Gene expression

Recovering key biological constituents through sparse representation of gene expression

Yosef Prat\textsuperscript{1}, Menachem Fromer\textsuperscript{1}, Nathan Linial\textsuperscript{1,2,\ast} and Michal Linial\textsuperscript{3,\ast}

\textsuperscript{1}School of Computer Science and Engineering, \textsuperscript{2}Sudarsky Center for Computational Biology and \textsuperscript{3}Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

Associate Editor: David Rocke

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 interrelationships 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 viewpoint 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 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.

Received on August 18, 2010; revised on December 27, 2010; accepted on January 1, 2011

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 (Badet 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 are 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 singular value decomposition (SVD) (Alter et al., 2000) and its principal component analysis 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) and 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 Section 3).

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 that are overpopulated 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. Formally, 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
article, we answer these questions by considering gene expression alone and testing datasets 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 Gene Ontology (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.

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:

$$\min_{x} \|x\|_0$$

subject to

$$Ax = b$$

Here $A \in \mathbb{R}^{d \times n}$ is a matrix of RNA expression levels of $n$ genes (the entire genome excluding the objective gene) measured in $d$ different experiments, $b \in \mathbb{R}^d$ is the vector of expression levels of the objective gene in the $d$ experiments, and $x \in \mathbb{R}^n$ 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. 1A 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:

$$\min_{x} \|x\|_1$$

subject to

$$Ax = b$$

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 $P_1$:

$$\min_{x} \|x\|_1$$

subject to

$$Ax = b, \quad ||Ax - b||_1 \leq \epsilon$$

where $\epsilon$ 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 (Klijn 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 dataset 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
We enhanced the mathematical findings of SPARCLE by direct geometric analysis of gene expression data. 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 \( (A) = s^* \). We then take \( \epsilon = ||A s^* - b||_1 \) as the degree of approximation on the unseen data. When \( \epsilon \) 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 \( s^* \), 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 \( \epsilon \) as before; repeating 10000 times, we estimated the background distribution of the \( \epsilon \) value, resulting in a \( P \)-value for each \( s^* \) value.

In the second test, we randomly select \( d \) genes, reducing the matrix \( A \) to contain only these \( d \) genes, which produced \( A \in \mathbb{R}^{d\times d} \) and solve:

\[
\begin{align*}
(P_{\epsilon}) \min_{x} & ||Ax - b||_1 \\
\text{s.t.} & ||Ax - b||_1 \leq \epsilon \\
\end{align*}
\]

For each gene, we obtained \( s \) and calculated \( \epsilon = ||A s - b||_1 \). 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 \( A \) 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 and (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 \( (\text{e.g. 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 \( \epsilon \). Clearly, a larger \( \epsilon \) 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 \( \epsilon \) values result in overfitting to the noise in the train data. In this article, we describe results obtained using \( \epsilon = 0.5 \). The \( \epsilon \) value was selected to be \( \leq 5\% \) 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 \( \epsilon \) are shown in Supplementary Figure S1.

### 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 at 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 (Barrett et al., 2009). Sets were considered enriched with an annotated \( P \)-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. \( x^* \)); 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 as follows: (i) coefficient of each gene in the expression profile of the other, as reconstructed by SPARCLE (non-zero if gene \( i \) is in the selected support for gene \( j \) and vice versa). (ii) The number of genes in the intersection of the two supports as recovered by SPARCLE. (iii) The number of supports containing both of the two genes. (iv) 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). (v) The Euclidean distance of the expression profile of gene \( i \) from the subspace spanned by gene \( j \)’s supporting profiles and vice-versa. (vi) Support size for each gene. (vii) Number of appearances in other supports for each gene. (viii) Average and SD for features (i)-(vii) over 20 perturbation runs of SPARCLE on the same data (where 25% of the genes were randomly removed each time). (ix) Pearson’s correlation between \( i \)’s and \( j \)’s expression profiles, for both the normalized and unnormalized expression data. (x) For each gene, the mean, median and SD of feature (ix) over the entire set. All listed features (i)-(ix) 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 (ix), (x) for a direct evaluation of the contribution of features that can be extracted directly from the raw data. We defined the analysis based on AdaBoost using features (ix), (x), collectively as Correlations+AB (Fig. 4, Figs S3–S7).

#### 3.8 Prediction of gene associations

The 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 Supplementary 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 \( \leq 50\% \) 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
We compared SPARCLE-based learning by AdaBoost to three other methods. We trained the AdaBoost method (Freund and Schapire, 1997) to classify retrieved using Pathway Palette (Askenazi et al., 2001). Further analysis included the use of PPI correlation-based transitive shortest path (SPath) evaluation method (Zhou et al., 2009). 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's correlation between genes as a distance measure.

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 $P$ is the Pearson's 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's 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 (TNs)). This process was done for all three GO subontologies (CC, BP and MF); hence, 60 pairs were manually investigated (Supplementary Table S3). The statistical enrichments 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 enrichment of functional annotation.

4 RESULTS

To demonstrate the utility of SPARCLE on gene expression data, we analyzed two very large experimental datasets: from the yeast S.cerevisiae and from the malaria parasite P.falciparum composed of 170 and 208 experiments and covering 6254 and 4365 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 dataset 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 $P_1$) 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 five times with essentially identical results (see Section 3). 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 (Supplementary Table S1). The statistical enrichments of GO annotations for a sample of genes supports are shown (Supplementary 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).
covering CC ontology. For detailed analyses of accuracy coverage tradeoff, Section 2). Profiles (Correlations), and a transitive correlations method (SPath, see Section 2). (C) Prediction of associations for the GO Slim annotations, covering CC ontology. For detailed analyses of accuracy coverage tradeoff, see Supplementary Figure S5 (GO slim) and Figure S5 (PPi).

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 (FP) rate (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 MF, CC and 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 (Fig. 3B and C). While the GO hierarchical database covers different descriptive resolutions (Supplementary Fig. S2), our protocol exhibited accurate predictions at all resolution levels (Supplementary Figs 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; Fig. 4A–C and Supplementary Figs S3–S5). For 100% coverage, we still achieved 65–72% accuracy for all ontology branches at low resolution (SPARCLE+AB; Fig. 4A–C), and 73–80% for the more specific terms of the high resolution of GO annotations (SPARCLE+AB; Supplementary 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 receiver operating characteristic curves. In all tests (for PPI, GO low and high levels and GO Slim), when SPARCLE+AB and Correlation+AB are compared a higher sensitivity is measured for the same specificity (data not shown).

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 \textit{P.falciparum} genes expression levels, from cells exposed to \sim30 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, Supplementary Fig. S7). The SPARCLE-based protocol again demonstrated high predictive power (Fig. 4D–F; Supplementary Fig. S7). Finally, we systematically tested the novel knowledge gained from the above-described protocols (Figs 3 and 4; Supplementary Figs 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 (FP) and, for comparison, pairs of genes which were annotated as unrelated and predicted to be unrelated (TNs). We manually examined each such pair of genes for functional connections. Remarkably, we verified our predictions for interrelations in \sim80% of all FP samples, yet could only detect relations in about a third of the TN set (Supplementary 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 and 4; Supplementary Figs 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 subexponential 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 dataset 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 dataset 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 dataset, viewed geometrically, is stretched in certain directions it tells you something that 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 dataset 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 and 2; Supplementary Fig. S1) and the generality for multiple model organisms (Figs 3 and 4; Supplementary Figs 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 that are overpopulated 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 subdimensional 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 (ISF 592/07). Y.P. and M.F. are supported by the Sudarsky Center for Computational Biology.

**Conflict of Interest**: none declared.

**REFERENCES**


