BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 24 2007, pages 3343–3349
doi:10.1093/bioinformatics/btm528

Gene expression

Unsupervised feature selection under perturbations: meeting the challenges of biological data

Roy Varshavsky[1,*], Assaf Gottlieb[2], David Horn[2] and Michal Linial[3]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem 91904, [2]School of Physics and Astronomy, Tel Aviv University 69978 and [3]Deptartment of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem 91904, Israel

Received on July 23, 2007; revised on September 12, 2007; accepted on October 15, 2007

Advance Access publication November 7, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: Feature selection methods aim to reduce the complexity of data and to uncover the most relevant biological variables. In reality, information in biological datasets is often incomplete as a result of untrustworthy samples and missing values. The reliability of selection methods may therefore be questioned.

Method: Information loss is incorporated into a perturbation scheme, testing which features are stable under it. This method is applied to data analysis by unsupervised feature filtering (UFF). The latter has been shown to be a very successful method in analysis of gene-expression data.

Results: We find that the UFF quality degrades smoothly with information loss. It remains successful even under substantial damage. Our method allows for selection of a best imputation method on a dataset treated by UFF. More importantly, scoring features according to their stability under information loss is shown to be correlated with biological importance in cancer studies. This scoring may lead to novel biological insights.

Contact: royke@cs.huji.ac.il

Supplementary information and code availability: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Computational biology has undergone a revolution in the last decade. One of the prominent characteristics of this revolution is the development of high-throughput technologies, allowing for gathering of large-scale data, both in the number of samples and in their features. Examples are microarray gene-expression experiments (Beer *et al.*, 2002; Khan *et al.*, 2001) and comparative genomic hybridization (CGH) (Snijders *et al.*, 2005).

A popular strategy for facilitating the analysis and interpretation of such large-scale data is selecting informative features from the thousands measured in each experiment (Guyon and Elisseeff, 2003; Herrero *et al.*, 2003). Feature selection methods are divided into two types: *supervised*, when a target function is known, and *unsupervised*, in which one has no, or limited, information regarding the samples. Supervised feature selection methods are abundant, in particular in the computational biology field, where they were found useful in improving classifications tasks (Bø and Jonassen, 2002). Nevertheless, it was argued that such methods do not lead to a unique set of selected features (Ein-Dor *et al.*, 2006). This is probably due to the fundamental variability within the data and the small number of samples (which is further reduced due to train-test partition), in comparison to the number of features.

Less studied approach is the unsupervised feature selection. Selection methods that are applied before clustering are often referred to as *filter* methods. Most methods of unsupervised feature filtering include ranking of features according to different criteria: correlation with the first principal component, range, fold-change, threshold, entropy and variance calculated on each feature individually (Guyon and Elisseeff, 2003; Herrero *et al.*, 2003). An underlying assumption for these selection methods is that only features that significantly vary along the samples carry the relevant information. Although it seems that unsupervised methods are scarce and less powerful than the supervised ones, most analysts (often inattentively), do apply some unsupervised schemes: in practice, almost every microarray analysis starts with filtering out thousands of genes with small variance or those that are below a predetermined fold-change threshold.

Recently, we have suggested an unsupervised feature filtering (UFF) framework (Varshavsky *et al.*, 2006) that was successfully applied to several datasets with various representations (e.g. gene-expression, amino-acid composition counts). UFF differs from other popular unsupervised selection schemes by (1) not involving a target function as the selection criterion [e.g. optimizing clustering results (Dy and Brodley, 2004)] and (2) considering the interplay of all features. It has been shown on several datasets of different types that a selection of only a few features according to the UFF method leads to improved clustering results relative to other unsupervised methods or to using the complete set.

Here, we investigate the effect of missing information on feature selection strategies. We employ UFF and study whether it remains valid when fractions of data are eliminated. In particular, we put emphasis on the stable features that continue to be selected under these conditions.

*To whom correspondence should be addressed.

Experimental data are prone to errors or information loss because of two major reasons: (i) missing or untrustworthy samples (Wang *et al*., 2006); (ii) missing values: unarguably, this is one of most bothering issues when handling gene-expression microarray datasets (de Brevern *et al*., 2004; Scheel *et al*., 2005); other microarray-based technologies (e.g. tiling array, ChIP on Chip and CGH screening) impose similar challenges. There exists a continuous drive to overcome these problems by improving the hardware (Shi *et al*., 2006), and developing imputation methods to replace missing values (Gan *et al*., 2006; Hua and Lai, 2007; Troyanskaya *et al*., 2001; Tuikkala *et al*., 2006). 'White noise' was shown to have negligible effect on the analysis (Klebanov and Yakovlev, 2007) and thus should not be considered.

Facing the fact that any data may be afflicted by missing information, we argue that a feature selection method should be relatively stable with respect to such errors. This assertion can be tested by simulating information loss and studying its effect on the method at hand. We evaluate UFF under such conditions, suggest viewing stability as a new criterion for feature selection, and study its use on biological data, leading to interesting new insights.

2 DATA AND METHODS

Figure 1 summarizes the analysis protocol. The original dataset (Section 2.1) is perturbed (Section 2.2) and filtered by UFF (Section 2.3). The selected features are then evaluated (Section 2.4) and tested with respect to their biological relevance (Section 2.5).

2.1 Datasets

A comparative analysis is performed on two (complete) gene-expression benchmarks, with known classifications, and a practical application is then applied to a Comparative Genomic Hybridization (CGH) dataset that inherently contains some missing values.

(1) SRBCT: the small round blue cell tumor gene-expression dataset includes glass-based cDNA microarray measurements of 2308 genes (features) for 83 patients (samples). The samples are categorized into four types of tumors: Burkitt lymphoma, Ewing sarcoma, Neuroblastoma and Rhabdomyosarcoma (Khan *et al*., 2001).

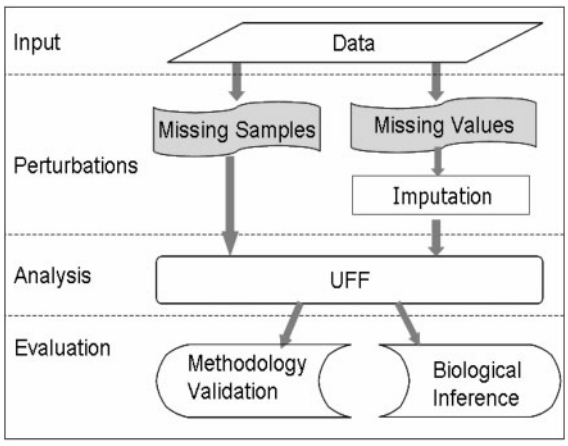

Fig. 1. Schematic representation of the analysis protocol.

(2) Lung: this HUGeneFL Affymetrix oligonucleotide gene-expression dataset (Beer *et al*., 2002), includes 86 primary lung adenocarcinomas and 10 non-neoplastic lung samples. Total 4966 genes are measured for each sample (features).

(3) CGH: this dataset (Snijders *et al*., 2005) comprises 1979 clones (features) for 89 instances (samples). The expression value of each record is the $\log_2$ratio normalized to the genome median $\log_2$ratio. The dataset contains 5807 missing values (3.3%).

2.2 Perturbations

Assuming the complete dataset is a full $[m \times n]$ matrix A, with m features describing n samples (or observations) we simulate information loss in two ways:

(1) Missing samples (Wang *et al*., 2006) are simulated by eliminating some of the columns in the matrix. We consider cases where 1%, 2%, 5%, 10%, 20% and 50% of all samples are randomly removed. Total 50 random eliminations were applied to each group size (in the leave-one-out case, all possibilities are considered).

(2) Missing values are modeled by randomly eliminating 1%, 2%, 5%, 10%, 20% and 50%, of all matrix elements. Total 50 random deletions were selected for each group size. The removed matrix elements are then imputed according to one of three imputation methods:

   (a) Standard average: each missing value is replaced with the average of all present values in the set.

   (b) Weighted average: each missing value is replaced by: [*average* (*row*) * *average* (*column*)]/*average* (*matrix*).

   (c) KNNImpute according to Troyanskaya *et al*. (2001), each missing value is replaced by the standard average of samples of the K nearest neighbors of a relevant feature ($K = 10$).

For clarity, (1) description of the KNNImpute method, (2) results of 50% data loss and (3) SDs appear in Supplementary Material.

2.3 Unsupervised feature filtering (UFF)

UFF scores each one of the features according to its contribution to the SVD entropy of the dataset. Computation of the score is based on a leave-one-out principle [for a complete description see Varshavsky *et al*. (2006)].

Let A denote a matrix, whose elements A_{ij} are the measurement of feature i on sample j, e.g. expression of gene i under condition j. We base our method on the Singular Value Decomposition (SVD) procedure. It decomposes the original matrix A into $A = USV^{\mathrm{T}}$, where U and V are unitary matrices whose columns form orthonormal bases. The diagonal, non-negative matrix S is composed of singular values (s_k).ordered from highest to lowest. Let l be the rank of the matrix [$l \leq \min(m, n)$], Using the normalized relative values, ρ_k

$$\rho_k = \frac{s_k^2}{\sum_{i=1}^{l} s_i^2} \tag{1}$$

a SVD-entropy (H) can be defined (Alter *et al*., 2000):

$$H = -\frac{1}{\log(l)} \sum_{k=1}^{l} \rho_k \log(\rho_k) \tag{2}$$

SVD-entropy varies between 0 and 1. Low entropy datasets are characterized by only a few high singular values whereas the rest are significantly smaller. This pattern reflects a great redundancy in the dataset. In contrast, non-redundant datasets result in uniformity in the singular values spectrum and in high entropy.

UFF scores each feature i using a leave-one-out calculation of the SVD-entropy: H is calculated for the entire matrix and for the matrix without feature i. The difference in the values defines the score of feature i. Figure 2 displays the results after applying the UFF algorithm to the SRBCT dataset, and sorting the features according to decreasing UFF scores. Clearly, one can divide the features into three groups:

(1) Features with positive score. These features increase the entropy.

(2) Neutral features that have negligible influence on the entropy.

(3) Negative score features. These features decrease the entropy.

Note that a majority of features falls into group 2 (~92%), while groups 1 and 3 represent minorities (~4% in each). The features selected according to the UFF approach are the positive score features [lying above the threshold of mean(score) + SD(score)]. The rationale behind picking group 1 features is that, because they increase the entropy, they decrease redundancy. Hence, we may expect samples to be better separated in the space spanned by these features.

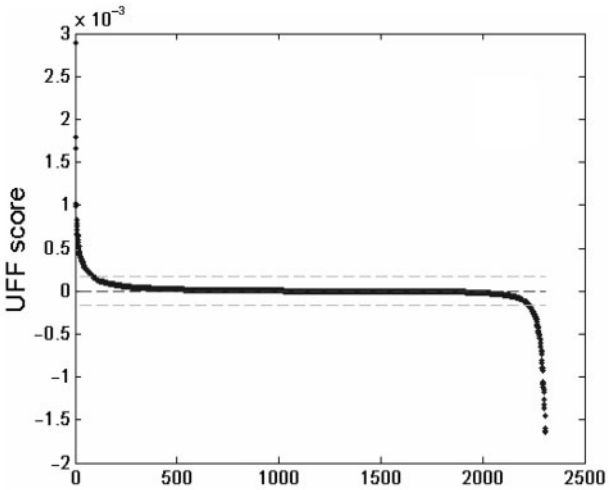

Fig. 2. UFF Scores of the 2308 genes of SRBCT features, ordered by decreasing scores. Dashed lines represent mean(score) ± SD(score).

2.4 Methodology evaluation

Given a set of selected features we evaluate it according to how successful it is in clustering correctly the set of samples, and how much it overlaps with the set of UFF selected features of the unperturbed data.

- Clustering quality. Clustering quality is measured both on perturbed and on perturbed-then-filtered datasets. Cases where the latter representation leads to higher quality indicate that the filtering is effective even though the dataset is damaged. This quality is measured using the Jaccard score: $J = n_{11}/(n_{11} + n_{10} + n_{01})$, where n_{11} is the number of pairs of samples that are classified together, both in a known classification and in the clusters obtained by the algorithm; n_{10} is the number of pairs that are classified together in the true classification, but not in the clustering and n_{01} is the number of pairs that are classified together by clustering but not in the true classification. In order to ensure that the evaluation is not biased by the clustering method, two clustering methods were compared and shown to provide consistent behavior patterns. In the two microarray datasets both QC [$\sigma = ½$, dims = 5, (Horn and Axel, 2003)] and hierarchical (Euclidian distance, average linkage) methods were considered.
- Filtering stability. Filtered features of the original and perturbed datasets are compared (Scheel *et al.*, 2005). The degree of intersection (similarity score) indicates the method's stability under the perturbation.

2.5 Stability scores

On average, each dataset has undergone ~1200 perturbations. Stability of a feature is defined as the probability of this feature to be selected under all perturbations. The features may be then ranked according to this criterion.

3 RESULTS

3.1 Methodology validation: filtering quality and stability

3.1.1 Smooth degradation of clustering quality under perturbations Figure 3 displays the clustering quality of the perturbed SRBCT and Lung datasets (missing samples and missing values with three imputation methods). UFF always

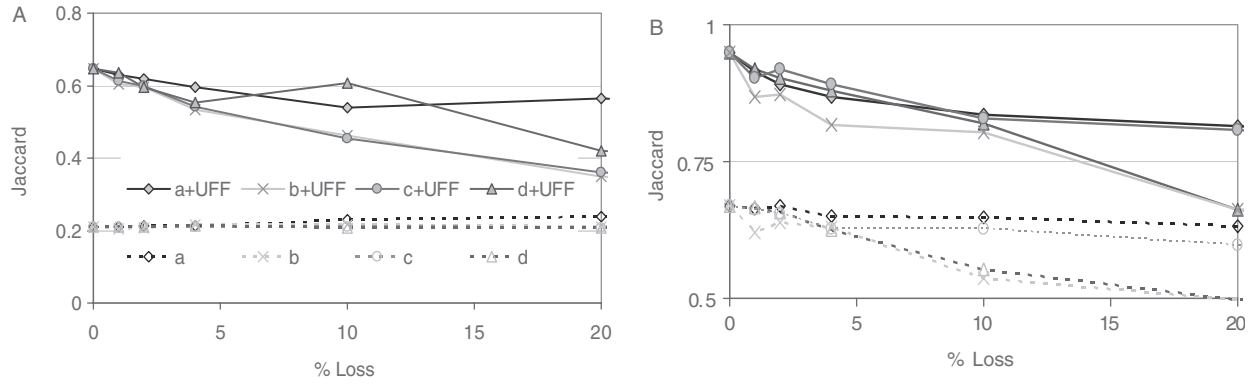

Fig. 3. Clustering results of the (**A**) SRBCT and (**B**) Lung datasets, following perturbations: missing samples (a) and missing values (with three imputation methods: (b) average, (c) weighted average and (d) KNNImpute). Dashed lines denote the clustering quality of the perturbed datasets after various levels of information loss and the continuous lined denote the corresponding quality of perturbed and then filtered sets (results shown are averages of 50 random perturbations). Detailed results for the two datasets appear in Supplementary Material.

improves clustering quality. The results degrade smoothly as a function of the amount of missing data. This allows us to draw two important conclusions: (1) UFF continues to be a good filtering method even under severe information loss. (2) There does not seem to exist a critical amount of loss beyond that clustering quality suffers a sudden drop.

In all *missing sample perturbations* cases, application of UFF improves considerably the clustering quality even under substantial information loss. This is also the case with *missing values perturbations*. Clustering after UFF outperforms clustering without UFF. Comparing between three imputation methods, we learn that the best method for the SRBCT dataset is the KNNImpute while for the Lung dataset it is the weighted average.

3.1.2 UFF is stable under perturbations The stability of filtering is measured by the similarity between the original list of features (selected when the information is complete) and the lists that are generated from the perturbed sets. The lists for the SRBCT and Lung datasets (comprising 88 and 62 genes, respectively) appear in the Supplementary Material.

Figure 4 displays the similarity scores of the perturbed SRBCT and Lung datasets as a function of the lost data. As shown, in the missing samples perturbation, the intersection levels remain high even after substantial loss. This means that UFF is stable under missing samples perturbations.

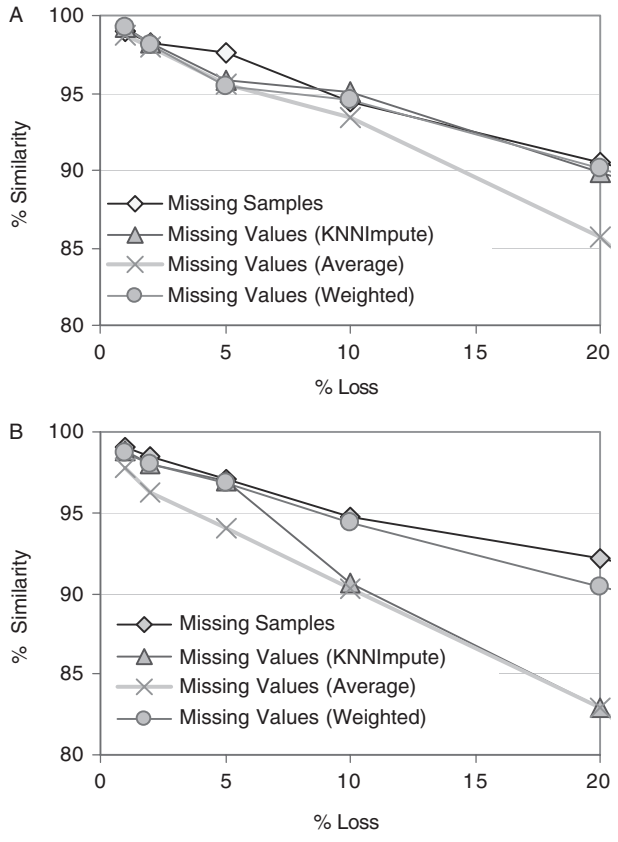

Fig. 4. Similarity levels as a function of lost data of the (**A**) SRBCT and (**B**) Lung datasets. Detailed results for the two datasets appear in Supplementary Material.

In the missing values perturbation, not all imputation methods perform equally. In both cases the simple average method performs relatively bad, while the weighted imputation method performs very well. In the SRBCT dataset the KNNimpute yields high similarity results, yet in the Lung dataset this method is found to result in less stable lists. Overall, similarity is seen to decrease linearly with information loss. In both perturbation schemes the intersection is high (~85%) even after substantial loss (20%). Similar qualitative results have been obtained by Scheel *et al*. (2005) in a supervised selection task.

3.2 Application to a faulty dataset

Given the CGH dataset that contains 3.3% missing values (see Section 2.1), we apply to it further artificial information loss in order to estimate (1) how damaging is the 3.3% original loss, and (2) which is the best imputation method.

The analysis starts with applying the three imputation methods to the dataset. Applying UFF to the three reconstructed forms, results in three lists of selected features, comprising 88, 83 and 85 clones for the average, weighted average and KNNImpute, respectively. These three lists, that are referred to as baselines, have 72 clones in common (Table S3). As shown in Figure 5, the dataset is further perturbed, both by missing values and by missing samples protocols. The resulting lists of features are then compared with their corresponding baseline lists. Figure 5 displays the

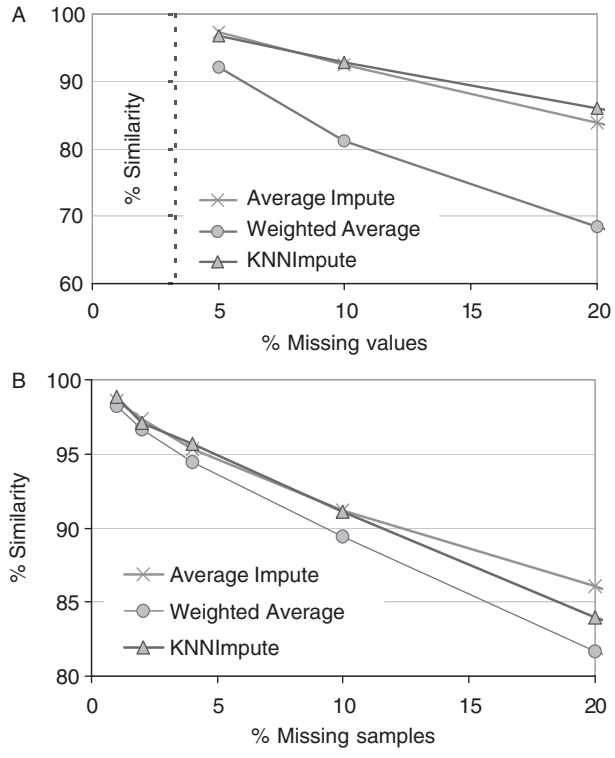

Fig. 5. Similarity scores as a function of lost data of the CGH dataset with (**A**) missing values and (**B**) missing samples perturbations. Note that the missing values analysis starts with the original 3.3% loss. Detailed results appear in Supplementary Material.

similarity scores as a function of the information loss. Note, that since three baseline lists are defined, three comparisons are applied to both protocols.

Clearly, under all perturbations, the similarity levels degrade smoothly (almost linearly), retaining high intersections (~85%) with the original lists even after substantial loss (20%). The high similarity levels may testify that, as far as clones selection is considered, the original 3.3% damage is not crucial. This observation matches the one found in the gene-expression case, which suggests that the stability characteristic of UFF is generic. Furthermore, both protocols lead to similar ranking of the different methods with weighted average inferior to the other two imputation methods.

4 BIOLOGICAL INFERENCE

In this section, we wish to study whether the stability criterion is also biologically meaningful, i.e. are the stable features causally related to the biological problem at hand?

4.1 Ranking stable features

Figure 6 displays the stability scores of the 88 first UFF genes in the SRBCT dataset (according to 0 and Varshavsky *et al.*, 2006). There exists a positive correlation between the rank order of the UFF score and stability. They are compared to the ranking of Khan *et al.* (2001) based on a supervised criterion. Out of 88, 37 of the UFF genes are common to the two lists (hypergeometric enrichment P-value of $1.7E^{-12}$).

Among the 10 and 20 top stable genes, 8 and 13, genes appear in the supervised-selection based list, respectively. The 20 most stable genes are listed in Table 1 (complete lists of the two datasets appear in the Supplementary Material, Tables S1A,B and S2).

4.2 Comparing stable and 'less-stable' SRBCT genes

4.2.1 Statistical analysis We conducted a statistical comparison of top 20 stable genes, with the 20 genes that were originally selected by the UFF algorithm, but found to be less stable (with stability score ranging from 0.85 to 0.51). The top stable genes have relatively low skewness and kurtosis, compared to the less stable genes. Since imputation methods tend to smooth distributions, wide symmetrical distributions should indeed be more resistant to perturbations.

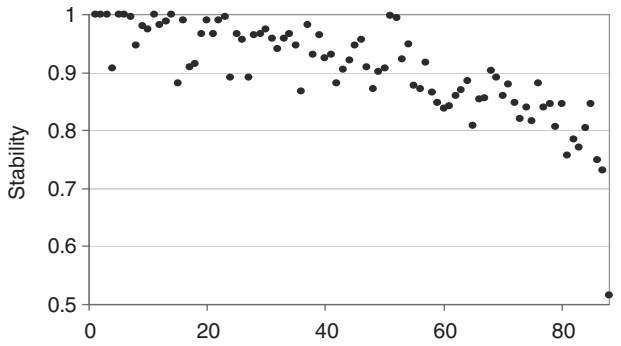

Fig. 6. Stability scores of the top scored UFF-based selection (88 genes) in the SRBCT dataset.

4.2.2 Functional analysis for the most stable genes The malignant tumors analyzed tend to occur in childhood. From a morphological view, subtle clues distinguish between the tumors. At present, analysis for chromosomal abnormalities and molecular probes are being used to assist the pathologists. The list of most stable features in the SRBCT set is intriguing. Among the top stable genes, several genes corroborate each other. Figure 7 illustrates protein–protein interactions that were experimentally validated. Several of the top 20 stable genes appear in these networks. The appearance of representative

Table 1. Top 20 stable genes in the SRBCT dataset

Stability ranking	Stability score	Genes name	UFF ranking	Khan's ranking
1–11	1	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF	1	2
1–11	1	Insulin-like growth factor 2 (somatomedin A)	2	1
1–11	1	Collagen, type III, alpha 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)	3	40
1–11	1	Insulin-like growth factor binding protein 2 (36kD)	5	8
1–11	1	Human insulin-like growth factor binding protein 5 (IGFBP5) mRNA	6	62
1–11	1	SMA3	11	–
1–11	1	Actin, alpha 2, smooth muscle, aorta	14	83
1–11	1	Antigen identified by monoclonal antibodies 12E7, F21 and O13	51	73
1–11	1	IM-379708	23	–
1–11	1	Growth-associated protein 43	7	31
1–11	1	Spectrin, beta, non-erythrocytic 1	52	–
12–15	0.99	Regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	20	57
12–15	0.99	Nucleolin	22	–
12–15	0.99	Gelsolin (amyloidosis, Finnish type)	16	–
12–15	0.99	Troponin T2, cardiac	13	25
16–19	0.98	Crystallin, alpha B	12	79
16–19	0.98	Secreted protein, acidic, cysteine-rich (osteonectin)	37	–
16–19	0.98	Collagen, type I, alpha 2	9	–
16–19	0.98	Follicular lymphoma variant translocation 1	30	75
20	0.97	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)	10	3

In addition, the ranking of the genes according to Khan *et al.* (2001) is given. '–' denote that a gene is not included in the reported 96 genes list (Khan *et al.*, 2001).

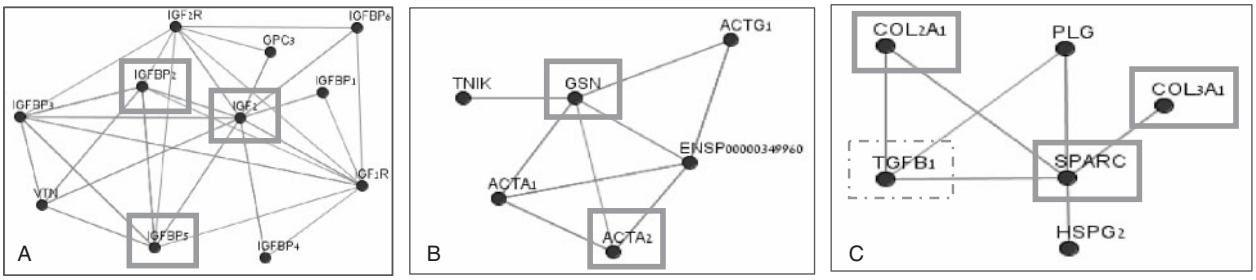

Fig. 7. Experimentally identified gene networks (von Mering *et al*., 2007). (**A**) IGF-2 and interacting proteins; (**B**) Actins (ACT) and Gensolin (GSN) and (**C**) Collagen (COL), Osteonectin (SPARC) and TFGβ (TGFB). Genes included in Table 1 are framed (dashed frame indicates a UFF selected gene, but not among the top 20).

genes within protein networks is an indication for the importance of the identified biological process in the classification. The most evident property is that the stable genes are strongly involved in regulatory networks. In general, several genes are involved in signal transduction (i.e. IGF response), regulation of cytoskeleton and extracellular signaling.

Some genes, listed among the top ranked genes, belong to cytoskeleton elements and their regulators (including actin, gelsolin, troponin, cardiac actin alpha 2, alpha B crystalline and beta spectrin). Their roles as tumor subtype classifiers are not evident and should be experimentally validated.

The biological properties of the less stable genes are different from the top ranked 20 genes. In general, many of these genes associate with a nuclear function and thus may belong to the tumorigenesis process. Among these genes are H2A histone, DEAD/H hnRNP K, FMR1 interacting protein 2, Cyclin-dependent kinase 2-associated protein 1 and more. It is possible that they are altered in tumors, but play a weaker role in distinguishing among the different types.

5 DISCUSSION

We have subjected UFF to a perturbation-based analysis and found it to obey the condition of stability. A similar perturbation-based selection was shown to be efficient in supervised tasks (selection and classification) (Chen *et al*., 2007). Ours is the first unsupervised perturbation-based selection procedure. We recommend using stability under perturbations as an important diagnostic tool when searching for a feature selection method.

Although for practical reasons, perturbation of even 10% should be already considered as significantly severe, in this study we extended our analysis to much higher damage levels (up to 50% of the data, see Supplementary Material). The reason for doing so is twofold: (i) acquire a deep understanding of the nature of the method and the data. It is of interest to investigate whether extensive damage, beyond some critical amount, leads to a collapse of our method (known as critical transition or percolation in various physical systems). In the problem studied here we observe a smooth, almost linear degradation in performance. (ii) In the context of gene expression, the number of unreliable or suspicious samples might often reach a significant fraction of the entire dataset. Often these samples are not literally missing but result from unreliable RNA extraction, low quality labeling, etc. We were therefore motivated to examine how removing many samples influences the lists of selected features (genes).

We have found that the effect of missing samples is very similar to the one of missing values (followed by imputation). In both, even a substantial loss of data does not significantly alter the list of the selected features, reaching a similarity of $\sim$85%. Nevertheless, it should be emphasized that this argument should be limited to datasets with no inherent dependency among the samples. Examples for such dependencies are: time series, cell-cycle and pre-post treatment for the same individuals.

Differences in the imputation methods are identified, emphasizing that imputation method needs to be data-driven. For instance, KNNImpute is usually found perform best in the low loss region while the two average-based imputations achieve higher similarity levels at the high loss region. This last finding can be explained by the local nature of the KNNImpute method (relying only on nearest neighbors). This understanding may assist in selecting among the various imputation methods.

In the cases analyzed, a high correlation between the external and internal criteria (clustering quality and filtering stability, respectively) is reported. Specifically, in both gene-expression benchmarks the two evaluation criteria rank the imputation methods identically. This observation can be exploited to select an imputation method given a dataset. Interestingly, when applying the NRMSE (Normalized Root Mean Square Error), the standard internal criterion for evaluating imputation methods, a different methods–ranking is reported (see Supplementary Material). This suggests that our unsupervised, internal, similarity measure may be a more reliable criterion for selecting an imputation method. We therefore suggest testing the imputation method in conjunction with an unsupervised feature selection method, such as UFF. Not only does it test stability of the selected features, it also points out the best imputation method to be used under these conditions.

Identifying genes as biomarkers for tumor detection and classifications and for the multiple neurological malfunctions is of ultimate importance. Many genes selected by our stability criterion are in agreement with the ones that were found in a supervised manner. However, some potential new features are suggested. Identifying new potential markers may be due to the lack of bias in our analysis, neither from sample labeling nor

from pre-selected classifier algorithm. Moreover, by applying the method on the entire dataset (without train-test splitting), we manage to reduce the well-known pitfall of over-fitting.

ACKNOWLEDGEMENTS

This research is supported by EU FR6 DIAMONDS consortium. R.V. is awarded a fellowship by the SCCB, the Sudarsky Center for Computational Biology of the Hebrew University of Jerusalem.

Conflict of Interest: none declared.

REFERENCES

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

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Bø,T.H. and Jonassen,I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biol.*, **3**, 1–17.

Chen,L. *et al.* (2007) Noise-based feature perturbation as a selection method for microarray data. In: Mandoiu,I. and Zelikovsky,A. (eds.), *ISBRA*. Springer-Verlag, Atlanta, GA, pp. 237–248.

de Brevern,A. *et al.* (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, **5**, 114.

Dy,J.G. and Brodley,C.E. (2004) Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, **5**, 845–889.

Ein-Dor,L. *et al.* (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl Acad. Sci. USA*, **103**, 5923–5928.

Gan,X. *et al.* (2006) Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Res.*, **34**, 1608–1619.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Herrero,J. *et al.* (2003) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.

Horn,D. and Axel,I. (2003) Novel clustering algorithm for microarray expression data in a truncated SVD space. *Bioinformatics*, **19**, 1110–1115.

Hua,D. and Lai,Y. (2007) An ensemble approach to microarray data-based gene prioritization after missing value imputation. *Bioinformatics*, **23**, 747–754.

Khan,J. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Klebanov,L. and Yakovlev,A. (2007) How high is the level of technical noise in microarray data? *Biol. Direct*, **2**, 9.

Scheel,I. *et al.* (2005) The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, **21**, 4272–4279.

Shi,L. *et al.* (2006) The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.

Snijders,A.M. *et al.* (2005) Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, **24**, 4232–4242.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tuikkala,J. *et al.* (2006) Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*, **22**, 566–572.

Varshavsky,R. *et al.* (2006) Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**, e507–e513.

Mering,C. *et al.* (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–362.

Wang,D. *et al.* (2006) Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene-expression profiles and functional modules. *Bioinformatics*, **22**, 2883–2889.