BIOINFORMATICS

Vol. 23 ECCB 2006, pages e91–e98
doi:10.1093/bioinformatics/btl298

A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data

Nir Yosef[1], Zohar Yakhini[3], Anya Tsalenko[3], Vessela Kristensen[4], Anne-Lise Børresen-Dale[4,5], Eytan Ruppin[1,2] and Roded Sharan[1,*]

[1]School of Computer Science and [2]School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel,
[3]Agilent Technologies and [4]Department of Genetics, Institute of Cancer Research,
Rikshospitalet-Radiumhospitalet Medical Center, Montebello, 0310 Oslo, Norway and
[5]Medical Faculty, University of Oslo, 0316 Oslo, Norway

ABSTRACT

Motivation: Large-scale association studies, investigating the genetic determinants of a phenotype of interest, are producing increasing amounts of genomic variation data on human cohorts. A fundamental challenge in these studies is the detection of genotypic patterns that discriminate individuals exhibiting the phenotype under study from individuals that do not posses it. The difficulty stems from the large number of single nucleotide polymorphism (SNP) combinations that have to be tested. The discrimination problem becomes even more involved when additional high-throughput data, such as gene expression data, are available for the same cohort.

Results: We have developed a graph theoretic approach for identifying discriminating patterns (DPs) for a given phenotype in a genotyped population. The method is based on representing the SNP data as a bipartite graph of individuals and their SNP states, and identifying fully connected subgraphs of this graph that relate individuals enriched for a given phenotypic group. The method can handle additional data types such as expression profiles of the genotyped population. It is reminiscent of biclustering approaches with the crucial difference that its search process is guided by the phenotype under consideration in a supervised manner. We tested our approach in simulations and on real data. In simulations, our method was able to retrieve planted patterns with high success rate. We then applied our approach to a dataset of 72 breast cancer patients with available gene expression profiles, genotyped over 695 SNPs. We detected several DPs that were highly significant with respect to various clinical phenotypes, and investigated the groups of patients and the groups of genes they defined. We found the patient groups to be highly enriched for other phenotypes and to display expression coherency among their profiles. The gene groups displayed functional coherency and involved genes with known role in cancer, providing additional support to their involvement.

Availability: The program is available upon request.

Contact: roded@post.tau.ac.il

1 INTRODUCTION

The dissection of complex diseases is one of the greatest challenges of human genetics with important clinical and scientific applications. High-throughput technologies yield large-scale datasets on genomic variation in diverse populations, allowing the study of these variations and their association with disease and other complex traits. A fundamental problem in interpreting this wealth of data is to pinpoint genotype patterns that discriminate phenotype classes from one another (Moore and Ritchie, 2004).

Traditionally, associations were sought between single genetic markers and disease (Thomas, 2004; Martin *et al*., 2001). Very few methods exist for identifying larger sets of single nucleotide polymorphisms (SNPs) that are associated with a phenotype of interest (Marchini *et al*., 2005). However, existing empirical evidence from model organisms (Segre *et al*., 2005) and human (Zerba *et al*., 2000) studies suggests that interactions among loci contribute broadly to complex traits.

Recently, several groups have conducted association studies that combine genotype data with expression profiling of the population under study (Sklan *et al*., 2004; Tanahashi *et al*., 2005). While much of our understanding of the genetic base of disease comes from identifying polymorphisms that affect protein structure or integrity, it is clear that RNA and protein abundance also drive disease processes (Takamizawa *et al*., 2004; Sorlie *et al*., 2001). In this setting, the discrimination challenge is even more involved and calls for pinpointing subsets of SNP and gene expression states (features) that allow the separation of a given phenotypic group from all others.

A possible approach to the problem is to apply biclustering algorithms to the data, searching for subsets of individuals that are coherent in their behavior across a subset of the features, as is routinely done for gene expression datasets (Tanay *et al*., 2002; Klugar *et al*., 2003). However, the obvious caveat of this approach is that the phenotype information is ignored in the search process, possibly yielding biclusters that exhibit no phenotypic coherence.

In this work, we present a graph theoretic approach to the discrimination problem, which directly searches for subsets of individuals that are both coherent in their phenotype information and in their feature data. Our approach is based on representing the feature data as a weighted bipartite graph in which vertices on one side represent individuals and their associations with the phenotype of interest, and vertices on the other side represent feature states. Edges in this graph connect individuals and the feature states they exhibit. The discriminating pattern (DP) problem can then be recast into that of finding high weight bicliques (fully connected subgraphs) in this graph. We first lay the theoretical foundations to this computational problem and prove its computational hardness; we then give a branch and bound like algorithm for the problem, which is shown to have good retrieval properties in simulations.

*To whom correspondence should be addressed.

Finally, we apply our approach to a combined SNP-expression data collected from 72 breast cancer patients (Kristensen *et al*., 2006). The data consist of 695 genotyped SNPs in selected genes from the reactive oxygen species (ROS) biochemical and signaling pathways, and expression levels of 3351 transcripts in tumor biopsies (Sorlie *et al*., 2001, 2003). We detect several DPs that are highly significant with respect to various clinical phenotypes, and investigate the groups of patients and the groups of genes they define.

The rest of this paper is organized as follows: Section 2 presents the DP problem and analyzes its complexity. Our algorithmic approach to the problem is described in Section 3. Finally, Section 4 presents our simulation experiments and the application of the method to real genotype and expression data.

2 PROBLEM DEFINITION

Let P denote the population under study and let (P^+, P^-) be a partition of the population w.r.t. a certain phenotype, where P^+ denotes the phenotypic class of interest, and $P^- = P \backslash P^+$. We call the individuals in P^+ (P^-) positives (negatives). Any biological attribute measured across the population, such as SNP states or gene expression levels, is called a feature. We assume that features can attain a set of discrete values, called states. The dataset can be represented as a feature matrix whose rows correspond to individuals and whose columns correspond to features. The entries of the matrix reflect the values of the biological attributes across the population. Given such a feature matrix, and a phenotypic group of interest, the goal is to identify a subset of features and an assignment of states to those features, such that individuals satisfying this assignment are enriched with the phenotype in question.

To formally define the DP problem, we first model the data using a bipartite graph $G = (P, F, E)$, where P is the set of individuals in the studied population and F is the set of all feature states. The edges in E connect each individual to the feature states he/she possesses. For a vertex $v \in P \cup F$ denote its set of neighbors in G by $N(v)$. For a subset $V \subseteq P \cup F$, let $N(V) = (\cup_{v \in V} N(v)) \backslash V$. For each individual $p \in P$, we assign a weight $w(p)$ to its corresponding vertex in G. $w(p)$ is set to 1 if the individual belongs to the phenotypic group under study, and -1 otherwise. We focus on bicliques of G, which represent subsets of individuals sharing certain feature patterns. We define the size of a biclique as the number of individuals it contains and the weight of a biclique as the sum of weights of the individuals it contains, i.e. the number of positives it contains minus the number of negatives it contains.

How does one measure pattern discrimination? An exact score is the hypergeometric p-value of the set of individuals satisfying the pattern, computed according to the phenotype in question. An approximate score that is much more efficient to compute is the weight of the set of individuals satisfying this pattern, i.e. the number of positives minus the number of negatives among them. Accepting the latter score, the discrimination problem is reduced to that of finding high scoring maximal bicliques in G. In fact, the crucial point is that each biclique is maximal w.r.t. individuals. This ensures that the pattern defined by the biclique is not satisfied by negative individuals outside the biclique. The problem is formally defined as follows:

DEFINTION 1. *(DP). Given a bipartite graph $G = (P, F, E, w)$ with $w{:}P \rightarrow \{-1,1\}$, find a feature subset $F' \subseteq F$ such that the biclique defined by F' and $P' = \{p \in P{:}N(p) \supseteq F'\}$ has maximum weight.*

In Appendix A we show that the decision version of DP is NP-hard. A branch and bound like approach to the problem is presented in the next section.

3 ALGORITHMIC APPROACH

3.1 Biclique identification

While there are previous studies of biclique identification in the literature (Tanay *et al*., 2002; Hochbaum, 1998; Bar-Yehuda *et al*., 2002), our instances differ in three important aspects: (1) vertex weights can be both positive and negative; (2) vertex degrees on both sides of the bipartite graph are not bounded and (3) solution bicliques are maximal with respect to the elements of P, i.e. no individual outside the biclique is fully connected to the features of the biclique. Thus, we cannot easily adapt approaches that search for maximum node bicliques (Hochbaum, 1998), or approaches that assume some degree bound on one of the sides of the bipartite graph (Tanay *et al*., 2002). Instead, to tackle DP we use a greedy search algorithm, which we describe next.

The input to the algorithm is a weighted bipartite graph $G = \{P, F, E, w\}$ with $w{:}P \rightarrow \{-1,1\}$. The algorithm returns a collection of feature subsets, each inducing a high scoring biclique. It consists of a series of recursive steps, gradually building the feature subsets to be returned. At each step the algorithm greedily chooses possible extensions to the current subsets and creates a collection of smaller instances of the problem to be passed on to subsequent recursion steps. The search starts at all possible single features and builds a search tree around each of them. The search tree is characterized by two parameters: κ, denoting the maximum degree of a node in the tree, and ϕ, denoting the maximum depth of the tree.

Starting with a single feature node, $v_{\text{root}} \in F$, the algorithm chooses the κ best candidate extensions $v_1 \ldots v_\kappa \in F \backslash \{v_{\text{root}}\}$, where each candidate v_i is scored based on the weight of the biclique it induces with v_{root}, i.e. $\text{score}(v) = \sum_{p \in N(v_{\text{root}}) \cap N(v)} w(p)$. For each pair $(v_i, v_{\text{root}}), 1 \leq i \leq \kappa$, the algorithm combines the two nodes and creates a new, smaller instance of the problem $G' = \{P', F', E'\}$ where $F' = F \backslash \{v_i\}, P' = N(v_{\text{root}}) \cap N(v_i)$ and $E' = E \cap (P' \times F')$. Each new instance is then passed on to the next recursive step. This process continues until the recursion reaches a depth of ϕ. Throughout its execution, the algorithm maintains a hash table where every feature subset visited during the search process is recorded in order to avoid repetitive computations. When the algorithm terminates, all subsets in the hash table whose corresponding biclique's score exceeds a certain threshold (Th) are output. The algorithm is summarized in Figure 1.

In our implementation ϕ was set to 5, and κ changed with the tree level: its value was set to 10 at the first level and 5 at all other levels. The running time of the algorithm for a typical input matrix with 50 individuals and 2000 features was <3 min, on average (using a single processor).

3.2 Significance assessment and post-processing

The computed feature subsets are subjected to several filtering steps that produce a set of significant, non-redundant patterns. Each subset induces a biclique whose set of individuals are enriched with the phenotype of interest. To evaluate the chance of observing

Algorithm 3.1: (c)

```
$BIC(G)$
    $H \leftarrow init\ hash\ table$
    for each $v \in F$ :
        $Call\ RecBic(\{v\}, G, 1, H)$
    for each $S \in H$ :
        if $|S| > 1$ and $HG\_Score(S) < Th$
            then $output\ S$

$RecBic(root, G, depth, H)$
    if $root \in H$ or $depth > \phi$
        then return
    $Add\ root\ to\ H$
    for each $v \in F \setminus root$ :
        $score(v) = \sum_{p \in N(root) \cap N(v)} w(p)$
    $\{v_1 \dots v_\kappa\} \leftarrow ChooseCandidates(score, \kappa)$
    for $i = 1$ to $\kappa$
        $F' \leftarrow F \setminus \{v_i\}$
        $P' \leftarrow N(root) \cap N(v_i)$
        $E' \leftarrow E \cap (P' \times F')$
        $G' \leftarrow \{P', F', E'\}$
        $Call\ RecBic(root \cup \{v_i\}, G', depth + 1, H)$
```

Fig. 1. Pseudo code of the biclique identification algorithm. HG_Score(·) computes the hyper geometric score of a biclique induced by a given feature subset. ChooseCandidates(·, κ) returns the best κ candidate extensions according to a given score.

such phenotype enrichment at random, we compare the returned bicliques to those obtained on randomized instances, in which the phenotype labeling is randomly permuted. For each random instance we record the highest scoring biclique of each size produced by our algorithm, and use these to rank and assign empirical p-values to the bicliques we identified in the original instance. The ranking is based on an exact hypergeometric score for each biclique rather than the less-accurate weight score. Formally, a biclique of size n that contains k positives is assigned the hypergeometric score HG($|P|$,$|P^+|$, n, k), where

$$\mathrm{HG}(N, B, n, b) = \sum_{m=b}^{\min\{n,B\}} \frac{\binom{B}{m}\binom{N-B}{n-m}}{\binom{N}{n}}.$$

Finally, we use a greedy process to filter the significant bicliques obtained, so that no bicliques share >50% of their edges (computed w.r.t. the smaller biclique). We also add to each output biclique additional features that are consistent with its patient set, whenever possible, thus obtaining maximal bicliques.

The process that we apply identifies the significant bicliques of each size. To estimate the false discovery rate (FDR) when combining all sizes together we use a procedure by Yekutieli and Benjamini (1999). For a given threshold p, on empirical p-values, the expected FDR is pK/R, where K is the number of sizes possible, and R is the total number of discoveries with empirical p-value less than p. We set p so that the expected FDR of the discovered bicliques is at most 10%.

The discrimination power of individual features within each feature subset reported by the algorithm may vary. Thus, for each subset F' we compute a core, which consists of a maximal subset of the features F'_C such that the enrichment of positives induced by F'_C is significant given the enrichments attained by all proper subsets of F'_C. Formally, let $n = |F'_C|$ and let $P_1, \dots, P_n$ denote the subsets of P induced by each of the ($n-1$)-size subsets of F'_C. Let P_C be the subset of individuals induced by F'_C. For a subset $Q \subseteq P$, let $Q^+(Q^-)$ denote the subset of positives (negatives) in Q. We are interested in the probability to randomly draw n subsets of P with the same sizes and scores as $P_1, \dots, P_n$ such that their intersection contains at least $|P_C^+|$ positives and at most $|P_C^-|$ negatives. To compute this score we use a method by S. Kaplan (personal communication). For simplicity, we describe it for $n = 2$, but the same technique generalizes to any n. For $n = 2$, this excess significance is $score\ (F'_C) = \frac{\gamma(|P_C^+|, \min(P_1^+, P_2^+), 0, |P_C^-|)}{\gamma(0, \min(P_1^+, P_2^+), 0, \min(P_1^-, P_2^-))}$, where

$$\gamma(\alpha_1, \alpha_2, \beta_1, \beta_2) = \sum_{\mathrm{pos}=\alpha_1}^{\alpha_2} \sum_{\mathrm{neg}=\beta_1}^{\beta_2} \binom{|P^+|}{\mathrm{pos}} \cdot \binom{|P^-|}{\mathrm{neg}} \cdot \binom{|P^+| - \mathrm{pos}}{|P_1^+| - \mathrm{pos}} \cdot \binom{|P^-| - \mathrm{neg}}{|P_1^-| - \mathrm{neg}} \cdot \binom{|P^+| - \mathrm{pos} - (|P_1^+| - \mathrm{pos})}{|P_2^+| - \mathrm{pos}} \cdot \binom{|P^-| - \mathrm{neg} - (|P_1^-| - \mathrm{neg})}{|P_2^-| - \mathrm{neg}}$$

4 EXPERIMENTAL RESULTS

To test our approach we first benchmarked it on simulated data which preserves some of the attributes of the real data. Next, we analyzed a genotype dataset of breast cancer patients. We began by applying a standard unsupervised biclustering algorithm to find bicliques without considering the phenotypic information. We then proceeded to show the added value provided by considering the phenotypic information available and applying our method. Finally, we analyzed a combined genotype-expression dataset.

4.1 Simulations

To study the performance of our algorithm, we performed a comprehensive set of simulation experiments. The simulated datasets were obtained by randomizing the real data that we subsequently analyzed. The randomization process was applied to the bipartite graph representing the real data, shuffling its edges while preserving vertex degrees. We randomly assigned half of the individuals in each simulation to be positives and the rest to be negatives.

The simulation tests were aimed at testing the ability of the algorithm to retrieve true significant bicliques. To this end, we planted within the randomized graphs bicliques of varying significance levels (hypergeometric scores). The algorithm was considered successful whenever it retrieved the planted biclique, or a biclique containing the planted one that attains a higher significance level. Figure 2 depicts the mean success rate in retrieving the planted bicliques as a function of their significance levels. The figure demonstrates that for high significance levels the planted biclique is retrieved with high success rate, while for lower significance levels ($p > e^{-6}$) the performance drops; the latter drop in performance is likely owing to a ‘masking’ effect of random bicliques that attain higher significance levels. Indeed, as can be

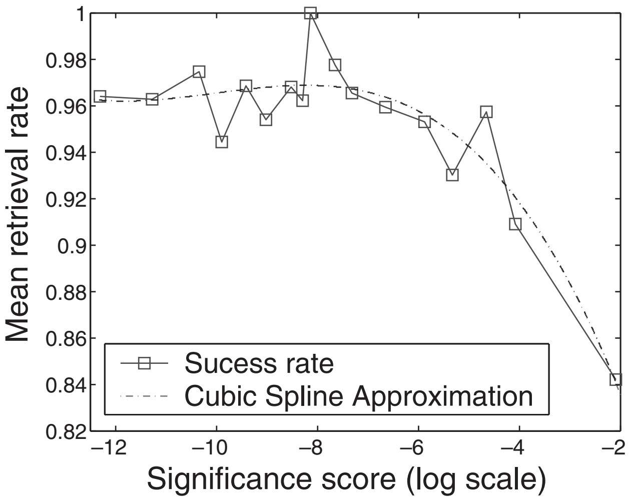

Fig. 2. Retrieval rate of planted bicliques is plotted as a function of their significance levels (hypergeometric score). The data are binned according to the significance scores of the planted bicliques, with ~100 data points in each bin. A polynomial spline of the success rate is provided to approximate the mean trend.

seen in Appendix B, the vast majority of random bicliques lie within this range of significance levels.

4.2 Application to breast cancer data

To demonstrate our approach in the settings of a large-scale association study, we applied it to a genotype-expression dataset collected from a group of 72 breast cancer patients (Kristensen *et al*., 2006). Blood samples of the patients were used to type 695 SNPs in selected genes from the ROS pathway. Gene expression profiles for 50 of these patients were collected from tumor tissues, and include the expression levels of 3351 genes (Sorlie *et al*., 2001, 2003). These expression data were normalized so that the mean expression level per individual is zero.

The patients were assigned with six clinical phenotypes, including response to treatment (ranging from 1 to 3, where 3 denotes a poorer response), grade of tumor (denoting its histological form, ranging from 1 to 3, with a higher grade being associated with poorer prognosis), the mutational status (either mutated or normal) of the TP53 gene (known to be associated with breast cancer and to affect the outcome of cytotoxic treatment), tumor classification according to Sorlie *et al*. (2001) (yielding five subtypes using a hierarchical clustering of expression data), lymph node status (ranging from 0 to 2, with a larger number denoting a more severe spread), and its T category (the staging level of the tumor, ranging from 1 to 4, denoting the tumor size, with category 4 denoting its spread to the chest wall). Each of these phenotypes induces a partition of the patients into several classes, or phenotypic groups. In the following we describe the application of the algorithm to each one of those classes.

4.3 Data modeling

We constructed a bipartite graph of individuals and features. Feature vertices correspond to SNP or expression states. Each SNP is represented by three feature vertices, corresponding to each of the states it can attain. Each gene is represented by two vertices reflecting up- or down-regulation of that gene compared with its mean expression level across all patients (determined as one standard deviation above or below the mean, respectively). Edges in this graph connect individuals to their corresponding features.

In a preprocessing step, we removed redundant features that shared all but at most five of their neighboring individuals in the bipartite graph. We also removed features that were connected to >60% of the individuals, to avoid unspecific features and reduce the computation time.

4.4 Treatment of missing data

Approximately 6.8% of the entries in the genotype-expression dataset were missing, mainly owing to poor signal to background ratios (Sorlie *et al*., 2001) or undetermined genotypes (Kristensen *et al*., 2006). We chose to treat missing data as if the corresponding edges did not exist (i.e. when the expression level of a certain gene in a certain sample is unknown, we assume that the gene was not significantly over- or under-expressed in that sample; similarly for missing genotype data). To ensure that the obtained bicliques are not influenced by biases in the missing data toward some of the phenotype classes, we test for such bias explicitly. Define the background set of a biclique as the set of of patients who do not miss any of the genotype or expression data of the biclique's features. We test whether this set is biased toward the phenotype class according to which the biclique was computed using a hypergeometric score. We discard bicliques whose hypergeometric *p*-value is <0.05.

4.5 Quality assessment

In order to evaluate the biological significance of the detected bicliques we examined whether they carry biological information in addition to the information related to the phenotype by which they were computed. We used three different criteria to evaluate the bicliques: enrichment of the patients in other phenotypic groups, coherency of the expression profiles of the patients, and functional enrichment of the genes participating in the bicliques. For all criteria, the computed score is compared with 100 scores for random sets of patients (in the first two cases) and of genes (in the third case) of the same size, and the resulting empirical *p*-value is estimated and reported.

Cross phenotype enrichment. Consider a biclique (P', F') constructed according to a phenotype T, and suppose we wish to test the enrichment of P' w.r.t. to a different phenotype Z. Since Z and T may be correlated, we cannot simply compute a hypergeometric score w.r.t. Z; instead we should take this possible correlation into account. To this end, we compute a hypergeometric-based p-value that conditions on the enrichment of the biclique w.r.t. T. Let (Z^+, Z^-) and (T^+, T^-) be the partitions induced by each of the two phenotypes, and let $k = |P' \cap T^+|$ and $m = |P' \cap Z^+|$. The cross phenotype enrichment score is computed as

$$\mathrm{Prob}(P', Z|T) = \sum_{m_1+m_2 \geq m} \frac{\binom{|T^+ \cap Z^+|}{m_1} \cdot \binom{|T^+ \cap Z^-|}{k - m_1} \cdot \binom{|T^- \cap Z^+|}{m_2} \cdot \binom{|T^- \cap Z^-|}{|P'| - k - m_2}}{\binom{T^+}{k} \cdot \binom{T^-}{|P'| - k}},$$

Table 1. Representative significant bicliques

Data	Phenotype	Features/patients	HG score	Expression coherency	Functional enrichment	Cross phenotype score/enriched phenotype
SNP	grade of tumor = 3	3/15	0.011	0.46	—	0.02/TP53 mut. Status = mutated
SNP	T category = 2 or 3	2/16	0.04	0.94	—	0.03/TP53 mut. Status = mutated
SNP	T category = 4	2/19	0.005	0.009	—	0.7/–
Combined	T category = 4	7/15	0.02	0.13	0.01	1/–
Combined	T category = 4	12/15	0.02	0.009	0.01	0.009/Array tumor classification = Luminal A

Columns indicate the dataset used, the phenotype according to which the biclique was computed, its size, its hypergeometric score w.r.t. its defining phenotype, the expression coherency of its patient set, the functional enrichment of its induced gene set (applicable only for bicliques containing expression state nodes), and the cross-phenotype enrichment score of the biclique together with the enriched phenotype.

where m_1 is restricted to the range $[\max\{0, k - |T^+ \cap Z^-|\}, |T^+ \cap Z^+|]$ and m_2 is restricted to the range $[\max\{0, |P'| - k - |T^- \cap Z^-|\}, |T^- \cap Z^+|]$. The cross phenotype score assigned to the biclique is taken as the minimum over its cross phenotype enrichments w.r.t. all phenotypes (other than T).

Gene expression coherency. We compare the pairwise correlations among expression profiles of patients in a biclique to those of random pairs of individuals using one-sided Wilcoxon rank-sum test.

Functional enrichment. For each biclique, we consider all feature nodes that represent expression states and are connected to >90% of the patients in the biclique. We test the functional enrichment of the corresponding gene set w.r.t. the gene ontology annotation using a hypergeometric score.

4.6 Performance of unsupervised biclustering

Our first task was to assess the potential advantages of our approach compared with extant unsupervised biclustering approaches. To this end, we analyzed the genotype data in an unsupervised manner using our in-house implementation of the SAMBA algorithm (Tanay *et al*., 2002), a state-of-the-art biclustering algorithm. SAMBA employs a likelihood ratio based scoring scheme to identify bicliques whose likelihood of occurrence in a random, degree preserving graph is small[1]. Since the seeding approach of SAMBA was infeasible in our setting (owing to the high degrees on each side of the bipartite graph), we used our recursive search procedure instead.

We evaluated the bicliques returned by SAMBA versus those returned on randomized, degree preserving instances. For each size separately, empirical p-values for bicliques of that size were evaluated using the randomized instances (comparing the score of the biclique to the maximum scores obtained for each of the randomized instances). Overall, 117 bicliques had significantly high scores. To test whether these bicliques were enriched for some phenotype we proceeded as follows: For each biclique we computed its enrichment w.r.t. each of the phenotypes using the standard hypergeometric score. We then picked the highest score and compared it with those obtained on 100 random subsets of individuals of the same size, obtaining an empirical p-value for the biclique. Only eight of the bicliques attained p-values <0.05; none of these passed an FDR threshold of 10%.

4.7 Analysis of the genotype data

Next, we applied our algorithm to the genotype data. The preprocessed bipartite graph contained 1410 feature nodes, representing SNP states, and 72 patient nodes. The algorithm identified nine significant bicliques in this graph, constructed w.r.t. four different phenotypes: response class, grade of tumor, TP53 mutational status and T staging category. Interestingly, seven out of the eight bicliques returned by the unsupervised analysis were detected in the course of the algorithm, although none was found to be significant.

To evaluate the biological significance of the bicliques, we computed their cross phenotype scores and the expression coherency of the patient sets they define. Two of the bicliques showed markedly high cross-phenotypic scores. A third biclique exhibited significant expression coherency. These bicliques are listed in Table 1 (top).

As an example, one of these bicliques was obtained w.r.t. the grade of tumor phenotype. It consists of three feature nodes spanning two heterozygous-state SNPs of *IGF2R* (insulin like growth factor 2 receptor), and one homozygous-state SNP of the *GCLC* gene (Glutamate-cysteine ligase, catalytic subunit). The biclique contains 15 individuals, 13 of which have a tumor grade of 3. Its core consists of two SNPs, one of each gene. It has significant cross-phenotypic score w.r.t. the TP53 mutational status (enriched for status mutated). Interestingly, *IGF2R* is known to be associated with poor patient prognosis in head and neck cancer (Jamieson *et al*., 2003), and we now find it to be associated with poor prognosis in breast cancer too, as 13 out of 15 of the patients in the biclique have the disseminated, grade 3 state of the tumor. More generally, *IGF2R* is involved in a variety of cancer types, including squamous cell carcinoma, lung cancer, hepatocellular carcinoma and prostate cancer, and has also previously been shown to be involved in breast cancer. However, it is known to play a malignant role only after attaining a heterozygous state followed by the occurrence of somatic point mutations in the remaining allele (Oates *et al*., 1998)—indeed, we find it in the pathogenic biclique in the form of two heterozygous-state SNPs, in full correspondence with the existing knowledge. Finally, statistically significant differences in M6P/*IGF2R* allelic variants have been identified between Japanese and American populations, but without any clear functional significance (Killian *et al*., 2001)—the involvement of the *IGF2R* variation in the biclique provides the first support that at least in

More generally, it can be applied to detect high-scoring bipartite subgraphs that are not necessarily complete.

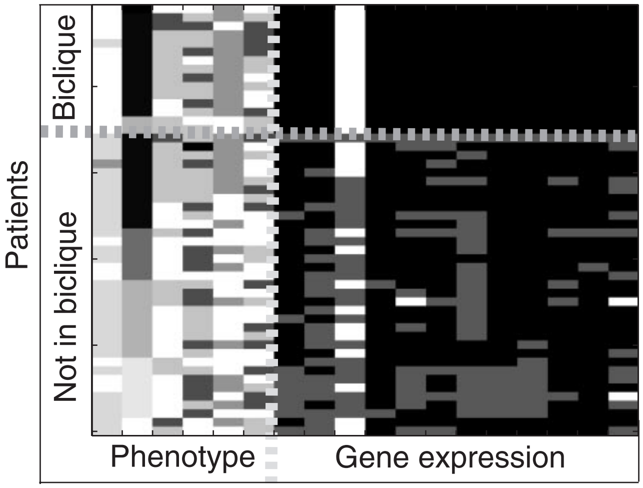

Fig. 3. A representative high scoring biclique. The figure depicts the fifth biclique listed in Table 1. Rows correspond to all individuals in the data and columns to phenotypes and features of the biclique. A horizontal line separate the patients in the biclique from all others. Columns 2 and 3 in the gene expression segment correspond to the genes that comprise the core of this biclique. The first column in the phenotype segment shows the T category annotation, according to which this biclique was constructed. Evidently, the biclique is also enriched w.r.t the array tumor classification phenotype, displayed on the second column and w.r.t the grade of tumor phenotype displayed on the third column. Interestingly, patients with the same T category outside the biclique have different array tumor classifications than those in the biclique.

breast cancer such variations may play an important pathological role.

4.8 The combined SNP-expression data

Finally, we applied our algorithm to the combined SNP-expression data. The bipartite graph we constructed contained 844 SNP state nodes, 1414 expression state nodes and 50 patient nodes. We excluded from the analysis the array tumor classification phenotype, as the latter was defined by Sorlie *et al*. based on its expression characteristics. Overall, the algorithm discovered seven significant bicliques w.r.t. two phenotypes: grade of tumor and T category. Two of these bicliques involved both SNP and expression features, while the other five contained only expression features. Expectedly, most of the bicliques were highly expression coherent. Two bicliques exhibited significant functional coherencies; one of them also showed a significant cross phenotype enrichment. These two bicliques are summarized in Table 1 (bottom).

Markedly, one of the identified bicliques, computed w.r.t. the T category phenotype, was found to be expression coherent, functional coherent and cross-phenotype enriched (Fig. 3). It consists of 12 feature nodes, all of which represent gene expression states, and 15 individuals, 14 of which have T category 4 (the highest and most invasive one in the data). The core of this biclique consists of nodes corresponding to over-expression of a zinc transporter gene *SLC39A6* and under-expression of *CKS2* (CDC28 protein kinase regulatory subunit 2). The *SLC39A6* gene has been previously associated with oestrogen-positive breast cancer and its metastatic spread to the regional lymph nodes, and it was claimed that it may play an important role in breast cancer progression (Taylor *et al*., 2003)—indeed, we find that 14 out of the 15 biclique patients have a T category of 4, i.e. larger tumors which have spread into the adjacent tissue. Interestingly, this biclique is highly enriched w.r.t. the array tumor classification phenotype, providing further support to the categorization devised by Sorlie *et al*.

5 CONCLUSIONS

Particular patterns of genotype variation and expression levels across multiple genes are believed to contribute to a predisposition to certain medical conditions and to underlie individual variation in response to medical treatments. One of the major difficulties in finding such patterns is computational; the processing time required for an exhaustive search over all combinations of genotype and expression states of a given size becomes prohibitively large even for relatively small sizes (3–4). In this study, we present a graph theoretic approach for this problem, designed to conduct an efficient search over a wide space of combinations. Unlike previous biclustering approaches, our method uses phenotype information specifically to find characteristic patterns for different phenotypic groups. We show that these patterns carry further clinical information, in addition to the phenotypic assignment according to which they were computed.

We apply our approach to a combined SNP-expression data collected from a group of breast cancer patients. We identify DPs w.r.t. several clinical phenotypes, and show that the induced patient sets exhibit coherent behavior w.r.t. other biological attributes, including different phenotypes and expression profiles.

The treatment of missing data remains a difficult challenge. In addition to the strategy we employed in this work, there are two other common strategies: data imputation and background adjustment. The first tries to complete the missing data. In our case, it seems that this strategy would have been less successful when applied to SNP data, since it relies on modeling the underlying haplotype data and such modeling would suffer from the sparse SNP data per gene in this study. The other strategy evaluates the significance of each biclique based on its corresponding background set. In our case, due to the relatively small number of individuals, this might yield ‘incomparable’ *p*-values that are based on substantially different background sets. Notably, when applying these two approaches to our data, the obtained significant bicliques substantially overlapped the ones reported here.

We believe that supervised identification of DPs will play a fundamental role in identifying novel pathogenic genomic alterations in future, larger datasets, requiring deeper (and albeit, computationally extensive) search processes. Here we already demonstrate the potential power of our approach in shedding new light on the putative role of genes that are likely to be involved in the pathogenesis and invasive spread of breast cancer.

ACKNOWLEDGEMENTS

The authors thank Oded Regev for proving the hardness of DP, Dani Yekutieli for his help with the FDR analysis and Shai Kaplan for help with the excess hypergeometric score. N.Y. is supported by the Tel-Aviv university president and rector scholarship. E.R. is supported by grants from the Yeshaya Howoritz Complexity Center, from the Tauber Foundation and the Israeli Science Foundation. R.S. is supported by an Alon Fellowship.

REFERENCES

Bar-Yehuda,R. and Rawitz,D. (2002) Approximating element-weighted vertex deletion problems for the complete *k*-partite property. *J. Algorithm.*, **42**, 20–40.

Hochbaum,D.S. (1998) Approximating clique and biclique problems. *J. Algorithm.*, **29**, 174–200.

Jamieson,T.A. *et al.* (2003) M6P/IGF2R loss of heterozygosity in head and neck cancer associated with poor patient prognosis. *BMC Cancer*, **3**, 4.

Killian,J.K. *et al.* (2001) Mannose 6-phosphate/insulin-like growth factor 2 receptor (M6P/IGF2R) variants in American and Japanese populations. *Hum. Mutat.*, **18**, 25–31.

Kluger,Y. *et al.* (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, **13**, 703–716.

Kristensen,V.N. *et al.* (2006) Genetic variation in putative regulatory loci controlling gene expression in breast cancer. *Proc. Natl Acad. Sci. USA*, **103**, 7735–7740.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

Martin,E.R. *et al.* (2001) Association of single-nucleotide polymorphisms of the *Tau* gene with late-onset Parkinson disease. *JAMA*, **286**, 2245–2250.

Moore,J.H. and Ritchie,M.D. (2004) STUDENTJAMA. The challenges of whole-genome approaches to common diseases. *JAMA*, **291**, 1642–1643.

Oates,A.J. *et al.* (1998) The mannose 6-phosphate/insulin-like growth factor 2 receptor (M6P/IGF2R), a putative breast tumor suppressor gene. *Breast. Cancer Res. Treat.*, **47**, 269–281.

Segre,D. *et al.* (2005) Modular epistasis in yeast metabolism. *Nat. Genet.*, **37**, 77–83.

Sklan,E.H. *et al.* (2004) Acetylcholinesterase/paraoxonase genotype and expression predict anxiety scores in health, risk factors, exercise training, and genetics study. *Proc. Natl Acad. Sci. USA*, **101**, 5512–5517.

Sorlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Sorlie,T. *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Takamizawa,J. *et al.* (2004) Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.*, **64**, 3753–3756.

Tanahashi,H. *et al.* (2005) Association of Lys173Arg polymorphism with CYP11B2 expression in normal adrenal glands and aldosterone-producing adenomas. *J. Clin. Endocrinol. Metab.*, **90**, 6226–6231.

Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, S136–S144.

Taylor,K.M. *et al.* (2003) Structure-function analysis of LIV-1, the breast cancer-associated protein that belongs to a new subfamily of zinc transporters. *Biochem. J.*, **375**, 51–59.

Thomas,D.C. (2004) *Statistical Methods in Genetic Epidemiology*. Oxford University Press.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Inf.*, **82**, 171–196.

Zerba,K.E. *et al.* (2000) Complex adaptive systems and human health: the influence of common genotypes of the apolipoprotein E (ApoE) gene polymorphism and age on the relational order within a field of lipid metabolism traits. *Hum. Genet.*, **107**, 466–475.

APPENDIX A

Complexity of the discriminating pattern problem

THEOREM 1. *Discriminating Pattern (DP) is NP-hard.*

Proof. Consider an instance $G = (P, F, E, w)$ of DP and let $F' \subseteq F$ induce an optimum solution, i.e. a biclique (P', F') where $P^{'} = \cap_{f \in F'} N(f)$ and $W^{*} = \sum_{p \in P'} w(p)$ is maximum. For $P^{''} \subseteq P$, denote by $W(P'')$ the sum of weights of the elements in (P''). Examine now the bipartite complement graph $\bar{G} = (P, F, \bar{E}, w)$, where $\bar{E} = (P \times F) \backslash E$. It is easy to see that $W(\cup_{f \in F'} N_{\bar{G}}(f)) = W(P) - W(\cap_{f \in F'} N_G(f))$. In other words, DP is equivalent to the following problem: Given a graph $G = (P, F, E, w)$, find a subset $F' \subseteq F$ such that $W(\cup_{f \in F'} N(f))$ is minimum. We prove the NP-hardness of this problem by reducing Set Cover to it.

Let (U, C) be an instance of Set Cover, where $C = \{S_1, \ldots, S_m\}$ is a collection of subsets of $U = \{1, \ldots, n\}$. W.l.o.g. we assume that U admits a set cover.

Construct an instance $G = (P, F, E, w)$ of DP as follows: Let $F = \{f_1, \ldots, f_m\}$ and let $P = \{p_{1,1}, \ldots, p_{1,m+1}, \ldots, p_{n,1} \ldots, p_{n,m+1}\} \cup \{z_1 \ldots z_m\}$. We define, $E = \{(f_i, z_i) : 1 \leq i \leq m\} \cup \hat{E}$, where $(f_i, p_{j,k}) \in \hat{E}$ iff $j \in S_i$. We set $w(p_{i,j}) = -1$ for all i,j and $w(z_i) = 1$ for all i. We prove that the Set Cover instance has a solution of cardinality at most r iff the DP instance has a solution of weight at most $r - (m+1)n$.

Let $H \subseteq C$ be an optimal solution for the Set Cover instance with cardinality r. Let $F' \subseteq F$ be the corresponding set of features. Clearly, $W(\cup_{f \in F'} N(f)) = r - (m+1)n$.

Conversely, let F' be a solution for the DP instance with weight at most $r - (m+1)n$. Then every vertex $p_{i,j}$ must be adjacent to some vertex $f \in F'$, implying a set cover of cardinality at most r.

APPENDIX B

Expected number of bicliques

Given a bipartite graph $G = (P, F, E, w)$ with w:$P \rightarrow \{-1,1\}$, denote by $\Psi(G, p)$ the expected number of bicliques whose hypergeometric p-value w.r.t. the partition of P to positives and negatives is at most p (Fig. A1). To estimate $\Psi(G,p)$ we first define the expected number of bicliques with exactly F' feature nodes, $|P^{'+}|$ positive nodes and $|P^{'-}|$ negative nodes:

$$\Upsilon(|F|, |F^{'}|, |P|, |P^{'+}|, |P^{-}|) = \binom{|P^{+}|}{|P^{'+}|} \cdot \binom{|P^{-}|}{|P^{'-}|} \cdot \binom{|F|}{|F^{'}|} \cdot \alpha^{|F^{'}| \cdot |P^{'}|} \cdot (1 - \alpha^{|P^{'}|})^{|F| - |F^{'}|} \cdot (1 - \alpha^{|F^{'}|})^{|P| - |P^{'}|},$$

where α is the fraction of edges in the bipartite graph, and the last two terms ensure that the chosen biclique is maximal. $\Psi(G,p)$ is then taken as the sum over all $\Upsilon(|F|, |F^{'}|, |P|, |P^{'+}|, |P^{'-}|)$ such that $\mathrm{HG}(|P|, |P^{+}|, |P^{'}|, |P^{'+}|) \leq p$.

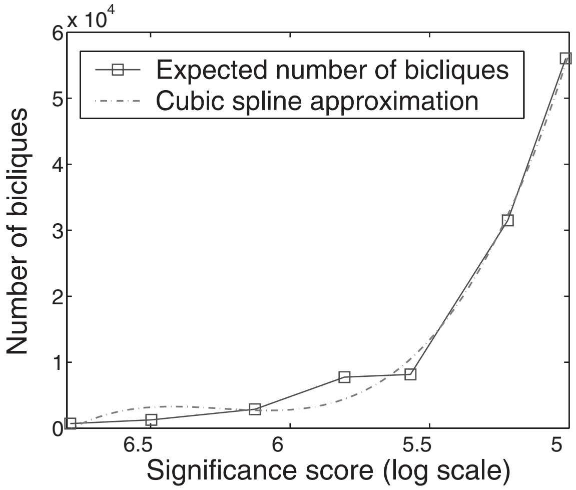

Fig. A1. Expected number of bicliques. The figure depicts the function Ψ (G, p) for different values of p, where G is the graph induced by the breast cancer genotype data.