predictor genes that had not previously been reported as being IFN-β modulated. The identity of these genes (data not shown) is not critical to the validity of the Markowitz model methodology.

The E_i change after treatment of each gene was statistically significant at $\alpha = 0.05$ in a two-tailed, paired *t* test. We conducted power calculations using Cohen's *d* with pooled standard deviations based on independent groups (6). The minimum value of Cohen's *d* was 0.9 (range 0.9–2.0), and for 14 samples, the estimate for power is approximately 0.85.

For each patient, *j*, the composite neighborhood analysis predictor, η_j, was calculated using P_i and the $E_{i,j}$ using:

$$\eta_j = \sum_{i=1}^{i=40} P_i E_{i,j}$$

Thus, in neighborhood analysis, logarithm of the expression level for each gene was weighted in proportion to the strength of its P_i value.

Markowitz Portfolio Calculations

The portfolio problem was solved using a quadratic programming approach. The up-regulated and down-regulated genes in the neighborhood analysis composite were treated as "assets," and the $\overline{E}_i$ values were used for the vector of asset returns.

The program available online at *http://www.iimahd.ernet.in/~jrvarma/software/* was used. The algorithm used by the program is based on Markowitz's approach described in Ref. 7.

Assessing Predictive Capabilities

The predictive capabilities of the Markowitz portfolio model were compared to neighborhood analysis using a variation of the leave-one-out method.

The ability of each composite to predict whether a given patient's sample belonged to the pretreatment or posttreatment class was assessed by withholding data corresponding to the individual patient. The weights of the genes in the Markowitz portfolio and neighborhood analysis composites were kept fixed at values calculated using the composite dataset. For each sample *j*, from the *k*th patient, the following predictor value, *Predictor* $_j$, was calculated:

$$Predictor_j = \sum_{i=1}^{i=40} w_i (X_{i,j} - \overline{X}_i)$$

The w_i are the weights of the *i*th gene in the composite of interest; $X_{i,j}$ is the log-transformed expression value of the *i*th gene in the *j*th sample; and $\overline{X}$ is the "leave-one-out mean" of the *i*th gene in the dataset of pretreatment and posttreatment expression values but without the *k*th patient.

The sample was assigned to the pretreatment class if the predictor was negative and to the posttreatment class if the predictor was positive.

DERIVATIONS AND RESULTS

Assessment of Treatment Effects in a Model for Two Genes

For illustrative purposes, I first analyze a drug study in which a pretreatment sample and a posttreatment sample are obtained at time, $t = T$, after drug administration. The presentation of this case is adapted from the treatment by Wilmott (8) and is intended to facilitate further development of the derivation, allow visualization of the results and to highlight the critical variables.

If there are only two genes, 1 and 2, in the composite M with mean fractional changes, $\overline{E}_1$ and $\overline{E}_2$, weights w_1 and w_1, standard deviation of mean fractional changes, σ_1 and σ_2, and correlation coefficients, ρ, in the composite, the mean, $\overline{M}$, and variance, $\sigma^2{}_M$, of the composite containing logarithm-transformed expression values are:

$$\overline{M} = w_1\overline{E}_1 + (1 - w_1)\overline{E}_2 \quad (1)$$

$$\sigma_M^2 = w_1^2\sigma_1^2 + 2(1 - w_1)w_1\rho\sigma_1\sigma_2 + (1 - w_1)^2\sigma_2^2 \quad (2)$$

Because uncertainty is present, one cannot focus on maximizing $\overline{M}$ alone because this could potentially result in a highly variable composite measure. This makes it necessary for the optimization to identify points that have the highest $\overline{M}$ value for a given σ_M or the lowest σ_M for a given $\overline{M}$.

The implications of this model are best visualized in a plot of $\overline{M}$ vs. σ_M (Fig. 1). The lines are often referred to as the "critical lines," and the analysis as "mean-variance analysis." The $\overline{M}$ vs. σ_M plot is a hyperbola for the two-gene case. The bold portion of the critical line is referred to as the "efficient frontier" because all combinations of the two genes on the bold line are efficient, and combinations not on the line are inefficient. For a given correlation coefficient value, the portion of the critical line shown in dashed gray line is not efficient because by moving to the bold portion of the line, the user can further reduce the standard deviation of the composite, σ_M, at a given M or increase $\overline{M}$ at a given standard deviation.

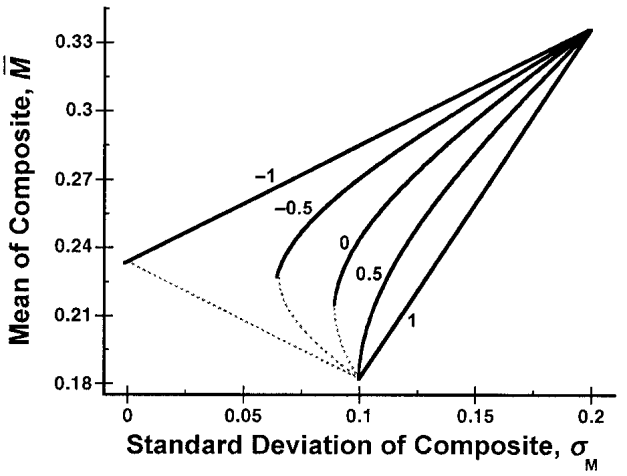

Fig 1. Figure 1 shows the Markowitz critical line for the two-gene case. The fold changes in expression levels of the two genes were set to 1.2 and 1.4, respectively. The corresponding E_i values were 0.182 and 0.336, and the standard deviations were 0.1 and 0.2, respectively. The Markowitz critical line is graphed for the five correlation coefficient values, 1, 0.5, 0, –0.5, –1, indicated. The solid line represents the efficient portion of the frontier, and the gray regions are not efficient.

Several counterintuitive results emerge from this simple analysis. If the goal is to maximize the mean change on treatment, a composite is not necessary, and only the gene with the highest predicted mean change is needed. Although all mean–variance efficient combinations of the two genes lie on the critical line, there is no single optimal set of weights: the individual has to select a weight based on his/her personal tolerance for variability. The individual may choose the point on the critical line associated with the lowest variability (called the minimum variance point) but may have to compromise on magnitude of the mean change in the composite. Importantly, because the correlation coefficient contributes to variability of the composite, a composite comprised of genes with negative or low correlation coefficients will have less variability than a composite comprised of genes strongly correlated with one another, all other factors remaining the same.

Extension to Multiple Genes

If the fractional change in expression of the ith gene is E_i, we define the composite measure function M containing the N informative genes as:

$$M = \sum_{i=1}^{N} w_i E_i \quad (3)$$

The optimization of the composite measure requires selecting the weights w_i for each gene so that M is maximal. The w_i are constrained to sum to unity. This linear constraint can be expressed as:

$$\sum_{i=1}^{N} w_i = 1 \quad (4)$$

However, the variability associated with expression profiling measurements requires the addition of a layer of complexity to this simple formulation. Because of variability, the E_i are random variables, and we will ideally wish to maximize the $\bar{M}$, the mean value of the composite measure function subject to linear constraints:

$$\underset{W_1,\ldots,W_N}{Max}\left\{\overline{M} = \sum_{i=1}^{N} w_i\overline{E}_i\right\} \quad (5)$$

The $\overline{E}_i$ are mean fractional change in expression values, ideally population means, and the maximization is accomplished by varying weights. Likewise, the variance, $\sigma^2{}_M$, of the composite measure M is:

$$\sigma_M^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j V_{ij} \quad (6)$$

The V_{ij} is the covariance matrix and is related to the covariance, σ_{ij}, the standard deviations, σ_{ii} and σ_{jj}, for genes i and j, respectively, and the correlation matrix ρ_{ij} by:

$$V_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \sigma_{ii}\sigma_{jj}\rho_{ij} \quad (7)$$

Although the underlying analysis of the multigene case is very similar to that of the two-gene case, the computational com-

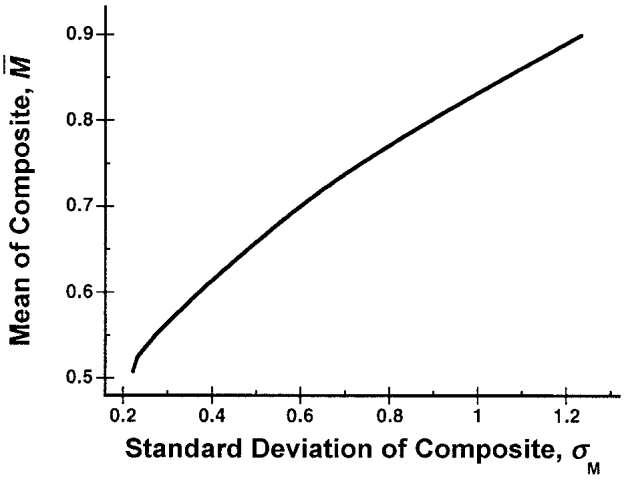

Fig 2. Markowitz efficient frontier for the five-gene case. The calculations used experimental values from 14 patients for the five genes in the neighborhood analysis predictor with the highest E_i values. The parameters in Table I were used.

plexity increases, and an optimization technique called quadratic programming is needed to determine the weights w_i. The quadratic programming algorithm reformulates the efficient frontier problem to minimizing the following objective function for each value of a parameter λ, subject to linear constraints:

$$\underset{w_1,\ldots,w_N}{Min} \{\sigma_M^2 - \lambda \overline{M}\} \quad (8)$$

For illustrative purposes, we used the five genes from the neighborhood analysis predictor with the highest E_i values. Table I summarizes the E_i values and the covariance matrix used as input to the Markowitz portfolio algorithm, and Fig. 2 shows the efficient frontier.

The weights for the five genes, genes 1–5, at the point of minimum variance were: 0.00, 0.222, 0.272, 0.443, 0.063, respectively. Interestingly, for this set of inputs, gene 1 was not needed for the minimum variance composite. The mean value, $\overline{M}$, and standard deviation, σ_M, at the minimum variance point were 0.508 and 0.222, respectively.

Comparison of the Markowitz Portfolio Model to Neighborhood Analysis

The weights predicted by neighborhood analysis were compared to those from the Markowitz portfolio model for the 14 patients for whom expression profiles after IFN-β1a treatment were available.

Table I. Inputs to the Markowitz Portfolio Model for the Five Genes with the Highest E_i Values

		Covariance Matrix				
Gene	E_i Value	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Gene 1	0.8992	1.5229	0.0862	0.3777	0.0660	0.0746
Gene 2	0.5693	0.0862	0.2151	–0.0081	0.0048	0.0280
Gene 3	0.5349	0.3777	–0.0081	0.1899	–0.0040	0.0193
Gene 4	0.4699	0.0660	0.0048	–0.0040	0.1010	0.0729
Gene 5	0.4384	0.0746	0.0280	0.0193	0.0729	0.0882

Figure 3A shows the weights assigned to each gene by the two methods. With a few exceptions, the neighborhood analysis method assigns approximately equal weights to each gene. In contrast, the Markowitz portfolio model selectively weights many fewer genes than neighborhood analysis. Figure 3B shows the corresponding efficient frontier, and the single point summarizes results from neighborhood analysis. The mean value $\overline{M}$ and standard deviation σ_M from neighborhood analysis were 0.296 and 0.078, respectively. In contrast, the mean value $\overline{M}$ and standard deviation σ_M at the minimum-variance point from the Markowitz portfolio analysis were 0.271 and 0.0044, respectively. The Markowitz efficient frontier has higher mean at a standard deviation of 0.078 and lower standard deviation at a mean of 0.296.

The neighborhood analysis predictor, which contained 40 genes, correctly predicted 13 of 14 pretreatment samples and all 14 posttreatment samples for an overall accuracy of 96.4%. The minimum-variance Markowitz portfolio, which contained only 12 genes, correctly predicted all 14 pretreatment samples and 13/14 posttreatment samples. The single gene with the

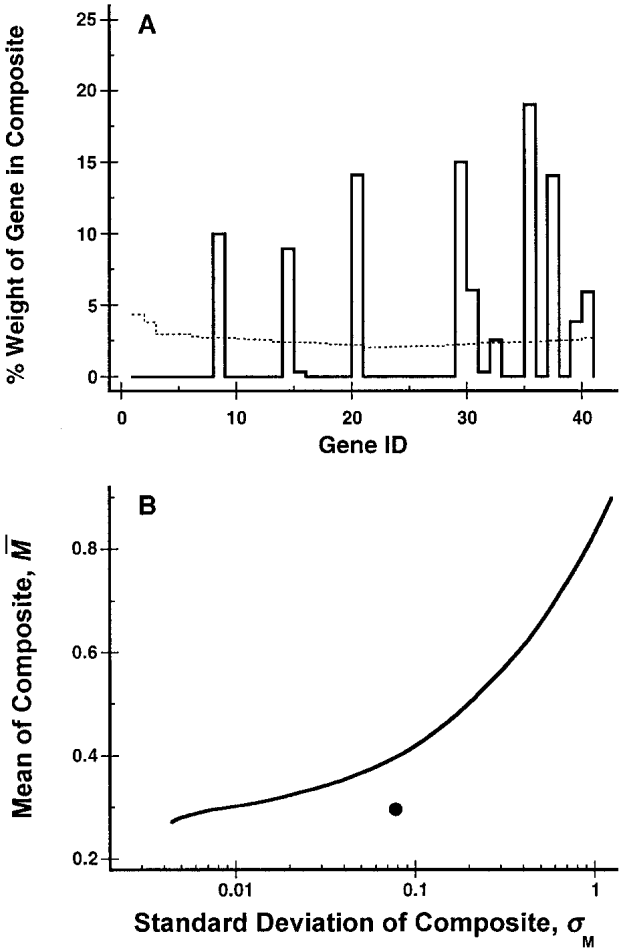

Fig 3. Comparisons of the results obtained from the Markowitz portfolio model to those from neighborhood analysis. (A) A step graph that shows the weights assigned to individual genes by the Markowitz model (solid line) and the weights assigned by neighborhood analysis. The weights are shown for the node with the minimum variance. (B) The Markowitz efficient frontier is the solid line, and the single point represents the results from neighborhood analysis. The *x*-axis is on logarithmic scale.

highest mean correctly predicted only 18 of 28 samples. In the quadratic programming solution for the Markowitz model, genes enter and leave at so-called "corner" portfolios, and all points in the efficient frontier can be represented as a linear sum of two corner portfolios. The prediction capabilities of over 20 Markowitz model corner portfolios nearest the minimum-variance portfolio were also examined. The four corner portfolios nearest the minimum-variance point all predicted 27 of 28 samples correctly; the next nearest set of eight corner portfolios predicted 26 of 28 samples correctly. Thus, the Markowitz model results in composites that are comparable to neighborhood analysis in predictive capabilities.

DISCUSSION

Here, the usefulness of a model for constructing composite biomarkers for treatments was demonstrated. The model is adapted from the Markowitz portfolio model of mathematical finance and provides useful insights into the requirements for efficient composites and the conditions under which they will outperform single endpoints in clinical studies.

The model implicitly assumes log-normal distributions for ratio of post- and pretreatment expression levels. This requirement is not particularly limiting because the log-normal distribution describes a wide range of therapeutic responses. However, the logarithm of the ratio of post- to pretreatment expression levels for the 40 genes in the neighborhood analysis composite were calculated and tested for normality with the Kolmogorov–Smirnov test. The statistical testing did not reject any of the null hypotheses, indicating that log-normal requirements are likely to be met for gene expression data from arrays.

The important methodologic distinction between the model proposed and the neighborhood analysis is the role of the gene–gene covariances. Unlike the neighborhood analysis method, which only includes associations between the individual genes and the treatment effects, the adaptation of the Markowitz model includes gene–gene correlations and is able to generate more efficient results with fewer genes. Although sample estimates for the covariance and other inputs were used, appropriate Bayesian priors or Stein estimators for these inputs can also be used when appropriate. The quadratic programming procedure involves numerical optimization, and it is important to have the best available estimates for the inputs.

Although the model was applied to gene array data, the principles involved are quite general and can readily be adapted for composite endpoints from other measurements such as those from proteomics and from physiologic measurements from multichannel recorders.

ACKNOWLEDGMENTS

This work was supported in part by a grant RG3258A2 from the National Multiple Sclerosis Society. I thank Linh Nguyen for conducting the array experiments and Li Zhang for assistance with the software.

REFERENCES

1. M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7**:819–837 (2000).
2. M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* **98**:8961–8965 (2001).
3. M. K. Kerr and G. A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **77**:123–128 (2001).
4. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**:531–537 (1999).
5. P. Chomczynski and K. Mackey. Short technical reports. Modification of the TRI reagent procedure for isolation of RNA from polysaccharide- and proteoglycan-rich sources. *Biotechniques* **19**: 942–945 (1995).
6. W. P. Dunlop, J. M. Cortina, J. B. Vaslow, and M. J. Burke. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol. Methods* **1**:170–177 (1996).
7. H. M. Markowitz. *Portfolio Selection; Efficient Diversification of Investments.* John Wiley & Sons, New York, 1959.
8. P. Wilmott. *Derivatives: The Theory and Practice of Financial Engineering.* John Wiley & Sons, New York, 1998.