BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 16 2009, pages 2013–2019
doi:10.1093/bioinformatics/btp357

Gene expression

PROMISE: a tool to identify genomic features with a specific biologically interesting pattern of associations with multiple endpoint variables

Stan Pounds[1,*], Cheng Cheng[1], Xueyuan Cao[1], Kristine R. Crews[2], William Plunkett[3], Varsha Gandhi[3], Jeffrey Rubnitz[4], Raul C. Ribeiro[4], James R. Downing[5] and Jatinder Lamba[6]

[1]Department of Biostatistics, [2]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, [3]Department of Experimental Therapeutics, M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, [4]Department of Oncology, [5]Department of Pathology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105 and [6]Department of Experimental and Clinical Pharmacology, University of Minnesota, 308 Harvard St. S.E., Minneapolis, MN 55455, USA

Received on January 30, 2009; revised on June 1, 2009; accepted on June 4, 2009

Advance Access publication June 15, 2009

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: In some applications, prior biological knowledge can be used to define a specific pattern of association of multiple endpoint variables with a genomic variable that is biologically most interesting. However, to our knowledge, there is no statistical procedure designed to detect specific patterns of association with multiple endpoint variables.

Results: Projection onto the most interesting statistical evidence (PROMISE) is proposed as a general procedure to identify genomic variables that exhibit a specific biologically interesting pattern of association with multiple endpoint variables. Biological knowledge of the endpoint variables is used to define a vector that represents the biologically most interesting values for statistics that characterize the associations of the endpoint variables with a genomic variable. A test statistic is defined as the dot-product of the vector of the observed association statistics and the vector of the most interesting values of the association statistics. By definition, this test statistic is proportional to the length of the projection of the observed vector of correlations onto the vector of most interesting associations. Statistical significance is determined via permutation. In simulation studies and an example application, PROMISE shows greater statistical power to identify genes with the interesting pattern of associations than classical multivariate procedures, individual endpoint analyses or listing genes that have the pattern of interest and are significant in more than one individual endpoint analysis.

Availability: Documented R routines are freely available from www.stjuderesearch.org/depts/biostats and will soon be available as a Bioconductor package from www.bioconductor.org.

Contact: stanley.pounds@stjude.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

1 INTRODUCTION

Microarrays have opened exciting new possibilities for biological research by enabling investigators to simultaneously measure thousands or even millions of genomic features in a biological specimen. Statistical analysis is used to identify features that are associated with an endpoint of interest. The statistical analysis often includes one or more hypothesis tests for each genomic feature to explore the association with the endpoint of interest. Gene-set enrichment analysis (GSEA; Jiang and Gentleman, 2007) can be a useful complement to the feature-by-feature analysis. Given a collection of predefined sets of genes that share a common biological function or jointly participate in a specific biological process, GSEA performs a statistical test for each gene set to determine whether the member genes are 'enriched' among the most statistically significant results. The feature-by-feature and GSEA each lead to a multiple-testing problem, so that several false discoveries may result even if very stringent *P*-value thresholds are applied. Pounds (2006) reviewed several methods that address this multiple-testing problem by estimating or controlling the false discovery rate (FDR; Benjamini and Hochberg, 1995; Storey and Tibshirani 2003). Nevertheless, after the statistical analyses are complete, the biologist is still left with the problem of using the results to select the most promising candidates for follow-up research.

One strategy to select the most promising leads from a list of genes that show a statistically significant association with one endpoint is to identify genes that show a biologically plausible pattern of association with related endpoints. For example, Yang *et al*. (2009) explored the association of the genotype of 600K single nucleotide polymorphisms (SNPs) with the level of minimal residual disease of acute lymphoblastic leukemia after initial treatment with chemotherapy. To refine the list of SNPs identified in the initial genome-wide analysis, they explored the association of the genotypes of the significant SNPs with pharmacokinetics (PK) endpoints that are associated with response to therapy. The SNPs with the most biologically plausible pattern of association with the

clinical and pharmacologic endpoints are now considered promising candidates for future pharmacogenetic research in this field. In this study, prior biological knowledge regarding the relationships of endpoints to one another helps to define patterns of association for genomic variables that are biologically most plausible.

However, to our knowledge, there is not a published statistical procedure that is designed to identify genomic features that show a specific pattern of association with multiple endpoint variables. In general, it may be difficult to incorporate prior biological knowledge regarding the relationships among multiple endpoint variables when trying to interpret the results of several analyses that each explores the association of genomic variables with an individual endpoint. For example, how does one characterize the statistical significance of a gene that has $P=0.0001$ for association with one endpoint, but $P=0.04$ for association with another endpoint in the context of multiple testing? If each individual endpoint analysis is corrected for multiple testing, then there may not be *any* gene that meets the stringent criteria for statistical significance in more than one individual endpoint analysis. Thus, to identify genes with interesting patterns of association, the *P*-value threshold must be made less stringent, opening the possibility for a large number of false discoveries.

Here, we propose projection onto the most interesting statistical evidence (PROMISE) as a statistical procedure to identify genomic features that show a specific pattern of association with multiple endpoints that is biologically most plausible or biologically most interesting. PROMISE performs one hypothesis test for the specified pattern for each genomic variable, thus avoiding the inferential problems described above. Additionally, PROMISE is a flexible procedure that can accommodate various types of endpoints, which classical multivariate procedures are not designed to manage. In Section 2, we describe the PROMISE procedure. Section 3 describes how PROMISE differs from other procedures. Section 4 presents simulation studies, and Section 5 presents the results from an example application. Finally, Section 6 provides the discussion and concluding remarks.

2 THE PROMISE METHOD

Suppose that $g=1,\ldots,m$ genomic features are measured for $i=1,\ldots,n$ subjects. Also, suppose that data on $j=1,\ldots,k$ endpoint variables are available for these subjects. For $i=1,\ldots,n$ and $g=1,\ldots,m$, let y_{ig} represent the value of genomic feature g for subject i. Additionally, let x_{ij} represent the value of endpoint variable j for subject i. For $j=1,\ldots,k$, let $\boldsymbol{x}_j$ represent the vector $(x_{1j},x_{2j},\ldots,x_{nj})$ of values of endpoint variable j for all subjects. Similarly, for $g=1,\ldots,m$, let $\boldsymbol{y}_g$ represent the vector $(y_{1g},y_{2g},\ldots,y_{ng})$ of values for genomic variable g for all subjects. Let $\boldsymbol{Y}$ represent the set of all y_{ig} and let $\boldsymbol{X}$ represent the set of all x_{ij}.

PROMISE is a general procedure to identify genomic variables with the strongest statistical evidence for the biologically most interesting pattern of associations with the endpoint variables. For $j=1,\ldots,k$, let $T_{jg}(\boldsymbol{x}_j,\boldsymbol{y}_g)$ be a statistic measuring the association of genomic feature y_g (for any g) with endpoint variable x_j. Now, define

$$\mathbf{T}_g=\left(T_{1g},T_{2g},\ldots,T_{kg}\right) \tag{1}$$

as the vector of the statistics measuring the association of genomic feature g with the endpoint variables $j=1,\ldots,k$.

In many applications, biological knowledge can be used to define a vector $\boldsymbol{d}=(d_1,\ldots,d_k)$ that corresponds to the biologically most interesting values for $\mathbf{T}_g$. For example, suppose that a set of subjects is treated with a drug that inhibits DNA synthesis. Drug levels and DNA synthesis rates after the treatment are measured as endpoint variables on these subjects. Genomic data are also collected on the same set of subjects. In this example, there are $k=2$ endpoints. Also, for each genomic variable g we have $\mathbf{T}_g=\left(T_{1g},T_{2g}\right)$ where T_{1g} and T_{2g} are the correlations of genomic variable g with drug level and DNA synthesis rate, respectively. Conceptually, a biologically most interesting result would be $\mathrm{T}_{1g}=+1$ and $\mathrm{T}_{2g}=-1$. Thus, the vector $\boldsymbol{d}=(+1,-1)$ defines a biologically most interesting statistical result for $\mathbf{T}$. Additionally, the result $-\boldsymbol{d}$ would be another biologically most interesting statistical result. Thus, $\boldsymbol{d}$ defines a vector in the space of $\mathbf{T}_g$ that corresponds to the biologically most interesting result.

The dot-product of $\boldsymbol{d}$ and $\mathbf{T}_g$,

$$R_g(\boldsymbol{X},\boldsymbol{y}_g)=\boldsymbol{d}\cdot\mathbf{T}_g=\sum_{j=1}^{k}d_j\mathrm{T}_{\mathrm{jg}} \tag{2}$$

is a statistic that measures the similarity of the vector $\mathbf{T}_g$ of observed associations to the vector $\boldsymbol{d}$ of the biologically most interesting statistical results. The sign of R_g indicates the direction of $\mathbf{T}_g$ relative to $\boldsymbol{d}$ and the magnitude of R_g is proportional to the length of the projection of $\mathbf{T}_g$ onto $\boldsymbol{d}$. Thus, (2) defines the PROMISE statistic.

The significance (i.e. *P*-value) of the PROMISE statistic can be determined via permutation. Let r_g represent the value of R_g computed from the observed data. Also, let $\boldsymbol{X}_l^{\star}$, represent an endpoint dataset obtained by $l=1,\ldots,b$ permutations of the subject indices of $\boldsymbol{X}$. Let $R_{gl}^{\star}=R(\boldsymbol{X}_l^{\star},\boldsymbol{y}_g)$ and let ρ_g^0 be a specified 'null value' of R_g. Then, for $g=1,\ldots,m$,

$$p_g=\frac{1}{b}\sum_{l=1}^{b}\mathrm{I}\left(\left|R_{gl}^{\star}-\rho_{\mathrm{g}}^0\right|\geq\left|r_g-\rho_{\mathrm{g}}^0\right|\right), \tag{3}$$

where $\mathrm{I}(\cdot)$ is the indicator function that equals 1 if the enclosed statement is true and equals 0 otherwise. Note that (3) defines a two-sided permutation *P*-value giving the probability that $\left|R_g-\rho_g^0\right|\geq\left|r_g-\rho_g^0\right|$ under the complete (permutation) null.

Here, the null hypothesis is exchangeability of the assignments of genomic data to endpoint data. Also, we note that the endpoint data are permuted jointly; each endpoint variable is not permuted individually. Permuting endpoint variables individually breaks the correlation of the endpoints with one another and thus tests a different null hypothesis than permuting the endpoint data jointly. As a consequence, permuting endpoints individually is likely to yield results that are very statistically significant, but not biologically meaningful because the small *P*-values may indicate that the endpoints are strongly correlated with one another instead of indicating that the genomic variable has the interesting pattern of association with the endpoint variables.

3 OTHER APPROACHES

There are other approaches that could be taken to identify genomic variables that exhibit a specific pattern of association with multiple endpoints. A seemingly straightforward approach would be to screen the association of the genomic variables with each endpoint

individually and then identify genes that are significant in each analysis and have the desired pattern of association. This approach is problematic because it lacks statistical power, and the results are difficult to interpret statistically. The analysis for each endpoint involves multiple testing. After adjusting each endpoint's results for multiple testing, it is quite likely that no gene will meet the criteria for inclusion as a 'significant' result because it is unlikely that a genomic variable could meet the stringent P-value threshold for each endpoint. Additionally, with this approach, if any genes are identified, it would be difficult to assign a meaningful FDR estimate to the result. For example, what FDR estimate should be ascribed to a set of genes that are inferred to have the association pattern of interest if these genes are selected because they meet a certain FDR or P-value threshold in several single-endpoint analyses? PROMISE avoids these problems by performing a single test for the pattern of association for each gene. For each genomic variable, PROMISE performs one test that directly addresses the question of whether a gene shows the association pattern of interest. This improves the statistical power and simplifies the interpretation, as seen in the example application of Section 5.

Classical multivariate methods, such as principal components (PC) or canonical correlation (CC), are other potential approaches to the problem. For example, one could determine the first PC (PC1) for the endpoint data and then test for the association of each genomic variable with the PC. However, in general, there is no reason to expect that the association of the genomic variable with the PC is a measure of evidence for a specific pattern of association with the endpoints. Clearly, a PC differs markedly from the definition of the PROMISE statistic in (2). The first PC is the linear combination of the endpoint variables that explains the greatest variation in the set of endpoint variables. Unlike PCs, PROMISE does not define a new endpoint variable by a linear combination of the individual endpoint variables. Instead, the PROMISE statistic in (2) is a linear combination of the *statistics characterizing the associations* of the individual endpoints with the genomic variable.

Also, one could compute the CC of each gene with the endpoint variables and test whether the CC is non-zero. However, CC also differs from the definition of the PROMISE statistic in (2). The CC of two sets of variables is the maximum possible correlation of a linear combination of the first set of variables with a linear combination of the second set of variables. The index selection method of quantitative genetics (Falconer and Mackay, 1996, pp. 240–245) is similar to CC. These approaches do not measure a specific pattern of association of one variable with a set of other variables as in (2).

Furthermore, in many applications, classical multivariate methods such as PCs and CC are not well suited to handle the endpoints of interest. For example, in clinical studies, the endpoint of greatest interest may be a censored time-to-event variable such as relapse-free survival. A censored time-to-event variable is one that is only partially observed in some subjects. For a subject of a clinical trial, relapse-free survival is the duration of time between study enrollment and death or relapse. Many subjects are living and remain relapse-free at the conclusion of the study. For these subjects, it is known that the relapse-free period is greater than the length of time they were followed, but the full duration of the relapse-free period is unknown. The relapse-free period for these subjects is considered *censored*. This type of endpoint does not fit into the classical multivariate normal framework. As seen in the example application of Section 5, the definition of the PROMISE statistic in (2) is flexible enough to accommodate censored time-to-event variables and ordered categorical variables.

GSEA and other gene-set analysis approaches [such as that of Nettleton *et al.* (2008)] could be used to determine whether gene sets identified from the analysis of association with one endpoint are associated with another endpoint. For example, one could identify genes that are associated with one endpoint and then explore whether the set of identified genes is associated with another endpoint. While this exercise may provide useful biological insights, it does not give results with the same interpretation as PROMISE. PROMISE provides a P-value for each gene, whereas gene-set methods give a P-value for each gene set. Additionally, the interpretation of gene-set results may be difficult. For instance, what if the list of genes associated with endpoint A are associated with endpoint B, but the list of genes associated with endpoint B are not associated with endpoint A? Such questions could easily become quite frustrating when more than two endpoints are involved. Nevertheless, permutation-based gene-set analyses can be performed *in conjunction with* PROMISE to identify gene sets that are enriched among genes that show the association pattern of interest. Integrating gene-set methods with PROMISE may prove to be a synergistic combination in terms of improving statistical power to reveal important biological insights.

4 SIMULATIONS

Simulations were performed to compare the statistical power of PROMISE to that of other approaches in a collection of simple settings. Let (Y, X_1, X_2) be a random vector corresponding to a genomic variable and two endpoint variables. Suppose that (Y, X_1, X_2) follows a multivariate normal distribution with mean $(0,0,0)$ and that Y, X_1 and X_2 each have unit variance. Let $\rho_x = \text{cor}(X_1, X_2)$, $\rho_1 = \text{cor}(Y, X_1)$ and $\rho_2 = \text{cor}(Y, X_2)$. Suppose that the interesting pattern of association is for Y to show the same direction of association with X_1 and X_2. In terms of the framework above, the test is performed using $\boldsymbol{d} = (+1, +1)$ in Equation (2). For the PROMISE procedure, the statistic R is defined as the average of Pearson's correlation of Y with X_1 and Pearson's correlation of Y with X_2.

Several alternative approaches were considered in the simulation study. The classical t-test for Pearson's correlation coefficient was used to test the association of Y with X_1 and the association of Y with X_2. Also, Pearson's correlation of Y with the PC1 of (X_1, X_2) was computed. Additionally, the CC of Y with (X_1, X_2) was computed. In each simulation, two-sided P-values for the PROMISE statistic, the correlation of Y with PC1 and the CC statistic were determined using the same set of 1000 permutations of Y. An overlap (OV) analysis identified genes that were significant in each of the single-endpoint analyses. A total of 1000 independent replications were performed for each simulation setting defined by a unique set of values for the sample size n, ρ_1, ρ_2 and ρ_x. Simulations were performed for all combinations of $\rho_1 = \{-0.4, -0.3, \ldots, 0.3, 0.4\}$, $\rho_2 = \{-0.4, -0.3, \ldots, 0.3, 0.4\}$, $\rho_x = \{-0.1, 0.0, 0.1, 0.2\}$ and $n = 10, 20, 50, 100$. The Supplementary Materials include tables and contour plots with the results of all simulations.

PROMISE had the greatest probability to give a P-value less than $\alpha = 0.01$ for values of (ρ_1, ρ_2) along the line $\rho_1 = \rho_2$ (Fig. 1A).

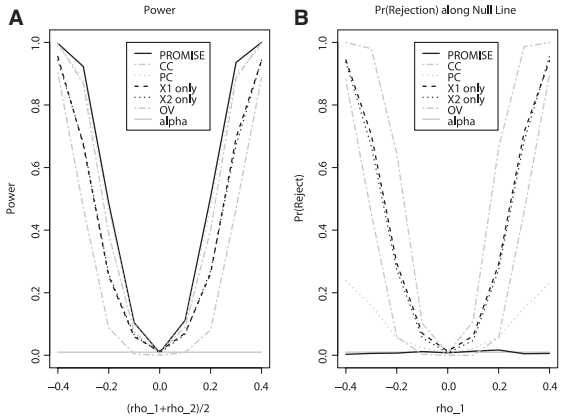

Fig. 1. Simulation results. Each plot gives simulation results for $\rho_x = 0$ and $n = 50$. (**A**) The power of PROMISE, PC, CC, OV and individual endpoint analyses (X_1 only and X_2 only) at the $\alpha = 0.01$ level along the line $\rho_1 = \rho_2$. Note that the line $\rho_1 = \rho_2$ is perpendicular to the null line $\rho_1 + \rho_2 = 0$ of the PROMISE procedure. (**B**) The probability that the procedures reject their respective null hypotheses along the null line $|\rho_1 + \rho_2| = 0$ of the PROMISE procedure. The Supplementary Materials provides the results of all simulations.

Recall that the interesting pattern of association is for Y to have the same direction of association with X_1 and X_2. The values of (ρ_1, ρ_2) that match this pattern of interest are the bottom-left and upper-right quadrants of the plots in the Supplementary Materials. The line $\rho_1 = \rho_2$ passes through the center of this region of interesting patterns of association. Simulations using other combinations of n and ρ_x give qualitatively similar results regarding the performance of PROMISE relative to that of other procedures (Supplementary Materials).

Additionally, PROMISE maintains its level along the line H_0: $\rho_1 + \rho_2 = 0$ (Fig. 1B). For values of ρ_1 and ρ_2 along the line $|\rho_1 + \rho_2| = 0$, other procedures test a different null hypothesis and therefore have a greater than $\alpha = 0.01$ probability of giving a P-value less than $\alpha = 0.01$ (Fig. 1B). For example, the X_1-only analysis tests $H_0 : \rho_1 = 0$, and the CC analysis tests the null hypothesis that the CC of Y with (X_1, X_2) is zero. However, the null hypothesis $H_0 : \rho_1 + \rho_2 = 0$ is the null hypothesis of interest, and thus the high probabilities of small P-values for values of (ρ_1, ρ_2) along this line are undesirable in this setting.

It is quite interesting that PROMISE outperforms classical multivariate procedures in this setting. The classical multivariate procedures were developed for this type of setting and are known to perform very well in this setting. The key observation here is that the classical multivariate procedures were designed to detect *any* non-zero correlations whereas PROMISE is designed to detect a *particular pre-specified pattern* of correlations. The simulations clearly show that PROMISE performs better for the latter purpose than other methods.

5 EXAMPLE APPLICATION

The example application uses data from the St. Jude AML97 clinical trial (Crews *et al*., 2002; Rubnitz *et al*. 2009). Ross *et al*. (2004) used Affymetrix U133A microarrays to measure gene expression in the leukemic cells of diagnostic bone marrow samples of 42 subjects in this trial. Additionally, PK, pharmacodynamic (PD) and clinical endpoint data were collected for these same subjects. Concentrations of ara-CTP (the active form of the chemotherapy drug cytarabine; ara-C) in leukemic cells of the bone marrow were obtained on the first day (CTP1; after ara-C alone) and the second day (CTP2; after combining cladribine with ara-C) of therapy. The rate of DNA synthesis in leukemic cells from the marrow was measured at diagnosis (baseline) and on days 1 and 2 of the therapy. From these measurements of DNA synthesis rates, we computed the log-ratio of the day 1 rate to the baseline rate (DNA1) and the log-ratio of the day 2 rate to the baseline rate (DNA2). The white blood count (WBC) in peripheral blood is a measure of tumor burden. The log-ratio of the WBC after 48 h of therapy to the WBC at the initiation of therapy was determined. Initial response (RESP) was determined by morphologic examination of a bone marrow aspirate collected after completion of the first course of chemotherapy and categorized as no response (RESP = 0), partial response (RESP = 1) or complete response (RESP = 2). Also, event-free survival (EFS) was defined as zero and considered uncensored for patients who did not achieve complete remission after two courses of therapy. For the remaining patients, EFS was defined as the time elapsed from study enrollment to relapse, development of a second malignancy or death, with patients having experienced none of those events censored at the date of last follow-up. For the purposes of PROMISE analysis in this application, the gene expression data were considered as genomic features and the other variables as endpoint variables. All 22 215 probe sets represented on the microarray were included in the analysis. We have previously noted that excluding probe sets on the basis of present–absent calls may be of limited value (Pounds and Cheng, 2005).

The correlation of gene expression with each endpoint can be measured using published statistical methods. The association of expression with CTP1, CTP2, DNA1, DNA2, WBC and RESP is measured with Spearman's correlation coefficient. For this application, these statistics are denoted T_{ctp1}, T_{ctp2}, T_{dna1}, T_{dna2}, T_{wbc} and T_{resp}. The association of expression with the risk of relapse and death (EFS) is measured using the rank-based statistic developed by Jung *et al.* (2005) that accounts for follow-up and censoring. For this application, this statistic is denoted T_{efs}. The statistic T_{efs} has the form of a dot product and was scaled so that $-1 \leq T_{\text{efs}} \leq 1$. For each endpoint, the association statistics were computed using all subjects with pairwise complete data (i.e. having gene expression data and data for the specific endpoint). This technique for managing missing data allows us to use all available data. The same approach was used for computing permutation statistics.

All AML97 subjects were randomly assigned to receive one of two infusion schedules for ara-C during the initial course of therapy. An amendment to the study protocol added one dose of intrathecally delivered ara-C before the first course of intravenous chemotherapy. Thus, each patient received one of four distinct therapies (short infusion or continuous infusion of ara-C with or without intrathecal ara-C). The statistical analyses of the association of gene expression with the PK endpoints (CTP1 and CTP2), PD endpoints (DNA1, DNA2 and WBC) and clinical endpoints (RESP and EFS) must account for the different therapeutic strategies. Therefore, for each endpoint and expression probe set, the correlation with expression was computed separately for each of four therapy-defined groups of subjects, and then the final correlation statistic was the sample-size weighted average of the group-specific correlations. This type

of adjustment for therapy is called a *stratified analysis*, and the therapy-defined groups are called *strata* (or one group is called a *stratum*).

In this application, prior biological and clinical knowledge was used to define the most interesting result for the association of gene expression with the seven endpoints. First, it is most interesting if the correlation of expression with CTP1 and CTP2 are both equal to ±1. For purposes of constructing the vector $\boldsymbol{d}$, let $d_{\text{ctp1}} = d_{\text{ctp2}} = 1$. Given this selection for d_{ctp1} and d_{ctp2}, the most interesting correlation of expression with DNA1 and DNA2 is $d_{\text{dna1}} = d_{\text{dna2}} = -1$, because the PD effect of ara-CTP is to interfere with DNA synthesis. Interference with DNA synthesis results in cell death and therefore leads to a reduction in WBC. Thus, $d_{\text{wbc}} = -1$. Increased levels of ara-C in leukemic blasts should reduce the amount of tumor in the marrow (a better tumor response), and therefore $d_{\text{resp}} = 1$. Effective therapy should reduce the risk of relapse and death, thus $d_{\text{efs}} = -1$. Therefore, by (2), the PROMISE statistic for this application was defined as

$$R_g(\boldsymbol{X}, y_g) = \frac{1}{7}\left(T_{\text{ctp1}} + T_{\text{ctp2}}\right) - \frac{1}{7}\left(T_{\text{dna1}} + T_{\text{dna2}} + T_{\text{wbc}}\right) + \frac{1}{7}\left(T_{\text{resp}} - T_{\text{efs}}\right). \quad (4)$$

The subscript g is omitted from the right-hand side of (4) for simplicity of notation. The PROMISE statistic is scaled by $1/7$ so that $-1 \le R_g \le 1$. A positive R_g indicates that the expression of probe set g shows a therapeutically beneficial pattern of correlation with the endpoint variables, i.e. higher expression associates with therapeutically desirable values of the endpoint variables. Similarly, negative R_g indicates that the expression of probe set g shows a therapeutically detrimental pattern of association with the endpoint variables.

The statistical significance of each individual endpoint's association statistic and the PROMISE statistic were determined using the same set of 10 000 permutations of the assignment of expression data to endpoint data. The permutations were restricted so that data reassignments were performed separately within each therapy-defined stratum because the differences in therapy are important factors for ara-C pharmacology and clinical outcome, as previously described (Rubnitz *et al.,* 2009). For each gene g, the P-value was computed by letting $\rho_g^0 = 0$ in Equation (3). Several interesting biological findings will be reported in detail elsewhere. Here, we describe a few results that illustrate the advantages of PROMISE.

The results for the human equilibrative nucleoside transporter 1 (hENT1) gene (probe set 201802_at) clearly illustrate the interpretative advantage of PROMISE. hENT1 is a solute carrier that brings ara-C into the cell (Hubeek *et al*., 2005). Given this role in ara-C metabolism, one would expect hENT1 to show a therapeutically beneficial pattern of association. This was indeed the case (Table 1 and Fig. 2). The expression of hENT1 was positively associated with CTP1, positively associated with CTP2, negatively associated with DNA1, negatively associated with DNA2, negatively associated with WBC, positively associated with RESP and negatively associated with the risk of an EFS event (Table 1). However, it is difficult to interpret the statistical significance of the association pattern with the seven individual P-values.

Table 1. The association statistic, P-value and rank among 22 215 probe sets (by P-value) for hENT1 from each individual endpoint analysis and the PROMISE analysis

Analysis	Corr	P-value	Rank
CTP1 (day 1 ara-CTP level)	+ 0.16	0.2579	6723
CTP2 (day 2 ara-CTP level)	+ 0.17	0.1890	4374
DNA1 (day 1 DNA synth.)	−0.30	0.0082	1084
DNA2 (day 2 DNA synth.)	−0.39	0.0214	726
WBC	− 0.09	0.5662	13 485
RESP	+ 0.40	0.0091	343
EFS	−0.07	0.2765	6874
PROMISE	+ 0.23	0.0033	225

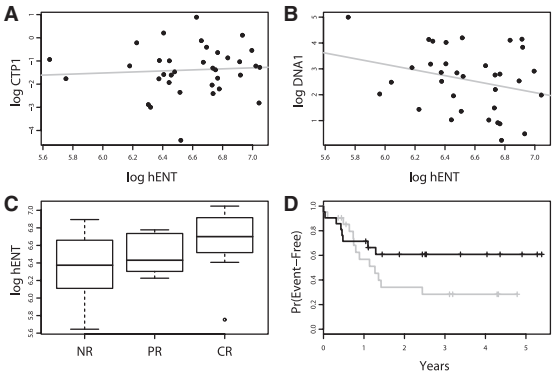

Fig. 2. Beneficial pattern of association for hENT1. (**A**) The log-CTP1 value versus the log-expression of hENT1. (**B**) The log-DNA1 value versus the log-expression of hENT1. (**C**) A boxplot of hENT log-expression for subjects with no response (NR), partial response (PR) and complete response (CR) to the initial course of chemotherapy. (**D**) The Kaplan–Meier estimates of EFS for subjects with hENT1 expression values greater than the median (black line) and those with hENT1 expression less than the median (gray line). Scatter plots illustrating the association of hENT1 expression with CTP2, DNA2 and WBC are not shown. Unlike the statistical analysis reported in Table 1, the above figures do not distinguish between subjects from different therapy groups.

The PROMISE analysis indicated that the beneficial pattern of association was very significant (R_g = 0.23, P = 0.0033, rank = 225). Thus, PROMISE identified this gene of known relevance to ara-C metabolism and provided a straightforward interpretation of statistical significance for the interesting pattern of associations with the endpoint variables. The individual endpoint analyses also provided insights that may be helpful for biological interpretation of the results. The associations with DNA1, DNA2 and RESP were strong contributors to the final value of the PROMISE statistic (correlations from 0.30 to 0.40). The associations with CTP1 and CTP2 were moderate contributors, and the associations with WBC and EFS were relatively weak contributors. Other genomic features had similar PROMISE statistics as hENT1, but for some of those features the associations with individual endpoints were substantially different.

In our application, PROMISE clearly had greater power to identify genes with interesting patterns of association. First, PROMISE identified a substantially greater proportion of all genes

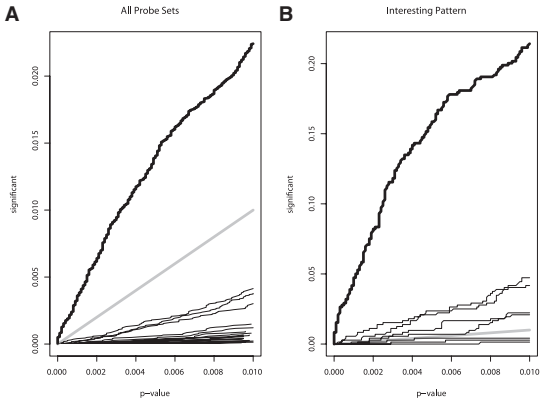

Fig. 3. Proportion of probe sets called significant as a function of P-value threshold. In (**A** and **B**), the results for PROMISE are shown by the thick black curve and the reference line $y=x$ is shown by the thick gray line. In (A), the results for pairwise overlap analyses are shown by the thin black lines. In (B), the results for individual endpoint analyses are shown by the thin black lines.

as significant than did any pairwise overlap analysis (Fig. 3A). In the PROMISE analysis, 498 probe sets were significant at $P=0.01$ level. In contrast, only 92 probe sets were significant at $P=0.01$ for DNA1 and DNA2, the greatest number of overlapping probe sets for any pair of individual endpoint analyses performed at the $P=0.01$ level. By definition, overlap of three or more endpoints detected fewer probe sets at the $P=0.01$ level. No probe set was significant in all seven individual endpoint analyses at a $P<0.15$. Second, PROMISE identified a greater proportion of genes with an interesting pattern of association than did any individual endpoint analysis (Fig. 3B). For 719 probe sets, the statistics measuring the association of expression with the individual endpoints all had signs matching (or all had signs mismatching) those of the interesting result vector $\boldsymbol{d}$. PROMISE identified 154 of these probe sets as significant at $P=0.01$. The individual endpoint analysis for DNA2 detected 34 of the probe sets with an interesting pattern of association at the $P=0.01$ level, the best among any individual endpoint analysis.

A permutation-based GSEA was also performed for each individual endpoint analysis and the PROMISE analysis. The pathway column of the Affymetrix annotation dataset was used to define 233 gene sets. For each gene set, the gene-set enrichment statistic was defined as the average of the absolute value of the member genes' correlation statistics. PROMISE identified 42 of these gene sets as significant at the $P=0.05$ level. Biologically interesting gene sets with significant PROMISE results included the DNA replication reactome gene set ($P=0.0248$) and the gene set for the G1-to-S phase of the cell cycle ($P=0.0113$). These gene sets are very interesting in this application because ara-C interferes with DNA synthesis (or replication), and clearly the cell cycle is very important in cancer. These two gene sets were significant at the $P=0.05$ level in the individual endpoint analysis for DNA1, but not in any other individual endpoint analysis. No gene set was significant in all seven individual endpoint analyses at $P<0.15$. Thus, GSEA based on PROMISE identified more gene sets than did searching for overlap among the results of gene-set analyses for individual endpoints.

6 DISCUSSION

PROMISE is a general procedure designed specifically to increase statistical power to identify genomic features that show a biologically most interesting pattern of association with multiple endpoint variables. PROMISE defines a test statistic that measures the evidence for the association pattern of interest by projecting the observed vector of association statistics onto the vector of conceptually most interesting values for those statistics. Permutation is used to compute P-values. Unlike classical multivariate statistical methods such as PC or CC, which are designed for data with a multivariate normal distribution, PROMISE can manage ordinal and censored time-to-event endpoints (Fig. 2). Furthermore, as observed in the simulation study in Section 4, CC and PC are not designed to detect a specific pattern of association and therefore do not have as much statistical power to detect the association pattern of interest as does PROMISE. PROMISE showed better power to identify genes with an interesting pattern of association in our example application than searching for such genes within lists of significant genes identified by individual endpoint analyses. Finally, GSEA can be incorporated into PROMISE so that the advantages of both approaches may be simultaneously realized. In our example, the PROMISE-based GSEA gave biologically interesting results and showed much greater statistical power than identifying overlap among the results of the individual-endpoint GSEAs.

Certainly, PROMISE is a very general procedure that must be customized to specific applications. The general concept presented here can be easily extended to accomodate stratified analyses by incorporating a stratum variable into the statistics $T_j(\cdot)$ and $R_g(\cdot)$ and restricting the permutations appropriately. As with other permutation-based methods, defining too many strata for stratified analysis may severely limit the statistical power of the analysis by reducing the number of available permutations. Additionally, the statistics $T_j(\cdot)$ and the most interesting results d_j must be defined in an application-appropriate manner. Also, PROMISE can be adapted so that SNP genotypes can be used as the genomic variables.

Future research should explore how to modify or generalize the correlation statistics and the way they are combined to form the PROMISE statistic. In this work, we used a geometric interpretation of the correlation vector to motivate the dot product as an objective way to uniquely define the PROMISE statistic given sufficient prior biological knowledge about the endpoint variables. Other ways to define the PROMISE statistic may prove useful in practice as well. Another approach to define the PROMISE statistic would be to define a vector $\boldsymbol{d}$ that subjectively weighs the correlations according to their practical relevance. For instance, in our example application, one may wish to give more weight to EFS due to its obvious importance for the patients. Additionally, it would be interesting to develop methods to define interesting result vectors and test statistics for applications with thousands of endpoint variables and thousands of genomic variables.

However, great caution should be exercised when using subjectively defined $\boldsymbol{d}$ because the statistical significance may be exaggerated if $\boldsymbol{d}$ is not selected a priori or if PROMISE is used as an exploratory procedure to perform many analyses with different $\boldsymbol{d}$ vectors. Users should definitely avoid using the observed correlations for specific genes to define $\boldsymbol{d}$. If $\boldsymbol{d}$ is defined to maximize the dot-product in (2) for a specific gene, the procedure will give a very small P-value for that gene. In this case, the small P-value

will greatly exaggerate statistical significance. In such a situation, the small P-value reflects the fact that the coefficients $\boldsymbol{d}$ were selected to maximize the statistic R_g instead of indicating that the observed correlations did not arise by chance. If a search is performed for $\boldsymbol{d}$, the P-value should adjust for that search in some manner. Nesting the search within each permutation round would be one way to perform such an adjustment and give meaningful P-values.

We do not recommend that the PROMISE procedure be used for applications in which there is not adequate prior biological knowledge about the endpoint variables to objectively define the vector $\boldsymbol{d}$. Our simulations suggest that canonical correlation is a good method for such a setting if the variables are appropriate for CC analysis.

ACKNOWLEDGEMENTS

We also thank Mr David Galloway for expert editorial assistance.

Funding: American Lebanese Syrian Associated Charities (ALSAC); National Institutes of Health (R01CA132946-01).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Crews,K.R. *et al.* (2002) Interim comparison of a continuous infusion versus a short daily infusion of cytarabine given in combination with cladribine for pediatric acute myeloid leukemia. *J. Clin. Oncol.*, **20**, 4217–4224.

Falconer,D.S. and Mackay,T.F.C. (1996) *Introduction to Quantitative Genetics*, 4th edn. Addison Wesley Longman, Essex.

Hubeek,I. *et al.* (2005) The human equilibrative nucleoside transporter 1 mediates in vitro cytarabine sensitivity in childhood acute myeloid leukaemia. *Br. J. Cancer*, **93**: 1388–1394

Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

Jung,S.H. *et al.* (2005) A multiple testing procedure to associate gene expression levels with survival. *Stat. Med.*, **24**, 3077–3088.

Nettleton,D. *et al.* (2008) Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, **24**, 192–201.

Pounds,S.B. (2006) Estimation and control of multiple testing error rates for the analysis of microarray data. *Brief. Bioinform.*, **7**, 25–36.

Pounds,S. and Cheng,C. (2005) Statistical development and evaluation of gene expression data filters. *J. Comput. Biol.*, **12**, 482–495.

Ross,M.B. *et al.* (2004) Gene expression profiling of pediatric acute myleogenous leukemia. *Blood*, **104**, 3679–3687.

Rubnitz,J.E. *et al.* (2009) Combination of cladribine and cytarabine is effective for childhood acute myeloid leukemia: results of the St. Jude AML97 trial. *Leukemia*, to appear.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Yang,J. *et al.* (2009) Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia. *J. Am. Med. Assoc.*, **301**, 393–403.