Recovering Key Biological Constituents through Sparse Representation of Gene Expression

Journal:	*Bioinformatics*
Manuscript ID:	BIOINF-2010-1240.R2
Category:	Original Paper
Date Submitted by the Author:	27-Dec-2010
Complete List of Authors:	Prat, Yosef; The Hebrew University, School of Computer Science and Engineering Fromer, Menachem; The Hebrew University, School of Computer Science and Engineering Linial, Nathan; The Hebrew University, School of Computer Science and Engineering Linial, Michal; The Hebrew University, Biological Chemistry
Keywords:	Microarray data analysis, Gene ontology, Gene prediction, Data integration, Knowledge representation, Information extraction

Gene expression

Recovering Key Biological Constituents through Sparse Representation of Gene Expression

Yosef Prat[1], Menachem Fromer[1] Nathan Linial[1,2] and Michal Linial[3*]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

[2]Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Jerusalem, Israel

[3]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Large-scale RNA expression measurements are generating enormous quantities of data. During the last two decades, many methods were developed for extracting insights regarding the inter-relationships between genes from such data. The mathematical and computational perspectives that underlie these methods are usually algebraic or probabilistic.

Results: Here we introduce an unexplored geometric view point where expression levels of genes in multiple experiments are interpreted as vectors in a high-dimensional space. Specifically, we find, for the expression profile of each particular gene, its approximation as a linear combination of profiles of a few other genes. This method is inspired by recent developments in the realm of compressed sensing in the machine learning domain. To demonstrate the power of our approach in extracting valuable information from the expression data, we independently applied it to large-scale experiments carried out on the yeast and malaria parasite whole transcriptomes. The parameters extracted from the sparse reconstruction of the expression profiles, when fed to a supervised learning platform, were used to successfully predict the relationships between genes throughout the Gene Ontology (GO) hierarchy and protein-protein interaction map. Extensive assessment of the biological results shows high accuracy in both recovering known predictions and in yielding accurate predictions missing from the current databases. We suggest that the geometrical approach presented here is suitable for a broad range of high-dimensional experimental data.

Contact: michall@cc.huji.ac.il

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

1 INTRODUCTION

High-throughput technologies have come to play a central role in biological and biomedical research in the last decade. Advances in large-scale technologies on a genome-wide scale produce enormous amounts of data (Bader, *et al.*, 2004; Barrell, *et al.*, 2009; Beyer, *et al.*, 2007; Desiere, *et al.*, 2005). Yet, a major goal of functional genomics is the quest for a comprehensive description of the functions and interactions of all genes and proteins in a genome.

Data such as large-scale gene expression is usually represented by a matrix, where n genes are examined in d experimental conditions. Here, we view such data as a set of n points (vectors) in d-dimensional space, each of which represents the profile of a given gene over d different experimental conditions. Many known methods that have yielded meaningful biological insights in fact seek geometric or algebraic features of these vectors. For example, analyzing the angles between vectors amounts to a correlation-based analysis. Similarly, the direction in space along which these points are most "spread out" correspond to SVD (Alter, *et al.*, 2000) and its principal component analysis (PCA) implementation (Raychaudhuri, *et al.*, 2000; Yeung and Ruzzo, 2001). These are powerful tools in providing biological inference (Misra, *et al.*, 2002). In general, methods and disciplines developed toward extracting information from expression data include pairwise properties (e.g., correlation, variance, entropy-based distance) (Amato, *et al.*, 2006; Jeffery, *et al.*, 2006), clustering (Alon, *et al.*, 1999; Eisen, *et al.*, 1998), Bayesian networks (Friedman, *et al.*, 2000), information theory, ordinary differential equations and other sophisticated distance measures (reviewed in (Quackenbush, 2006; Slonim, 2002)).

In this study, we applied a different approach to gene expression data analysis. The geometric principle that underlies it is very natural and different from existing methods, though it is close in spirit, and inspired by, recent advances in compressive sensing and sparse signal recovery (Candes, 2008; Candes and Tao, 2005; Donoho, 2006). A simple probabilistic consideration implies the following geometric claim: given a set of n randomly chosen points in the d-dimensional space, it is "very unlikely" that a linear subspace Y exists where more than dim(Y) points of the chosen points reside "very close" to Y (see Methods).

In this study, we present a natural, yet unexplored, approach for the seemingly exhausted problem of gene expression analysis. Adopting a sparse signals reconstruction mindset, we recover a support set of genes for each gene in a genome. Geometrically, we uncovered linear subspaces which are over-populated with expression-profiles in the multidimensional space of the experiments set. We could verify the robustness and significance of the sparse reconstructions using measures intrinsic to the method and data. For-

*To whom correspondence should be addressed.

mally, we are interested in subsets S of our n-point set that (nearly) resides on a subspace of dimension strictly smaller than $|S|$. Having found such sets, several immediate questions suggest themselves: (i) Are these findings robust? (ii) If they are robust, can we directly interpret their biological meaning? (iii) Can such representation uncover meaningful structures? (iv) Does the method generalize? In this paper, we answer these questions by considering gene expression alone and testing data sets coming from the transcriptomes of the budding yeast *Saccharomyces cerevisiae* and the malaria parasite *Plasmodium falciparum*.

A conceptually new method that we call SPARCLE (SPArse ReCovery of Linear combinations of Expression) is introduced. It is inspired by the plausible assumption that expression data, when considered over a broad range of experimental conditions, encodes profound layers of systematic (yet hidden) behaviors. We further confirmed the stability and robustness of SPARCLE results for entire transcriptomes under perturbations to the data. Extracting features from the geometric parameters of SPARCLE's results, and training AdaBoost, a machine learning platform, to exhaustively reveal pairwise associations between gene function (represented by GO annotations and by the protein-protein interaction (PPI) map) confirmed the principal information encoded by the geometric-based representation. The generality of the method is confirmed by applying it to both the knowledge-rich yeast model and the poorly annotated malaria parasite proteome.

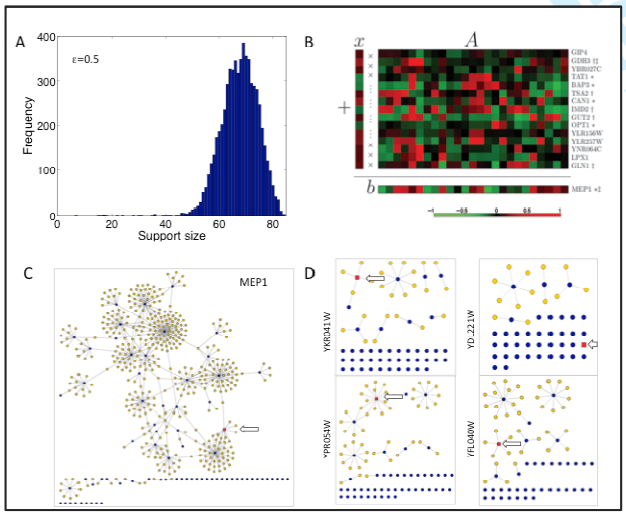

Fig. 1. Sparse reconstruction of yeast genes expression profiles by SPARCLE. (A) Support sizes of the solutions to the SPARCLE optimization problem (the number of genes used to reconstruct each particular gene), for all 6254 yeast genes analyzed. (B) The expression profile reconstruction for MEP1 (ammonium transporter) as recovered by SPARCLE. The expression profile of the gene (bottom) is displayed as a linear combination of the profiles of its supporting genes, with their corresponding coefficients (left). For comprehensibility, only the 15 genes with the largest absolute value coefficients are shown, as well as a third of the 85 conditions. *transmembrane transporters; †oxidation-reduction proteins; ‡ammonia-related processes. (C) Genes in the support of MEP1. The objective gene (MEP1) is indicated by an arrow. Note that the majority of the genes are part of a PPI network. (D) Sample of 4 objective genes (marked by arrows) whose supports are indicated by poor connectivity and a fragmented PPI network. PPI connectivity is retrieved from the BioGrid (http://thebiogrid.org/) repository. Graphics are based on Pathway Palette (Askenazi, *et al.*, 2010).

2 SPARSE REPRESENTATION OF EXPRESSION

We wish to discover linear dependencies within groups of expression profiles, using full transcriptome mRNA expression measured under a wide range of environmental conditions. Given an objective gene expression profile, one would seek, then, the smallest number of profiles, whose linear span contains the expression profile of the objective gene. Formally, this is expressed as the following problem:

$$(P_0) \quad \begin{array}{l} min\ \|x\|_0 \\ s.t.\ Ax=b \end{array}$$

Here $A \in \mathbf{R}^{d\times n}$ is a matrix of RNA expression levels of n genes (the entire genome excluding the objective gene) measured in d different experiments, $b \in \mathbf{R}^d$ is the vector of expression levels of the objective gene in the d experiments, and $x \in \mathbf{R}^n$ are the n optimization variables, which are n coefficients corresponding to the n genes in the genome. The $\|x\|_0$ notation stands for the L_0 "norm" of x, which is the number of non-zero entries in x (See example in Fig. 1, A and B). We should note here that we consider the common situation where n is much larger than d, hence $Ax=b$ is an underdetermined system of linear equations. In its general form, this optimization problem is NP-hard (Natarajan, 1995). Fortunately, theoretical developments in recent years imply that this problem can be efficiently solved in practice, or at least approximated well, in many practical cases. The theory developed around this problem (Candes and Tao, 2005; Donoho, 2006; Rudelson and Vershynin, 2008) shows that for generic instances of this problem, the solution of P_0 coincides, at least nearly, with the solution of the following problem:

$$(P_1) \quad \begin{array}{l} min\ \|x\|_1 \\ s.t.\ Ax=b \end{array}$$

The advantage is that P_1, where the L_0 "norm" has been replaced by the L_1 norm, can be stated as a linear programming problem and is hence efficiently solvable. In order to apply this method to noisy biological data, we use a relaxed form of P_1:

$$(P_\varepsilon) \quad \begin{array}{l} min\ \|x\|_1 \\ s.t.\ \|Ax-b\|_1 \le \varepsilon \end{array}$$

Where ε is a sufficiently small noise parameter. We use a linear programming solver to solve this optimization problem, for each gene in the dataset as an objective gene in its turn. This is followed by an intrinsic assessment of robustness. We refer to this combined procedure as SPARCLE.

3 METHODS

3.1 Datasets

Gene expression measurements were extracted from the GEO database GSE11452 (Knijnenburg, *et al.*, 2009) and consist of a microarray compendium of 170 steady-state chemostat cultures of *S. cerevisiae*, which encompass 55 unique environmental conditions. The full data consists of 9335 Affymetrix probes, representing the full *S. cerevisiae* transcriptome. We used a set of 6254 genes, after elimination of most non-coding transcripts including transposons, tRNAs, and rRNAs, and selecting one probe for each coding gene. The same filters were applied to GEO database GSE19468 of the malaria parasite *P. falciparum*. We used a set of 208

microarray experiments that cover 4365 genes from *P. falciparum* (Hu, *et al.*, 2009).

3.2 Solving the SPARCLE optimization problem

The expression data set was divided into two sets of experiments, where one was used for the unsupervised learning of sparse representations, and the other was left aside for a cross-validation test of robustness. The problem was solved using the matrix A of 85 (experiments) × 6254 (genes), for *S. cerevisiae*, and 104 (experiments) × 4365 (genes), for *P. falciparum*. Repeatedly, each column (85, or 104, coordinate gene expression profile) is chosen as b in (P_ε) and is removed (for this single iteration) from the matrix A. The optimization problem was solved as a linear programming problem using Matlab's linprog solver. The noise parameter ε in (P_ε) was set to 0.5 (Fig. S1). The noise was evaluated using the L_1 norm, permitting an efficient linear programming description. Random partitions of the data into learning and test sets (5 repetitions) resulted in almost identical outcome, verifying the independence of the results on the specific partitions chosen.

3.3 Robustness of expression profile representations

Biological robustness and validity of the solutions were measured by their degree of approximation in the unseen data of experiments. Specifically, we denote by A' the unobserved matrix excluding the objective gene, and b' as the objective gene's d-dimensional expression profile in the unseen data. The solution of the minimization problem using the first matrix (A) is x*. We then take $\varepsilon'=\|A'x^*-b'\|_1$ as the degree of approximation on the unseen data. When ε' is small, the solution may be considered as biologically robust, since the linear combination it describes holds true for a set of biological experiments not utilized by SPARCLE. In order to assess the quality of ε', we performed two different tests. In the first one we chose a random support set for each gene, of the same size as the support chosen by SPARCLE and calculated coefficients for each support member by solving: $min_x \|Ax-b\|_1$ where b is the objective gene's profile, and x is a vector of all zeroes but at the support's coordinates. Then, using the solution x, we evaluated ε' as before; repeating 10,000 times, we estimated the background distribution of the ε' value, resulting in a p-value for each ε' value.
In the second test, we randomly select d genes, reducing the matrix A to contain only these d genes, which produced $\hat{A} \in \mathbf{R}^{d\times d}$ and solve:

$$(P_{\hat{A}}) \quad \begin{array}{l} min\ \|x\|_1 \\ s.t.\ \|\hat{A}x-b\|_1 \le 0.5 \end{array}$$

For each gene, we obtained x and calculated $\varepsilon'=\|A'x-b'\|_1$. The choice of d genes was done in order to ensure the existence of a feasible solution in the optimization problem (as the biological data is noisy, we assume both matrices A and $\hat{A}$ have rank d); repeating this process 1000 times allows estimation of the corresponding p-value.

3.4 Normalization and setting the noise measure

The raw expression data were normalized in two ways: (i) the expression profile for each gene was divided by its maximal value (ii) for each experiment/condition, the mean expression value across the entire set of genes was subtracted from each gene. We further added a column (i.e., a new "gene") with a constant expression value of 1, and gave it a zero weight in the minimization problem; this step permitted the free use of a constant factor in the linear combinations found. We tested several values for the noise factor ε. Clearly, a larger ε yields sparser solutions (as the constraints of the optimization problem are relaxed) but with a less accurate reconstruction of the objective gene. On the other hand, tighter constraints of smaller ε values result in over-fitting to the noise in the train data. In this paper, we describe results obtained using ε=0.5. The ε value was selected to be less than 5% of the mean L_1 norm of the normalized profiles, and such that it will never exceed 20% of any profile's L_1 norm.
The assessments and influence on support sizes of using different values of ε=0.25, ε=0.75 are shown in Fig. S1.

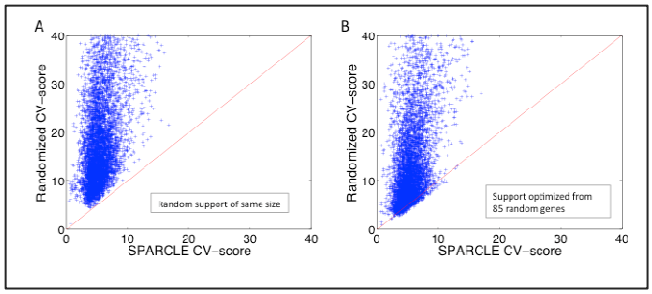

Fig. 2. Cross-validation tests for SPARCLE robustness (A) Comparison of the cross-validation (CV) scores for each reconstructed support for an expression profile with the score obtained for a random support of the same size; note that lower scores correspond to more robust predictive power. (B) Comparison of the CV scores for each reconstructed support for an expression profile with the score obtained by a restricted SPARCLE run over 85 random profiles (see Methods). The SPARCLE results are consistently better than random. For the first test (A) all 6254 results received p-value < 10^{-6}, for the second test (B) 4633 results received p-value < 10^{-6}, and another 445 received p-value < 0.05.

3.5 High-dimensional geometric analysis

We enhanced the mathematical findings of SPARCLE by direct geometric analysis of the raw input data. As mentioned above, we view each expression vector as a point in d-dimensional space. We analyzed the geometric properties of the data by investigating the convex hull of this set of vectors. This information was used to quantify the deviation of the expression vectors of genes from those of others. These quantities were included as features in leveraging the follow-up supervised learning of biological associations between genes.

3.6 Measuring GO enrichment

For a given set of support genes found by SPARCLE to reconstruct an objective gene, GO enrichment was calculated using a hypergeometric test, with the entire set as a background (Barrell, *et al.*, 2009). Sets were considered enriched with an annotation if the annotation received a p-value <0.05, corrected for a False Detection Rate (FDR) of 5%. Hypergeometric probabilities and FDR were computed directly using Matlab.

3.7 Extraction of feature vectors

The sparse representations found by SPARCLE were condensed into feature vectors for each pair of genes. These vectors contained both individual features of each member of the pair and pairwise features. Importantly, all the features were extracted from the input data (e.g., correlations, high-dimensional geometric analysis), the output solutions of SPARCLE (e.g., support sizes, mutual coefficient values), and their intrinsic assessment values (e.g., ε'); no external features were used. These feature vectors were used in a supervised learning platform in order to assess the significance of our results.
The following features were extracted from SPARCLE results and the raw data. They comprise a vector with 40 parameters for each pair of genes, which was used for the supervised learning. The features (for a pair of genes i and j) are: (a) Coefficient of each gene in the expression profile of the other, as reconstructed by SPARCLE (non-zero if gene i is in the se-

lected support for gene *j*, and vice versa). (*b*) The number of genes in the intersection of the two supports as recovered by SPARCLE. (*c*) The number of supports containing both of the two genes. (*d*) The L_1 distance of each gene's expression profile from the convex hull of the other genes' vectors, as recovered by the high-dimensional geometric analysis (Section 3.5). (*e*) The Euclidean distance of the expression profile of gene *i* from the subspace spanned by gene *j*'s supporting profiles, and vice-versa. (*f*) Support size for each gene. (*g*) Number of appearances in other supports for each gene. (*h*) Average and standard deviation for features (*a*)-(*g*) over 20 perturbation runs of SPARCLE on the same data (where 25% of the genes were randomly removed each time). (*i*) Pearson correlation between *i*'s and *j*'s expression profiles, for both the normalized and unnormalized expression data. (*j*) For each gene, the mean, median, and standard deviation of feature (*i*) over the entire set.

All listed features (*a*)-(*j*) were used in the supervised learning of shared GO annotations and PPI by the AdaBoost algorithm. To test the principal information from SPARCLE, we separated features (i), (j) for a direct evaluation of the contribution of features that can be extracted directly from the raw data. We denoted the analysis based on AdaBoost using features (i), (j) collectively as Correlations+AB (Fig. 4, Figs. S3-S7).

3.8 Prediction of gene associations

The Gene Ontology (GO) is structured as three directed acyclic graphs (DAG): the cellular component (CC), the biological process (BP), and the molecular function (MF) ontology. Each term, used to annotate genes, resides at a different depth with respect to its root (CC, BP, or MF). The deeper the term resides in the graph, the higher its annotation resolution, i.e., it is more specific (as illustrated in Fig. S2). In order to label two genes as associated by similar GO terms, one should first choose the resolution of interest. We choose to measure the depth of the term as the length of the shortest path from the root to the term in the DAG. We tested our predictions both at low resolution (close to the root) and at high resolution (deep in the GO structure, i.e., specific annotations). The low-resolution depth was chosen as the lowest level of description where less than 50% of the gene pairs would be considered as associated with a GO term (depths 5, 2, and 1 for CC, BP, and MF, respectively for the yeast data, and depths 3, 1, and 1 for the malaria parasite data). The high-resolution depth was chosen as the highest level of description where at least 1% of the gene pairs would be assigned the same annotation (depths 11, 8, and 7 for the yeast, and depths 7, 5, and 5 for CC, BP, and MF, respectively for the malaria parasite). In addition to using the depth measure for resolution, we also applied the GO-Slim (Barrell, *et al.*, 2009) set of manually selected GO terms, constructed to eliminate the hierarchical structure of GO.

3.9 Interpreting the results of supervised learning

We trained the AdaBoost method (Freund and Schapire, 1997) to classify the feature vectors as positive (i.e., same GO annotation) or negative for biological association. The training set included 15,000 randomly selected pairs, half positively and half negatively labeled. The test set contained 200,000 randomly selected pairs that were not used in the training set, again half positively and half negatively labeled. We applied a simple threshold on the AdaBoost raw classification values in order to assign confidence values to its classifications. The confidence level granted a tradeoff between coverage and accuracy. In essence, this requires higher confidence in making any classification at all, hence refusing to classify some of the examples. In order to obtain x% coverage, we ignore all but the x% highest positive classification values, and x% lowest negative values.

3.10 Comparing predictions

We compared SPARCLE-based learning by AdaBoost to three other methods of predicting associations among genes. First, we used AdaBoost to learn associations using only correlation-related features. Second, we used the correlation-based transitive shortest path (SPath) evaluation method (Zhou, *et al.*, 2002). Briefly, an undirected graph is constructed, with genes as nodes, and edge weights 1-*P*, where *P* = the Pearson correlation between the pair (for $P \geq 0.6$). A shortest path was then constructed between each pair, and its weight was used as an estimator for a distance between the genes. Lastly, we used the absolute value of the Pearson correlation between genes as a measure of their association, applying a confidence level.

3.11 Inspection SPARCLE-based predictions

We chose to manually test the possibility that the false predictions are due to incomplete labeling of gene products by GO annotations. To this end, we sampled a set of 10 predicted associations (gene pairs) from the yeast data, which were not annotated as being associated (false positives), and compared them with a random sample of 10 pairs predicted as not associated, conforming to GO annotation (true negatives). This process was done for all three GO sub-ontologies (CC, BP, and MF); hence, 60 pairs were manually investigated (Table S3). For each pair, a shared annotation (if found) was retrieved from a literature based association protocol (Jenssen, *et al.*, 2001). Further analysis included the use of PPI networks based on the BioGrid (Stark, *et al.*, 2006) and STRING (von Mering, *et al.*, 2003) experimental data servers. When the servers found an association, they also returned a p-value for the connection. The minimal number of intermediate nodes connecting a pair of genes in the network was retrieved using Pathway Palette (Askenazi, *et al.*, 2010).

4 RESULTS

To demonstrate the utility of SPARCLE on gene expression data, we analyzed two very large experimental data sets: from the yeast *S. cerevisiae*, and from the malaria parasite *P. falciparum* comprised of 170 and 208 experiments, and covering 6,254 and 4,365 genes, respectively. While the SPARCLE methodology is not restricted by the type or source of data, we used mRNA expression measurements from (Knijnenburg, *et al.*, 2009), which constitute a microarray compendium of chemostat cultures of *S. cerevisiae* that cover 55 unique growth conditions, including nutrient-limiting substrates, growth rate, aeration, pH, and temperature. This data set was divided randomly into two equal-sized sets of d=85 experiments covering n=6254 yeast genes. Our matrix has full row rank d=85 and linear algebra implies that the smallest support (of a solution to P_0) will never exceed d. Indeed, the coefficient vectors obtained were considerably sparser with an average support size of 67 (Fig. 1A). Thus our goal of achieving a 'short' compact linear representation is achieved. To ensure robustness, half of the experiments (85) were not used for such representation, and were reserved for the purpose of cross-validation and evaluation. Random partitions of the data into two parts were performed 5 times with essentially identical results (see Methods). Following this new geometrical representation of the data and confirming its stability to perturbations (Fig. 2), we turned to extracting valuable biological information for the entire proteomes.

The first functional test was based on searching enrichment in GO (Barrell, *et al.*, 2009) annotations. For 10% of the genes, significant enrichment of functional annotation could be found among their set of supporting genes retrieved by SPARCLE. An example is the gene MEP1 (Fig. 1B) for which many of the support members share annotations (Table S1). The statistical enrichments of GO annotations for a sample of gene supports are shown (Table

S2). Furthermore, MEP1 is interconnected with several of the support gene products, as reflected by the connected graph of the PPI network (Fig. 1C). However, for most genes (90%), an immediate biological interpretation could not be retrieved from the support set. Typically, the objective gene and its support gene products are isolated in a PPI network graph (examples are shown in Fig. 1D).

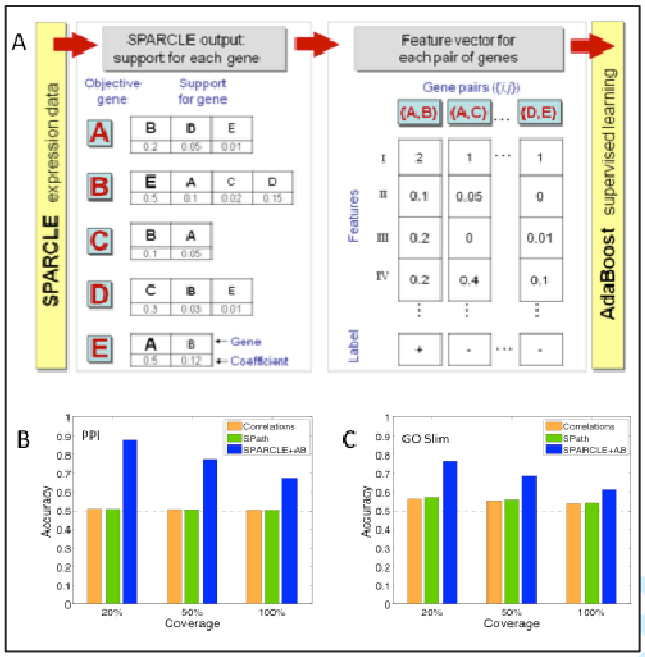

Fig. 3. Prediction of PPI and GO annotations. (A) Illustration of feature extraction for pairs of genes from SPARCLE. Each sparse representation includes a set of genes and their assigned coefficient. For each pair of genes, a feature vector was constructed from the properties of their representing sets. The feature vector also included another high-dimensional analysis, i.e., distances of each profile from the convex hull of the others. Other features were obtained directly from the input data (see Methods). Features in the illustration: I – co-occurrence in supports, II – gene *i*'s coefficient in gene *j*'s support, III – gene *j*'s coefficient in gene *i*'s support, IV – Pearson correlation of the expression profiles. (B) Prediction of PPI, as represented by the STRING database, by supervised learning from SPARCLE results (SPARCLE+AB). Accuracy is traded off with coverage by applying certainty thresholds on the classifier output. Other methods for predicting genes interrelationships are: Pearson correlation of the expression profiles (Correlations), and a transitive correlations method (SPath, see Methods). (C) Prediction of associations for the GO Slim annotations, covering cellular component ontology. For detailed analyses of accuracy-coverage tradeoff see Fig. S5 (GO slim) and Fig. S5 (PPI).

As SPARCLE results proved meaningful and robust by the cross-validation test (Fig. 2, Fig. S1), we expect the method to capture hidden information. To this end, we used SPARCLE results as input for a machine learning procedure (Fig. 3A). Specifically, we trained the AdaBoost framework (Freund and Schapire, 1997) to classify whether each pair of genes has a reported protein-protein interaction or not, using information that is only extracted from the input data itself (i.e., the expression matrix) and the SPARCLE analysis (see Methods). Together, the results of SPARCLE, with the input expression data, were condensed into feature vectors for each pair of genes (Fig. 3A).

We tested whether functional information that is encoded in the yeast PPI map can be successfully recovered. Using a confidence threshold for the classification, accurate performance can be traded off in exchange for providing lower coverage of the data. The results of the supervised learning were exceptionally good (Fig. 3B). For 50% coverage of the high confidence predictions, an accuracy of 78% was reached. Even for 100% coverage, the accuracy reaches 70% (Fig. 3B). Recall that the yeast unfiltered PPI map still exhibits a high false positive rate (FP) (Wu, *et al.*, 2006). The combined protocol of the unsupervised SPARCLE method and supervised learning platform (based on SPARCLE feature vector, Fig. 3A) was then tested for the task of recovering the GO associations between genes, with the three functional branches covering molecular function (MF), cellular component (CC), and biological process (BP) (Fig. 3C). Specifically, gene pairs were classified as sharing, or not sharing, similar GO annotations.

For comparison, we compare the prediction results to other correlation-based methods (Figs. 3B, 3C). While the GO hierarchical database covers different descriptive resolutions (Fig. S2), our protocol exhibited accurate predictions at all resolution levels (Fig. S3-S5). For example, with 20% coverage at high GO resolution the accuracy reached 97.6%, 91%, and 99% for CC, BP, and MF, respectively (SPARCLE+AB, Figs. 4A-4C and Figs. S3-S5). For full coverage, we still achieved 65-72% accuracy for all ontology branches at low resolution (SPARCLE+AB, Figs. 4A-4C), and 73-89% for the more specific terms of the high resolution of GO annotations (SPARCLE+AB, Figs. S3-S5). An additional perspective on the SPARCLE+AB method is retrieved from the tradeoff of sensitivity and 1-specificity as presented by the ROC (receiver operating characteristic) curves. In all tests (for PPI, GO low and high levels and GO Slim) when compared SPARCLE+AB and Correlation+AB, a higher sensitivity is measured for the same specificity (not shown).

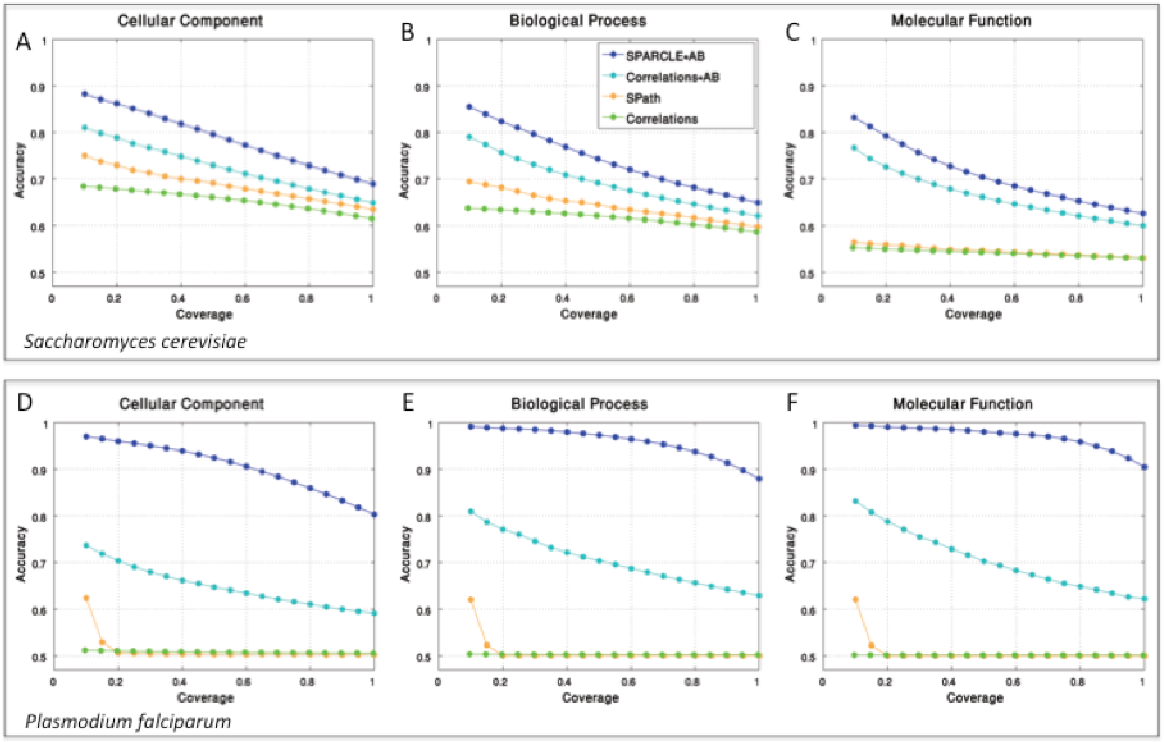

Fig. 4. Prediction of genes' associations according to GO, where accuracy is defined as in Fig. 3. A comparison of SPARCLE-based AdaBoost learning (SPARCLE+AB), correlation-based AdaBoost learning (Correlations+AB), correlations-based shortest path (SPath) (Zhou, *et al.*, 2002), and pairwise correlations for the raw data (Correlations) for *S. cerevisiae* (A-C) and *P. falciparum* (D-F) transcriptomes. The ontology branches CC (A,D), BP (B,E) and MF (C,F) were examined. A detailed analysis for all GO resolution levels is shown for *S. cerevisiae* (Figs. S3-S5) and *P. falciparum* in Fig. S7.

Next, we tested whether our inference method "happens" to do well on the yeast as a model system. Indeed, the yeast genome is extremely rich in annotations and currently 88% of its genes are associated with some informative GO annotation. Similarly, the quality and density of the yeast interactome exceed those of any other model system. We thus repeated the entire protocol for a set of 208 experiments (Hu, *et al.*, 2009) measuring 4365 *P. falciparum* genes expression levels, from cells exposed to ~30 anti-malaria drugs. Note that only 5% of the malaria genes are reviewed by SwissProt, 65% of the proteins are annotated as 'putative' and only 46% of the genes are associated with some GO annotations (often at a low resolution, Fig. S7). The SPARCLE-based protocol again demonstrated high predictive power (Figs. 4D-4F, Fig. S7). Lastly, we systematically tested the novel knowledge gained from the above-described protocols (Figs. 3-4, S3-S5). To this end, we randomly sampled pairs of yeast genes which were annotated as unrelated and yet which we predicted to be related (false positives, FP) and, for comparison, pairs of genes which were annotated as unrelated and predicted to be unrelated (true negatives, TN). We manually examined each such pair of genes for functional connections. Remarkably, we verified our predictions for interrelations in ~80% of all FP samples, yet could only detect relations in about a third of the TN set (Table S3). While this manual inspection cannot be considered to stand on solid statistical ground, it provides support for the relevance of SPARCLE based properties, when they are fed into a machine-learning platform to empower functional inference.

5 DISCUSSION

The value of the information retrieved by the SPARCLE approach was demonstrated by using its results as a basis for machine learning classification of gene associations. A systematic and comprehensive evaluation, ranging from PPI networks and going through all resolution levels of the GO annotation database, covering the immensely explored yeast transcriptome and the poorly annotated malaria-parasite genome, revealed the large potential of using such a poorly studied geometric approach to extract principal insights from gene expression data.

Many approaches aim to develop a systematic way to unravel hidden structure in data. Most studies that looked for biological coherence in gene expression data applied clustering (at different levels of sophistication), revealing the existence of some hidden 'structure' in the data. In the current research, comparisons to clustering results were not carried out, as our goal here is quite different. The high performance of SPARCLE-based AdaBoost learning should be considered as evidence for the principal information that is embedded in the geometric properties of the data. Therefore, a critical comparison was performed to evaluate the information that is embedded in correlation (a form of geometric representation, see below). We show that the correlation performed very poorly on the malaria data and somewhat better on the yeast data. In addition, by combining the AdaBoost learning protocol with the correlation (Correlation+AB), we isolated the contribution of the AdaBoost

learning itself. SPARCLE+AB outperformed these other approaches for the entire range of accuracy and coverage (Figs. 3,4 and Fig. S3-S7).

Several aspects of our approach differ from common practices, and should be elaborated. Most of the activity in the machine-learning area can be viewed as a modern-day approach to the classical questions of statistics. The data at hand is considered as being sampled from some distribution and the question is to get as accurate as possible a description of that distribution. Our approach is different.

When data items are (or can be naturally viewed as) points in space, it is possible to utilize any "unexpected" geometric properties that this set of points (corresponding to data items) has. In fact, many successful existing methods in machine learning can be viewed from this perspective. Thus, if S is a generic set of N points in d-dimensional space and if N is sub-exponential in d, then we do not expect to see any pairs of points (even nearly) in the same direction from the origin. If the set of points that is your data set violates this statement, you can conclude that it has a geometrically non-trivial structure. This structural property is very likely a reflection of an interesting (albeit not necessarily interpretable) property in the domain from which the data set came. This is our interpretation of correlation analysis, one of the most reliable workhorses of bioinformatics. Likewise, a generic point set in Euclidean space is not expected to be stretched in any special directions in space. Therefore if your data set, viewed geometrically, is stretched in certain directions, it tells you something, which can often be used to discover interesting phenomena, this is our interpretation of SVD analysis.

Correlations and stretch are only two of the numerous properties that one may consider in a point set in Euclidean space. Our work considers another very basic property that we know not to exist in generic sets: (Nearly) linearly-dependent sets of points of cardinality that is substantially smaller than the dimension of the host space. When such an unexpected property of the data set is discovered, two questions suggest themselves: (i) Is this phenomenon only coincidental? and (ii) How can this geometric property of the data help us learn something about the system which it represents? In this study we confirm the robustness of this property under multiple perturbations (Figs. 1-2, Fig. S1) and the generality for multiple model organisms (Figs 3-4, Fig. S3-S7). The SPARCLE based machine-learning analysis is a first step toward a deeper understanding of the underlying complexity of the biological gene associations.

In this study, we present a natural, yet unexplored, approach for the seemingly exhausted problem of gene expression analysis. Adopting a sparse signals reconstruction mindset, we recover a support set of genes for each gene in a genome. Geometrically, we uncovered linear subspaces which are over-populated with expression-profiles in the multidimensional space of the experiments set. We could verify the robustness and significance of the sparse reconstructions using measures intrinsic to the method and data.

A notable byproduct of the process is the observation that a biological interpretation of the support sets was mostly indirect. This is to be expected, since we only consider the smallest support size for each given vector while often many other representations of the same vector can be found with sub-dimensional supports. Another offshoot is the partial ability to identify unannotated genes, which somewhat contributed to the high precision in the case of the *P. falciparum* study. Such genes are mostly evolutionary branch-specific genes, and identifying them from expression data is stimulating in and of itself.

Funding: This work was supported by EU Framework VII Prospects consortium and a grant from ISF 592/07. YP and MF are supported by the Sudarsky Center for Computational Biology.

Conflict of Interest: none declared.

REFERENCES

Alon, U., *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc Natl Acad Sci U S A*, **96**, 6745-6750.

Alter, O., *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling, *Proc Natl Acad Sci U S A*, **97**, 10101-10106.

Amato, R., *et al.* (2006) A multi-step approach to time series analysis and gene expression clustering, *Bioinformatics*, **22**, 589-596.

Askenazi, M., *et al.* (2010) Pathway Palette: A rich internet application for peptide-, protein- and network-oriented analysis of MS data, *Proteomics*.

Bader, J.S., *et al.* (2004) Gaining confidence in high-throughput protein interaction networks, *Nat Biotechnol*, **22**, 78-85.

Barrell, D., *et al.* (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource, *Nucleic Acids Res*, **37**, D396-403.

Beyer, A., *et al.* (2007) Integrating physical and genetic maps: from genomes to interaction networks, *Nat Rev Genet*, **8**, 699-710.

Candes, E. (2008) The restricted isometry property and its implications for compressed sensing, *Comptes Rendus Mathematique*, **346**, 589-592.

Candes, E.J. and Tao, T. (2005) Decoding by linear programming, *IEEE Transactions on Information Theory*, **51**, 4203-4215.

Desiere, F., *et al.* (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry, *Genome Biol*, **6**, R9.

Donoho, D.L. (2006) For most large underdetermined systems of linear equations the minimal l(1)-norm solution is also the sparsest solution, *Communications on Pure and Applied Mathematics*, **59**, 797-829.

Eisen, M.B., *et al.* (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A*, **95**, 14863-14868.

Freund, Y. and Schapire, R. (1997) A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**, 119-139.

Friedman, N., *et al.* (2000) Using Bayesian networks to analyze expression data, *J Comput Biol*, **7**, 601-620.

Hu, G., *et al.* (2009) Transcriptional profiling of growth perturbations of the human malaria parasite Plasmodium falciparum, *Nat Biotechnol*, **28**, 91-98.

Jeffery, I.B., *et al.* (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data, *BMC Bioinformatics*, **7**, 359.

Jenssen, T.K., *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression, *Nat Genet*, **28**, 21-28.

Knijnenburg, T.A., *et al.* (2009) Combinatorial effects of environmental parameters on transcriptional regulation in Saccharomyces cerevisiae: A quantitative analysis of a compendium of chemostat-based transcriptome data, *Bmc Genomics*, **10**, -.

Misra, J., *et al.* (2002) Interactive exploration of microarray gene expression patterns in a reduced dimensional space, *Genome Res*, **12**, 1112-1120.

Natarajan, B.K. (1995) Sparse Approximate Solutions to Linear-Systems, *Siam Journal on Computing*, **24**, 227-234.

Quackenbush, J. (2006) Weighing our measures of gene expression, *Mol Syst Biol*, **2**, 63.

Raychaudhuri, S., *et al.* (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series, *Pac Symp Biocomput*, 455-466.

Rudelson, M. and Vershynin, R. (2008) On sparse reconstruction from Fourier and Gaussian measurements, *Communications on Pure and Applied Mathematics*, **61**, 1025-1045.

Slonim, D.K. (2002) From patterns to pathways: gene expression data analysis comes of age, *Nature Genetics*, **32**, 502-508.

Stark, C., *et al.* (2006) BioGRID: a general repository for interaction datasets, *Nucleic Acids Res*, **34**, D535-539.

von Mering, C., *et al.* (2003) STRING: a database of predicted functional associations between proteins, *Nucleic Acids Res*, **31**, 258-261.

Wu, X., *et al.* (2006) Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations, *Nucleic Acids Res*, **34**, 2137-2150.

Yeung, K.Y. and Ruzzo, W.L. (2001) Principal component analysis for clustering gene expression data, *Bioinformatics*, **17**, 763-774.

Zhou, X., *et al.* (2002) Transitive functional annotation by shortest-path analysis of gene expression data, *Proc Natl Acad Sci U S A*, **99**, 12783-12788.