BIOINFORMATICS

Vol. 24 ECCB 2008, pages i90–i97
doi:10.1093/bioinformatics/btn279

Clinically driven semi-supervised class discovery in gene expression data

Israel Steinfeld[1], Roy Navon[1], Diego Ardigò[2], Ivana Zavaroni[2] and Zohar Yakhini[1,*]

[1]Agilent Laboratories, Tel Aviv, Israel and [2]Departments of Internal Medicine and Biomedical Sciences, University of Parma, Parma, Italy

ABSTRACT

Motivation: Unsupervised class discovery in gene expression data relies on the statistical signals in the data to exclusively drive the results. It is often the case, however, that one is interested in constraining the search space to respect certain biological prior knowledge while still allowing a flexible search within these boundaries.

Results: We develop an approach to semi-supervised class discovery. One component of our approach uses clinical sample information to constrain the search space and guide the class discovery process to yield biologically relevant partitions. A second component consists of using known biological annotation of genes to drive the search, seeking partitions that manifest strong differential expression in specific sets of genes. We develop efficient algorithmics for these tasks, implementing both approaches and combinations thereof. We show that our method is robust enough to detect known clinical parameters in accordance with expected clinical values. We also use our method to elucidate cardiovascular disease (CVD) putative risk factors.

Availability: MonoClaD (Monotone Class Discovery). See http://bioinfo.cs.technion.ac.il/people/zohar/MonoClad/

Supplementary information: Supplementary data is available at http://bioinfo.cs.technion.ac.il/people/zohar/MonoClad/software.html

Contact: zohar_yakhini@agilent.com

1 INTRODUCTION

Two cell types with dramatically different biological characteristics are expected to yield very different gene expression profiles (e.g. normal cells versus tumor cells from the same tissue or endothelial cells from around blood vessel lesions versus endothelial cells from normal arteries). Indeed, genomic studies, specifically ones that are based on high-throughput molecular profiling, often focus on comparing two or more sample sub-populations included in the data. For example, Golub *et al.* (1999) studied the differential gene expression as measured when comparing acute myeloid leukemia (AML) to ALL samples. Bittner *et al.* (2000) applied several clustering methods on expression profiles of melanoma tumors. They discovered a classification that was then further substantiated by ascertaining phenotypical differences. Alizadeh *et al.* (2000) compared DLBCL cells to other types of lymphoma and to healthy T-cells and B-cells. Applying agglomerative clustering over genes they managed to find genes with similar behavior, across the different types of samples. From the resulting hierarchy they manually selected specific subsets of genes. Finally, by restricting to a particular subset of genes, they applied clustering on the samples to discover a partition of the DLBCL samples.

Such methodology of supervised class discovery suffers from the need for manual human curation. This intervention is required since typical clustering procedures, used in gene expression analysis, attempt to find groups of samples such that the overall expression profiles are similar within clusters and different between clusters. In addition, it is important to realize that very often the majority of the active cellular mRNA is not affected by biological differences, even by very significant ones. That is, a dramatic biological difference does have a gene expression level manifestation, but the set of genes that is involved, representing specific biological processes, can be rather small, as a fraction of the entire mRNA repertoire. Such differences are 'washed out' by uniform measures of similarity between samples (e.g. Pearson or Euclidean distance). For example, the classification discovered by Alizadeh *et al.* (2000) is not apparent when tissues are clustered using all the genes. In this particular case, a set of relevant genes was identified based on other considerations and prior hypotheses about potential sources of differences between DLBCL subtypes.

The need to identify sub-classes in nominally uniform data has led to the development of several unsupervised class discovery methods (Ben-Dor *et al.*, 2001b; von Heydebreck *et al.*, 2001). While these methods are totally unsupervised, the complexity of the problem in question is exponential and discovery relies on heuristics. More importantly, these methods are designed to find the strongest statistical trend in the data. In practice, the resulting partition might not be relevant to the research question and therefore is poorly understood.

Disease related studies are usually accompanied with significant clinical data that is missed in all aspects of expression profiling mediated class discovery. In the current work we address the use of external information to automatically drive the class discovery processes, thus overcoming several of the shortcomings of the aforementioned approaches. While clustering approaches are totally un-supervised and while classification (that is—the discovery of genes differentially expressed between two types of samples) is totally supervised, our proposed approach is, in effect, a form of clinically driven semi-supervised class discovery.

The process we develop in this article has two components:

(1) An optimized search process. In this work we introduce a search procedure that is constrained to respect some external

*To whom correspondence should be addressed.

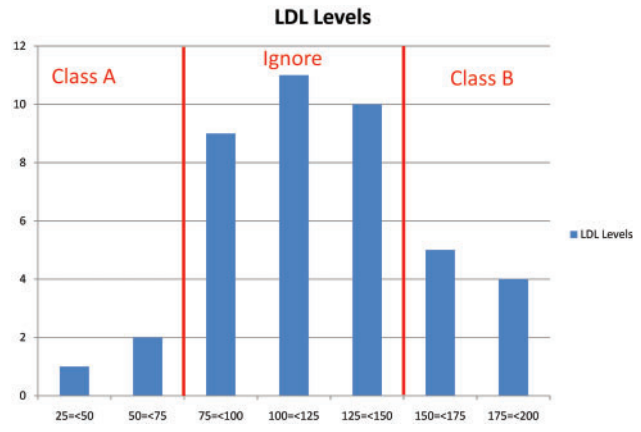

Fig. 1. Distribution of LDL levels in a cohort of 46 healthy people (see Section 3.1 for details of the data). Marked in red lines are two thresholds, t and u, that define the sample assignment to Class A and B. Where Class A will consist of all samples with LDL levels below t and Class B will consist of all samples with LDL levels above u. Once a partition is defined, one can use any method to assess differential expression. Note that methods that afford p-values will be better in handling the variance in class sizes when considering different pairs of thresholds and possibly different phenotypes.

quantitative measurement. For example, when LDL levels[1] are available for the subjects of the sample set, we seek a partition that defines high and low LDL. That is—we only allow partitions that respect the order of the quantitative measurement (Fig. 1).

(2) Figure of merit associated to putative partitions. We are guided by the fact that biologically meaningful partitions typically manifest either: (a) large (statistically meaningful) overabundance of differentially expressed genes in the different sample classes or (b) an enrichment of a meaningful subset of genes, usually commonly related to a specific biological process, amongst the differentially expressed genes. We use efficient statistical methods to assess enrichment in a ranked list without requiring a threshold to be set on the differentially expressed genes (Eden *et al.*, 2007).

A preliminary short abstract describing this approach appears in Steinfeld *et al.* (2007).

We exemplify our approach through the analysis of data comprising quantitative phenotypic measurements and expression profiling from healthy individuals. To analyze the data we developed a semi-supervised class discovery method, constraining the search space to patterns that respect an order induced by the rich quantitative annotations. We show that our method is robust enough to detect known clinical parameters with accordance to expected values. We also apply our method, using a community curated sets of genes, to elucidate novel cardiovascular disease (CVD) putative risk factors.

[1] **LDL** (low-density lipoprotein) are circulating particles of lipids, phospholipids and proteins acting as transporters of cholesterol to the peripheral body tissues. An excess of circulating LDL cholesterol is responsible for increased cholesterol deposition in the artery walls leading to atherosclerosis and cardiovascular diseases (myocardial infarction, stroke, etc.).

2 METHODS

2.1 Differential expression for quantitative phenotypes

One of the basic tasks in gene expression data analysis is finding differentially expressed genes between two classes (such as tumor versus normal or diabetics versus non-diabetics). A variety of methods were developed to address this task, such as TNoM (Ben-Dor *et al.*, 2001a), SAM (Tusher *et al.*, 2001) and others (see review of Cui and Churchill, 2003). In common practice bioinformaticians typically use categorical information on the samples to derive partitions. In this section, we assume that the gene expression data is such that each sample is associated with various clinical phenotypes. Some of these phenotypes are quantitative (numbers); e.g. blood pressure, BMI and LDL. In this case, one could partition the data using the quantitative information by either of the following approaches (Fig. 1):

Parametric: Setting two values as thresholds (t and u)—all samples whose chosen phenotype value is below t will be in Class A, and all samples whose chosen phenotype value is above u will be in Class B. The rest of the samples (between t and u) will be ignored.

Non-parametric: Setting two percentile values as thresholds.

Using the quantitative information to partition the samples maintains the biological context of the analysis and enables the researcher to better interpret the result. Furthermore, by enabling dual thresholds, we focus on extreme biological states while ignoring samples (with mid-range values) that might confound the analysis.

2.2 Evaluating partitions

As part of a class discovery process, one needs to assess the statistical significance of any partition considered, and to compare between partitions. In this article we consider two general approaches to evaluating the statistical significance of a partition:

Overabundance analysis: Using a differential expression score that affords an exact p-value and can be efficiently computed, one can estimate the expected number of differentially expressed genes for any given partition. By comparing the observed number of differentially expressed genes to the expected number, under a null model, we can calculate the overabundance of differentially expressed genes. This quantity can be used as a figure of merit, indicating a more profound change in the cell state. This approach has been described and used in Ben-Dor *et al.* (2001b), von Heydebreck *et al.* (2001) and others.

Other effective measures for estimating the overabundance of differential expression can include the number of genes that pass a Bonferroni correction, or the number of genes at a given false discovery rate (FDR) (Benjamini and Hochberg, 1995) level.

Set enrichment: Part of the semi-supervised approach of this article involves the use of external information in driving the statistical assessment of differential expression. Consider, for example, a universe gene set $G = \{g_i\}_{i=1...N}$, and a set of genes all participating in the same biological process, which we shall denote by $T : G$. Consider a candidate binary partition $Q = (A, B)$ of the mRNA expression data and assume that for every transcript g we computed a differential expression score, reflecting under/over expression in A as compared to B, denoted d(g). Rank the transcripts according to d(g), where the most significant transcripts are at the top of the list. We will assign a statistical score to the enrichment of T at the top of this list. This enrichment score, $\varphi(Q, T, G)$, is a figure of merit that can be used to find partitions where an activity of T is evident in the mRNA differential expression. $\varphi(Q, T, G)$ is computed using the minimal hyper geometric (mHG) statistics (Eden *et al.*, 2007, and Appendix A), as described in Section 2.4.

The enrichment procedure can be used with GO terms, TF cohorts, KEGG classes, miRNA target sets (as inferred from databases such as TARGETSCAN), genomic intervals and sets of genes derived from other studies. Some of these approaches are exemplified in the Section 3.

2.3 Monotone class discovery

Typical class discovery in gene expression data (Ben-Dor *et al.*, 2001b; von Heydebreck *et al.*, 2001) searches over all possible partitions of the set of samples. As this collection is exponential in the number of samples, L, heuristic methods, such as simulated annealing (Kirkpatrick *et al.*, 1983) are typically used.

When taking a semi-supervised approach we are seeking partitions that respect (or are monotone with respect to) some independently measured quantitative phenotype. In this case, we address a different biological question and reduce the search space from $\Theta(3^L)$ to $\Theta(L^2)$ making the search far more tractable.

A formal definition of monotone class discovery follows. Consider gene expression data given as a matrix D. Rows are genes $G=\{g_i\}_{i=1\ldots N}$ and columns are samples $S=\{s_j\}_{j=1\ldots L}$. Assume that we also have a quantitative phenotype measured for the set of samples. For each $s\in S$, we therefore have a number $q(s)$. Without loss of generality we further assume that $q(s_1)\leqslant q(s_2)\leqslant\cdots\leqslant q(s_L)$. A monotone partition of S is a pair of disjoint subsets $A=\{s_1, s_2, \ldots, s_t\}$ and $B=\{s_u, s_{u+1}, \ldots, s_L\}$, where $t<u$. Consider a statistical figure of merit φ that can be computed for any partition of S. Monotone class discovery seeks a monotone partition P for which $\varphi(P)$ is optimal. Examples are described in Section 2.2.

2.4 Search heuristics

The simplest, very naïve, approach to monotone class discovery, would be to consider all possible pairs of thresholds t and u, evaluate φ for the corresponding partition and return the optimum found. Considering we have N genes and L samples, this search will require $O(N*L^2)$ time complexity. For large datasets this can be prohibitive and we are interested in developing a more efficient method. For the case of the semi-supervised class discovery with set-enrichment as a figure of merit, we present the following heuristic approaches. This section can be skipped for general understanding of the article. It is specific to the faster search heuristic.

Following the description in Section 2.2, consider a gene universe $G=\{g_i\}_{i=1\ldots N}$, and a gene set of interest, T:G. For any candidate binary partition $Q=(A, B)$, and a fixed n, representing the top Q-differentially expressed genes, consider the hypergeometric tail distribution as a figure of merit to score Q. Namely, $\varphi(Q, T, G)=\mathrm{HGT}(N, B, n, b)$ (see Appendix), where $N=|G|$, $B=|T|$ and b is the number of elements in T that occur in the top n Q-differentially expressed genes, which we also denote by $b_n(Q, T, G)$; this latter notation emphasizes b's dependence on the universe G. Since N, B, and the threshold, n, are all fixed, the HGT score is monotone with $b=b_n(Q, T, G)$. Therefore, during an exhaustive search over all partitions, we maintain the maximum b so far, $b_{\max}$, and trace it back to the partition where it was obtained, at the end of the search. A simple pseudo-code for the algorithm BP-HGT$_{\mathrm{all}}$ (Best-Partition-HGT$_{\mathrm{all}}$) is presented:

```
BP-HGT_all(T,G) =
1:  b_max = 0
2:  for all partitions Q:
3:          b = b_n(Q,T,G)
4:          if (b > b_max)
5:                  b_max = b
6:  return b_max
```

The main idea of our heuristic approach is to avoid the repeated calculation of differential expression scores of all genes, which is needed in Line 3, for every partition. By working with a carefully selected reduced universe we only compute differential expression scores when the currently considered partition stands a chance of exceeding $b_{\max}$. A full description of the algorithm is described in the following pseudo-code:

```
BP-HGT_jump(T,G) =
1:  R = random gene set of size r, where n<r<N
2:  G_R = RUT                          // the reduced universe
3:  b_max = 0
4:  for all partitions Q:
5:          b = b_n(Q,T,G_R)
6:          if (b > b_max)
7:                  b = b_n(Q,T,G)
8:                  R = the top r Q-diff-expressed genes
9:                  G_R = RUT
10:                 if (b>b_max)
11:                         b_max = b
12: return b_max
```

CLAIM. $BP\text{-}HGT_{\mathrm{jump}}(T, G)=BP\text{-}HGT_{\mathrm{all}}(T, G)$

PROOF. Assume to the contrary that

(1) BP-HGT$_{\mathrm{jump}}(T, G)$ $<$ BP-HGT$_{\mathrm{all}}(T, G)$.

Let Q^* be the partition for which BP-HGT$_{\mathrm{all}}$ attains maximum enrichment. Let $b^*=b_n(Q^*, T, G)$. Our assumption (1) implies that when considering Q^*, BP-HGT$_{\mathrm{jump}}$ did not reach Line 7, for which b is computed exhaustively on all genes and b^* would have been found and replaced the then current $b_{\max}$.It follows that there exists a set R for which $b_n(Q^*, T, G_R)$ (computed in Line 5) $<b_{\max}<b^*$. For this to happen, in the reduced universe G_R there need to be strictly more than $(n-b^*)$ genes not from T that are in the top n Q^*-differentially expressed genes. Since $T\subset G_R\subset G$, in G there are strictly more than $(n-b^*)$ genes not in T amongst the top n Q^*-differentially expressed genes (there can be new genes in this list, but not from T). This in turn means that $b_n(Q^*, T, G)<b^*$. A contradiction. ■

Our selection of R is based on the assumption that similar partitions result in similar differential expression patterns. Following this assumption, R is consistently updated to be the top r differentially expressed genes, each time we calculate differential expression for all genes. We ran empirical tests to investigate optimal values for r. See Section 3.2 for details.

The above heuristic still requires calculating differential expression on $O(|G_R|)$ genes for each partition. For any selection of R our algorithm complexity would be $\Omega(N+(r+B)*L^2)$. The only definite upper bound we can provide is equivalent to the naïve exhaustive approach. However, in typical cases we can select $r<<N$ that still enables avoiding, for most partitions, the calculation of differential expression for all of G. Thus, we do get significant acceleration in practice, as described in Section 3.2.

Also, in practice we are usually only interested in partitions that yield a p that is better than a fixed threshold. $b_{\max}$ can, therefore, be initialized to be the minimum b for when $\mathrm{HGT}(N, B, n, b)<p$. We also found that improvement in running time is achieved by initializing R (Line 1) with the r genes that are most correlated with the quantitative phenotype.

For adaptation of BP-HGT$_{\mathrm{jump}}$ to the mHGT statistics, see Appendix A1.4.

2.5 GO enrichment and visualization

GO enrichment analyses were performed using GOrilla (http://cbl-gorilla.cs.technion.ac.il).

All gene expression heatmap visualizations were generated using Mayday software (Dietzsch*et al.*, 2006).

2.6 Samples and microarrays

Enrolled subjects were all offspring of participants in the Barilla study cohort, a longitudinal survey started in 1981 to investigate the impact of classical and novel risk factors on CVD development (Zavaroni *et al.*, 1989, 1999). All

volunteers were young adults [median age of 35 years, Inter-Quartile Range (IQR) = 7], without clinically relevant diseases or chronic medications. In addition, all subjects had a low CVD risk profile and major risk factors were within the normal range for most of the population.

Peripheral blood mononuclear cells (PBMCs) were isolated from fresh blood drawn in fasting conditions and DNA-free total RNA extracted with standard techniques. Hybridization to 44K oligonucleotide arrays by Agilent Technologies (Santa Clara, CA, USA) was performed according to standard protocols.

2.7 Description of available software (MonoClaD)

A program which performs this class discovery is available along with the Supplementary Material at the MultiKnowledge Project's website: http://bioinfo.cs.technion.ac.il/people/zohar/MonoClad/

The input to the program is:

- Gene expression matrix.
- Quantitative phenotype vector (one value for each sample and the order should be consistent with the order of the columns in the expression data matrix).
- Optional: a set of genes to drive enrichment-based class discovery.

The program performs monotone class discovery (using either over-abundance or set enrichment). It returns the thresholds of the quantitative phenotype for which differential expression is maximized and the list of genes ranked by their differential expression. GO enrichment can then be performed on this list by using the output file as input into GOrilla (http://cbl-gorilla.cs.technion.ac.il). To perform this either RefSeq or UniProt names should be used in the input files. A more detailed manual for running this software package can be found in the above URL.

3 RESULTS

3.1 CVD data

We applied our method to PBMCs gene expression profiling data, collected from 46 healthy subjects (see Section 2). PBMCs are a sub-population of circulating white blood cells highly involved in the inflammatory processes responsible for atherosclerosis and CVD (Libby *et al.*, 2002). Recent evidence suggests that PBMCs may act as 'biosensors' of systemic diseases and their response to CVD risk factors, assessed by gene expression profiling, provides a biological signature of atherosclerosis (Ardigo *et al.*, 2007).

Clinical, laboratory measurement and CVD prognostic indicators were also collected, adding more than 160 phenotypic quantitative parameters for each subject. Each one of the phenotypes can be used as a basis for computing the associated differential expression, as reflected in the subject's PBMCs, as well as for semi-supervised class discovery.

3.2 Tests of the algorithmic efficiency

In Section 2.2, we presented the set enrichment class discovery partition scoring method. Namely, for a given threshold n, the set enrichment method seeks the best partition, Q, which maximizes the enrichment of a gene set, T, in the top n Q-differentially expressed genes. Our set enrichment heuristic class discovery (see Section 2.4) is strongly relying on the ability to work with a reduced universe set of genes, G_R, for most partitions tested. The reduction of the universe is done by focusing on the genes in T and on the top r differentially expressed genes, for the current Q.

To investigate the optimal size of r we ran empirical tests on various values of n and B (=$|T|$, the size of the set driving the search).

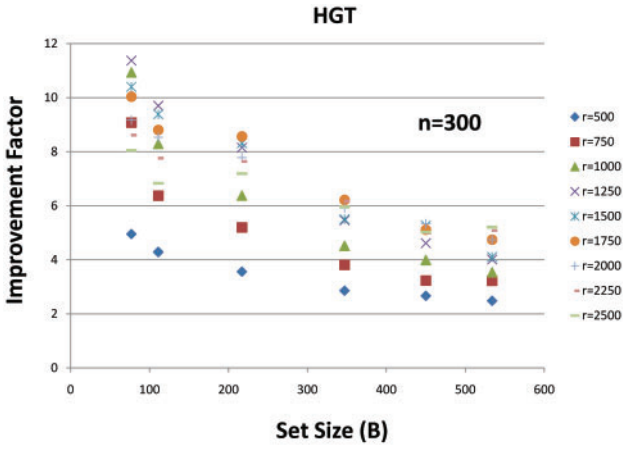

Fig. 2. The improvement in running time as a function of B and r (see text). In all tests the threshold, n, is fixed to 300. Note that we get strong efficiency factors when B is fairly small, sae is true for small n's (Supplementary Table A).

Table 1. Diffreent functions $F(n,B)$ for estimating an optional r (see text)

F(n,B)	Correlation to $r^*(n,B)$
sqrt(B^*n)	0.9
sqrt(B^*n)/log(B)	0.95
sqrt(B^*n)/log(n)	0.83
sqrt(B^*n)/log(n^*B)	0.90
$B+n$	0.92
B^+n	0.85
(B^+n)/log(B)	0.95
(B^+n)/log(n)	0.84

For each pair of values for n and B the optimal value of r, r^* (n,B), was calculated empirically (Supplementary Table B). The Pearson correlation coefficient between $r^*(n,B)$ and $F(n,B)$ is presented. The correlation was computed for data spanning $n=1$ [100, 300, 500] and B=[77, 111, 217, 347, 450, 534]. The highest correlation is obtained for $F(n,B)=(B+n)/\log(B)$.

For each size n and gene set, of size B, we run the naïve approach as well as the heuristic approach with multiple choices of r (Fig. 2). For each instance of r we profiled the time efficiency over 10 randomly selected quantitative sample annotations vectors from the data described in Section 3.1.

In addition to Figure 2 we describe more tests of n and B in Supplementary Table A. In practice, when analyzing a set T of size B using a threshold n, the user can infer an optimal $r=r^*(n,B)$ from the test results (Supplementary Table A). We also tried to fit a function $F(n,B)$ to the empirically computed values of $r^*(n,B)$. Correlation values of $r^*(n,B)$ and various functions are presented in Table 1.

3.3 Recapturing known quantitative traits

Our dataset, described in Section 3.1, is composed of 26 females and 20 males. To test our methods, we used available quantitative sample annotations that are known to have different value distributions for the two gender sub-populations. The quantitative annotations

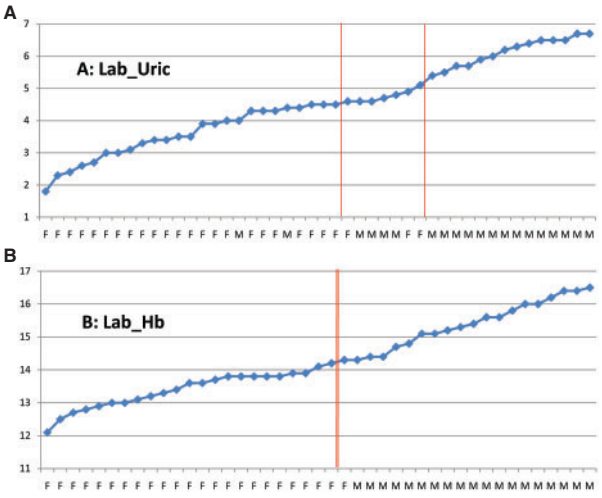

Fig. 3. Best partitions resulting from analyzing two quantitative sample annotations related to the subject's gender. Samples are ordered according to their quantitative values and are presented as a function of the sample subject gender (F – female, M – male). (**A**) Using the uric acid levels in the blood and seeking the highest overabundance of differentially expressed genes. (**B**) Using haemoglobin levels in the blood and seeking the partition that is most enriched with heterosome genes. In the partition attained we observe enrichment at mHG $p < 10^{-32}$(corrected by 46 choose 2, for multiple testing) of heterosomes in the top differentially expressed genes.

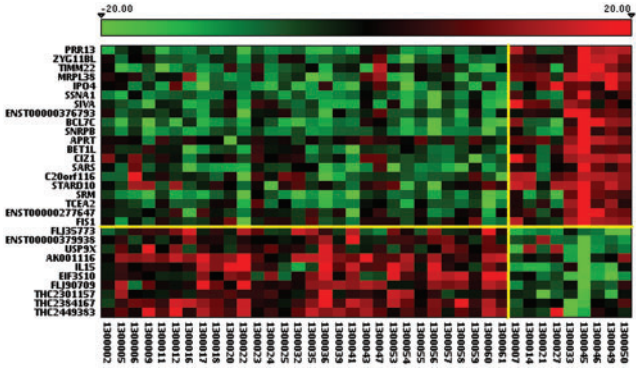

Fig. 4. Heatmap of 30 genes that are top differentially expressed in the optimal partition inferred by IMT_all_max quantitative sample annotation. Each line represents a gene and each column represents a sample. The scoring scheme, used for the partition search, was overabundance analysis (see Section 2). The nine samples on the right have high IMT level ($\geqslant$0.92). The 30 samples on the left have low IMT level ($\leqslant$0.9). The partition of the IMT levels closely agrees with the known levels in the literature.

selected are the hemoglobin and uric acid levels measured in peripheral blood of the subjects. Our data shows a perfect separation of the sub-populations by the hemoglobin sample annotation (TNoM $p < 10^{-11}$), and also a significant separation by the Uric acid sample annotation (TNoM $p < 10^{-7}$).

We used two methods for scoring partitions: overabundance of differentially expressed genes and gene set enrichment (see Section 2).

Figure 3A illustrates the monotone partition with the highest overabundance of differentially expressed genes, when using the uric acid quantitative sample annotation. The separation between males and females strongly agrees with the discovered partition. The partition induces the differential expression depicted in Supplementary Figure A. It is important to note that the separation of the two sub-populations was obtained solely by comparing their transcriptional profiles while respecting monotonicity.

The use of overabundance of differentially expressed genes to score partitions is very useful when such high differences are available and expected. In some cases, though, it is possible that quantitative phenotypic differences will yield a change only on a single biological process. In this case we would not expect an overabundance of differentially expressed genes, but only differential expression of a particular set of genes, representing the biological process we are interested in. To handle such cases we developed the set driven partition scoring (see Section). We first characterized the set of genes that are expected to differ between the gender sub-populations. A total of 2034 genes residing in either one of the sex chromosomes X or Y (Maglott *et al.*, 2005), were collected. On the expression profiling microarray 557 of the heterosome genes were found to be present, and therefore represent the set of heterosome genes to be used in the enrichment analysis.

Figure 3B illustrates the partition for which the heterosome genes are most enriched (mHG $p < 10^{-32}$, corrected for multiple testing) in the deferentially expressed genes, when using the hemoglobin sample annotation.

Again we see that by using pre-existing biological knowledge, of expected differentially expressed genes, we manage to drive the partition search and find the optimal partition that is inherent by the quantitative sample annotation.

3.4 Finding novel risk factors for CVD

In addition to clinical parameters that are risk factors for CVD, a few non-invasive tests can be used to assess the presence of vascular atherosclerosis at a preclinical stage—i.e. before the disease leads to major clinical manifestations such as myocardial infarction of stroke. Carotid intima-media thickness (IMT) is an ultrasound-based measurement of thickness of the inner layer of the carotid artery and provides information on the infiltration of LDL cholesterol and inflammatory cells in the artery wall. IMT is a powerful predictor of future CVD events (Hurst *et al.*, 2007) and has been imposing as surrogate end-point in short term clinical trials to assess the efficacy of a treatment in preventing atherosclerosis progression (Bots, 2006). Although at present no clear cut-off level of normality has been established for IMT (as it varies as a function of age, gender and BMI), an intima-media layer thicker than 1 mm should be considered as a pathological value at any age (Touboul, 2007).

Using semi-supervised class discovery based on overabundance and respecting monotonicity of the IMT values we received IMT threshold levels that are in close agreement with the known prognosis values (IMT levels below 0.9 in Class A versus IMT levels above 0.92 in Class B). The differentially expressed genes in this partition (Fig. 4) were enriched with GO terms related to vesicle-mediated transport ($p < 10^{-8}$) and glycolysis ($p < 10^{-6}$), giving mechanistic insights to the difference between the two cell states.

We observed even more interesting finding by applying the set enrichment partition scoring to drive the partition search. In this approach we used 300 genes that have a community established

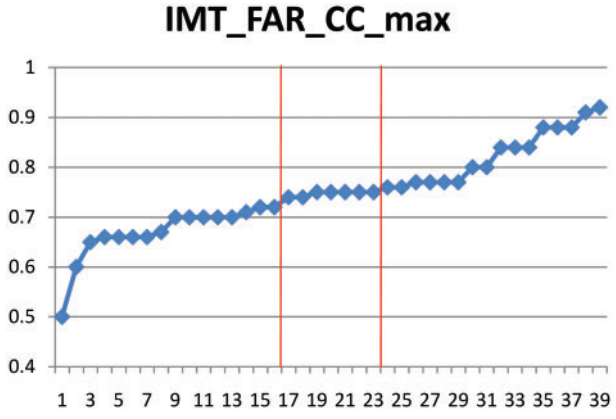

Fig. 5. Optimal partition obtained using set driven class discovery (see Section 2). A total of 300 genes related to CVD were collected and used to drive the partition search. Amongst the ∼160 available quantitative sample annotations, the annotation that received the highest enrichment was an IMT related one, previously shown to be correlated to CVD development (Lorenz *et al.*, 2007) (mHG $p < 10^{-4}$, after correction for the multiple partitions and quantitative annotation tested).

association with CVD (UCL web site—http://www.ucl.ac.uk/medicine/cardiovascular-genetics/geneontology.html). Searching over all available parameters (∼160, see Section 3.1), a particular IMT parameter gave highly informative partition regarding the CVD genes (Fig. 5). The 300 CVD-related genes were enriched at mHG $p < 10^{-9}$, which translate to $p < 10^{-4}$ after correction for the multiple partitions and quantitative annotation tested. We further ran 100 random sets of size 300 through the same analysis, using the real quantitative annotations and data. The best result obtained was mHG $p \sim 10^{-7}$,which translates to a corrected $p \sim 10^{-2}$, as expected.

This result is interesting since (1) an IMT parameter shows the highest correlation with the expression pattern of CVD-related genes. (2) Among the multiple parameters that are measured during an IMT test, the one showing the most informative correlation to gene expression is also the one considered to have the highest correlation with CVD events prediction (Lorenz *et al.*, 2007). (3) The identified IMT threshold values indicate that the expression of CVD-related genes begins at a lower than expected IMT value. (4) GO enrichment analysis of the differentially expressed genes in the depicted partition (Fig. 4) can provide mechanistic insights in the progression of the disease. We see a general activation of the immune response (Table 2), which is in line with the inflammatory hypothesis of atherosclerosis (Packard and Libby, 2008); moreover—we observe the regulation of IL6 biosynthesis that was previously associated with CVD.

We further note that in taking a simplistic approach and ranking the genes according to their Pearson correlation to any of the quantitative sample annotations (including the IMT annotations), no enrichment of the CVD genes was found (data not shown). This underscores the utility of methods based on differential expression.

3.5 Additional example

To assess the applicability of our methods in a broader biological context we analyzed several additional published datasets. As an example, we present the results obtained from van't Veer *et al.* (2002). The data consists of 117 primary breast tumor expression profiles. The authors report a gene expression signature that is strongly predictive of short interval to distant metastases. Many of the signature genes partake in processes such as cell cycle, invasion and metastasis. We used the GO term cell cycle process (GO: 0022402) to drive semi-supervised class discovery, with the length (in months) of the interval to metastasis as the constraining quantitative annotation of the samples. Using a single threshold the resulting partition was up to 61 months in Class A and more than 62 months in Class B, with an enrichment of $p < 10^{-17}$. This result is interesting as it is similar to the partition used in the article and to the common clinical use (<5 years and >5 years). More interestingly, using two thresholds (excluding samples with median values) we receive a partition of up to 18 months and more than 124 months with an enrichment of $p < 10^{-34}$.

Table 2. List of GO terms enriched in the differentially expressed genes induced by the optimal partition constrained by IMT value and searched by CVD related genes set class discover

GO Name	*P*-value	Enrichment (N, B, n, b)
Response to other organism	4.10E-07	2.73 (9825, 447, 354, 44)
Inflammatory response	5.00E-07	3.99 (9825, 174, 354, 25)
Response to pest, pathogen or parasite	2.10E-06	2.19 (9825, 442, 608, 60)
Cellular defense response	1.10E-05	3.85 (9825, 84, 608, 20)
Response to wounding	1.90E-05	2.35 (9825, 310, 608, 45)
Immune response	2.30E-05	2.18 (9825, 676 360, 54)
Response to external stimulus	3.90E-05	2.54 (9825, 400, 367, 38)
Positive regulation of immune response	7.80 E-05	5.99 (9825, 52, 347, 11)
Interleukin-6 biosynthesis	8.20E-05	21.20 (9825, 6, 309, 4)
Macrophage activation	8.20E-05	21.20 (9825, 6, 309, 4)

GO enrichment was computed using GOrilla (http://cbl-gorilla.cs.technion.ac.il).

4 DISCUSSION

In this article we introduce fast algorithmics for semi-supervised class discovery, where the search is constrained by clinical quantitative information. We show data-driven analysis of expression data and describe results that shed more light on the driving clinical parameters as well as on the associated expression profiles. We demonstrate gained biological insight by pointing out IMT values at which a CVD pathogenesis process might be on going, as evidenced by expression profiles that are more similar to disease profiles than to normal ones.

In the scope of this article we do not address the case of clinical data information represented as discrete classes rather than as numerical quantities. Our monotone class discovery (see Section 2.3) can be adapted to this scenario by exhaustively searching all possible trinary assignments. Namely—each original sample class can be fully assigned to either class of the partition or not assigned at all. This search is exponential in K $(3^K - 2^{K+1} + 1)$, the number of original classes. Furthermore, in our study the monotone class discovery was driven by a single quantitative sample annotation. This method can be further extended to take under consideration multiple quantitative sample annotations

(e.g. considering age and height, young and high subjects will be compared to old and short subjects). To add robustness, with respect to the noise in clinical data, it is useful to extend the methods to address less strict monotonicity conditions, by allowing a few exceptions (not all samples in one class need to have lower IMT than in the other, only most of them do).

In Section 3.4 we show how a predefined CVD related set of genes, assembled and characterized by the scientific community, can help in driving the analysis of gene expression data. The use of predefined sets of biologically related genes is very common in analyzing high-throughput genomic and genetic data. Gene Ontology (Harris *et al*., 2004) is a good example of the scientific community predefining biologically related sets of genes. Many other classifications are also frequently used (Kanehisa, 2002; Subramanian *et al*., 2005). Our method of set driven class discovery (Section 2.2) can be easily extended to handle a collection of multiple sets. In this case a partition will be scored according to the most enriched set amongst the collection of sets under consideration. We believe it to be of particular interest as these ensembles of sets are more often used to assess results rather than as a driving force in such analyses. Proper statistical corrections need to be applied for such an approach.

It is important to note that a predefined set of related genes is not limited to the available ensembles described earlier. In fact, in the set-driven class discovery, the results from a different study can be used to drive the analysis. For example, the set of differentially expressed genes between tumor and normal samples in one study can drive the analysis in a study evaluating a quantitative measure to test cancer progression.

It is often the case that the results of a differential expression study are represented as a ranked list of genes rather than as a fixed set of genes (Eden *et al*., 2007). An extension of our methods would score a putative partition, Q, according to the agreement between the differential expression ranks for a known condition and that implied by Q.

ACKNOWLEDGEMENTS

We thank Eran Eden and Amir Ben-Dor for useful discussions. We thank the ECCB anonymous reviewers for excellent comments.

Funding: This study is partially supported by the European Union with a grant in the framework of the FP6 Multi Knowledge Project (#FP6-IST-2004-027106).

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Ardigo,D. *et al*. (2007) Application of leukocyte transcriptomes to assess systemic consequences of risk factors for cardiovascular disease. *Clin. Chem. Lab. Med.*, **45**, 1109–1120.

Ben-Dor,A. *et al*. (2001a) Tissue classification with gene expression profiles. *J. Comput. Biol*., **7**, 559–583.

Ben-Dor,A. *et al*. (2001b) Class discovery in gene expression data. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB'01)*. ACM Press, pp. 31–38.

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.

Bittner,M. *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature,* **406**, 536–540.

Bots,M.L. (2006) Carotid intima-media thickness as a surrogate marker for cardiovascular disease in intervention studies. *Curr. Med. Res. Opin*., **22**, 2181–2190.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*., **4**, 210.

Dietzsch,J. *et al*. (2006) Mayday - a microarray data analysis workbench. *Bioinformatics*, **22**, 1010–1012.

Eden,E. *et al*. (2007) Discovering motifs in ranked lists of DNA Sequences. *PloS Comput. Biol.*, **3**, e39.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.

Hurst,R.T. *et al*. (2007) Clinical use of carotid intima-media thickness: review of the literature. *J. Am. Soc. Echocardiogr.*, **20**, 907–914.

Kanehisa,M. (2002) The KEGG database. *Novartis Found Symp*., **247**, 91–101.

Kirkpatrick,S. *et al*. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Libby,P. *et al*. (2002) Inflammation and atherosclerosis. *Circulation*, **105**, 1135–1143.

Lorenz,M.W. *et al*. (2007) Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis. *Circulation*, **115**, 459–467.

Maglott,D. *et al*. (2005) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33** (Database issue), D54–D58.

Packard,R.R. and Libby,P. (2008) Inflammation in atherosclerosis: from vascular biology to biomarker discovery and risk prediction. *Clin. Chem*., **54**, 24–38.

Steinfeld,I. *et al*. (2007) Semi-supervised class discovery using quantitative phenotypes – CVD as a case study. *BMC Bioinformatics*, **8** (Suppl. 8), S6.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA,* **102**, 15545–15550.

Touboul,P.J. *et al*. (2007) Mannheim carotid intima-media thickness consensus (2004–2006). An update on behalf of the advisory board of the 3rd and 4th watching the risk symposium, 13th and 15th European stroke conferences, Mannheim, Germany, 2004, and Brussels, Belgium, 2006. *Cerebrovasc Dis*., **23**, 75–80.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

von Heydebreck,A. *et al*. (2001) Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, **17**, S107–S114.

van't Veer,L. *et al*. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* **415**, 530–536.

Zavaroni,I. *et al*. (1989) Risk factors for coronary artery disease in healthy persons with hyperinsulinemia and normal glucose tolerance. *N. Engl. J. Med.,* **320**, 702–706.

Zavaroni,I. *et al*. (1999) Hyperinsulinemia in a normal population as a predictor of non-insulin-dependent diabetes mellitus, hypertension, and coronary heart disease: the Barilla factory revisited. *Metabolism*, **48**, 989–994.

APPENDIX A

A1 THE mHG STATISTICS

A1.1 HG: the hypergeometric distribution

Consider the following scenario. A closet contains two drawers, together containing N socks. One drawer contains n socks and the other $(N-n)$. Exactly B of the socks are black and the remaining $(N-B)$ are white. A question that can be asked is: Does the first drawer contain significantly more black socks than the second? In other words, are the black socks enriched in the first drawer? Under a uniform distribution over all configurations of socks in the drawers the probability of finding exactly b black socks in the first drawer is

described by the hypergeometric function:

$$\mathrm{HG}(N,B,n,b)=\frac{\binom{n}{b}\binom{N-n}{B-b}}{\binom{N}{B}}$$

The tail probability of finding b or more black socks in the first drawer is:

$$\mathrm{HGT}(N,B,n,b)=\sum_{i=b}^{\min(n,B)}\mathrm{HG}(N,B,n,i)$$

A1.2 mHG

In many scenarios a fixed partition of the set (e.g. first versus second drawer) is not known. If some ranking of the elements is given then we can consider all partitions that respect the given ranking—dividing the entire set of elements into a subset of high-ranking elements and a subset of low-ranking elements. We want to discover such partitions for which either of the subset is enriched with some attribute. Formally, consider a set of measurement values for elements $S=\{s_1,\ldots,s_N\}$, where $s_i<s_{i+1}$ and some binary labeling of the elements $\lambda=\lambda_1,\ldots,\lambda_N\in\{0,1\}^N$. The binary labels represent the attribute—say 1 for membership in a GO term and 0 otherwise. We define the mHG score as:

$$\mathrm{mHG}(\lambda)=\min_{1\leqslant n\leqslant N}\left(\mathrm{HGT}\left(N,B,n,b_n(\lambda)\right)\right)$$

where $b_n(\lambda)=\sum_{i=1}^{n}\lambda_i$.

In addition to the mHG score itself, it is also useful to note the rank n^* at which the minimal HGT was attained.

A variant of the mHG score is obtained when we limit the set of considered threshold to $1<n<n_{\max}$, where in practical uses $n_{\max}\ll N$.

A1.3 mHG p-values

It is important to note that the mHG score is not a p-value. To enable an accurate interpretation of the mHG score significance we employ an efficient dynamic programming procedure to fully characterize the distribution of mHG including exact p-values. We also employ effective bounds that can accelerate calculations. For details on the computational aspects and for applications to identifying transcription factor binding sites see Eden *et al.* (2007).

A1.4 Jump mHG heuristic

Subsequent to the description in Section 2.4 the following pseudo-code describes the naïve approach of finding the most enriched partition using mHGT:

```
BP-mHGT_all(T,G) =
    0: N = |G|
    1: p_min = 1
    2: for all partitions Q:
    3:         p = p_n(Q,T,G)
    4:         if (p < p_min)
    5:                 p_min = p
    6: return p_min
```

Where $p_n(Q,T,G)$ stands for the minimum HGT score over all possible thresholds up to n (= $n_{\max}$ the largest threshold under consideration).

$$\mathrm{mHG}(\lambda)=\min_{i\leqslant n\leqslant N}(\mathrm{HGT}(N,B,n,bn(\lambda))$$

The jump heuristic for mHGT is adjusted accordingly:

```
BP-mHGT_jump(T,G) =
    0: N = |G|
    1: R = random gene set of size r, where n<r<N
    2: G_R = RUT                        // the reduced universe
    3: p_min = 1
    4: for all partitions Q:
    5:         p = p_n(Q,T,G_R)
    6:         if (p < p_min)
    7:                 p = p_n(Q,T,G)
    8:                 R = the top r Q-diff-expressed genes
    9:                 G_R = RUT
    10:                if (p< p_min)
    11:                        p_min = p
    12: return p_min
```

Notice that in the reduce universe G_R, λ is padded with 0's:

$\lambda_R=\lambda_1,\ldots,\lambda_{|R|},\delta_1,..,\delta_{N-|R|}$ where $\lambda_i\in\{0,1\}$ and $\delta_i=0$.

CLAIM. $\mathrm{BP\text{-}mHGT}_{\mathrm{jump}}(T,G)=\mathrm{BP\text{-}mHGT}_{\mathrm{all}}(T,G)$

PROOF. Assume to the contrary that

(2) $\mathrm{BP\text{-}mHGT}_{\mathrm{jump}}(T,G)>\mathrm{BP\text{-}mHGT}_{\mathrm{all}}(T,G)$.

Let Q^* be the partition for which $\mathrm{BP\text{-}mHGT}_{\mathrm{all}}$ attains its maximum enrichment. Let $p^*=p_n(Q^*,T,G,N)$ and n^* be the threshold in which this enrichment was attained. Our assumption (2) implies that when considering Q^*, $\mathrm{BP\text{-}mHGT}_{\mathrm{jump}}$ did not reach Line 7, for which p is computed using G and p^* would have been found and replaced the then current $p_{\min}$.

It follows that there exists a set R for which $p_n(Q^*,T,G_R)$ (computed in Line 5) $>p_{\min}>p^*$. => $b_{n*}(Q^*,T,G_R)<b_{n*}(Q^*,T,G)=b^*$, in particular.

For this to happen, in the reduced universe G_R there need to be strictly more than (n^*-b^*) genes not from T that are in the top n^*Q^*-differentially expressed genes. Since $T\subset G_R\subset G$, in G there are strictly more than (n^*-b^*) genes not in T amongst the top n^*Q^*-differentially expressed genes (there can be new genes in this list, but not from T). This in turn means that $b_{n*}(Q^*,T,G)<b^*$. A contradiction. ■