Data Models for Exploratory Analysis of Heterogeneous Microarray Data

Jaewoo Kang

NC State University, Raleigh NC 27695, USA

1 Introduction

Microarrays are one of the latest breakthroughs in experimental molecular biology. It provides a powerful tool by which the expression patterns of thousands of genes can be monitored simultaneously and are already producing huge amount of valuable data. Analysis of such data is becoming one of the major bottlenecks in the utilization of the technology. The gene expression data are organized as matrices — tables where rows represent genes, columns represent various samples such as tissues or experimental conditions, and numbers in each cell characterize the expression level of the particular gene in the particular sample. Application of microarray technology to biological problems, ranges from understanding of metabolic responses of microbes, to cancer in humans.

The main challenge of analyzing microarray is the virtual explosion in the volume and complexity of gene expression data. Thousands of different research groups generate tens of thousands of microarray gene expression profiles. Different experiments utilize different tissue types, examine different treatment strategies, and consider different stages of disease development. This, along with differences in microarray platform, technology and protocols used in different labs, leads to difficulties in integrating microarray data across experiments.

How to combine the data (gene expression levels) in different microarrays is a challenging problem since these gene expression levels are not necessarily directly comparable.The same gene may exhibit different bias at different data sets. For instance, a gene in the liver tissue may have higher expression level (higher values in a microarray) than that in the skin tissue (lower values in another microarray) by the nature. As a result, directly integrating the microarrays according to the gene ids would result in inconsistency. In addition, microarrays may contain different (overlapping) sets of genes. This increases the difficulties in the integration of the microarray data sets.

The focus of this work is on building a unified data model for microarray data, which allows coherent interpretation of the independently generated heterogeneous microarray experimental data. In order to address this problem, we propose a novel **correlation signature** method. The correlation signature captures the data-set-wise characteristics of a gene in terms of its correlations to a set of landmark genes. Various methods can be used to choose the landmarks, e.g., genes from a particular pathway or deemed important by domain experts, etc. The expression level of a gene at a microarray table can be converted into

Input : Microarray table M ($n \times m$, n genes and m conditions),
set of k landmark genes $L = \{l_1, ..., l_k\}$
Output: Set of gene signature vectors $S = \{\overrightarrow{sig}(g_1), ..., \overrightarrow{sig}(g_n)\}$

for *each gene* g_i *in* M **do**
  for *each gene* l_j *in* L **do**
    $d_j \leftarrow dist(\overrightarrow{g_i}, \overrightarrow{l_j})$
  end
  $\overrightarrow{sig}(g_i) \leftarrow [d_1, d_2, ..., d_k]$
end

Fig. 1. SigCalc: signature computation algorithm.

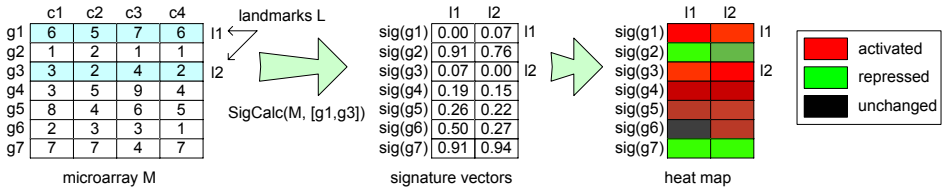


Fig. 2. Example of signature vector computation. Assume $l1$ and $l2$ are regulator genes with similar functions.

the similarity (or correlation) to the set of landmark genes. For example, if there were 10 landmark genes, then at each microarray table, a gene will have 10 correlation values each of which corresponds to a landmark. We call these correlation values as the *correlation signature vector* of the gene. The signature vector removes the bias in the expression values and can be used to compare genes across heterogeneous experiments.

2 Unified Data Model for Gene Expression Profiles

Figure 1 shows an overview of our signature calculation algorithm, **SigCalc**, and Figure 2 illustrates the signature computation process through an example. **SigCalc** takes as input a microarray table M and a set of k landmark genes. The landmark genes can be selected either manually by the user or automatically by the system. If user did not provide landmarks, system can automatically select candidate landmark genes. Different techniques can be used. For example, depending on the application, system may run a feature selection algorithm [60, 62] to choose a set of representative genes in the table, or simply choose a random set of genes and use them as landmarks. With random landmarks, the correlation signature model behaves similar to the random projection, a popular dimensionality reduction method [1, 3, 22, 31, 32, 45], except that the random projection projects the original high-dimensional space onto a random subspace while the correlation signatures project the original space onto a subspace whose coordinates correspond to the landmark genes (See Section 3 for more details).

When users provide landmarks to the system, they can either explicitly pass a hand-selected genes to the system, or they can just state what kinds of genes they want the system to use. For the latter case, system can guide users to make their choices on the group of genes, by providing information about gene annotations, functional groups, known regulator genes, or genes that are involved in a certain pathway, retrieved from some external sources such as GO ontology database (http://www.geneontology.org) and KEGG pathway database (http://www.genome.jp/kegg/).

Once landmark genes are selected, system calculates signature vectors of all genes in the table as shown in Figure 1. **SigCalc** uses a distance function, *dist*, to measure similarities and dissimilarities between gene vectors (rows of M). Any conventional distance metric can be used including standard metrics such as Euclidean or cosine distance, or some variants that are popular in microarray analysis such as correlation distance or mean-expression distance, as defined below.

- *Euclidean Distance:* Given two gene vectors $\overrightarrow{x}$ and $\overrightarrow{y}$, where $\overrightarrow{x} = [a_1, ..., a_n]$ and $\overrightarrow{y} = [b_1, ..., b_n]$, respectively, the Euclidean distance is : $euc(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{(a_1 - b_1)^2 + ... + (a_n - b_n)^2}$.
- *Cosine Correlation:* Given two gene vectors $\overrightarrow{x}$ and $\overrightarrow{y}$, the cosine correlation is: $cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$. The cosine correlation measures the similarity between gene vectors. For a dissimilarity measure, simply $1 - cos(\overrightarrow{x}, \overrightarrow{y})$.
- *Pearson Correlation:* Given two gene vectors $\overrightarrow{x}$ and $\overrightarrow{y}$, Pearson correlation is: $cor(\overrightarrow{x}, \overrightarrow{y}) = \frac{covariance(\overrightarrow{x}, \overrightarrow{y})}{\sqrt{covariance(\overrightarrow{x}, \overrightarrow{x}) \times covariance(\overrightarrow{y}, \overrightarrow{y})}}$. For a dissimilarity measure, $1 - cor(\overrightarrow{x}, \overrightarrow{y})$.
- *Mean-Expression Distance:* Given two gene vectors, the mean-expression distance is defined as: $dist(\overrightarrow{x}, \overrightarrow{y}) = mean(\overrightarrow{x}) - mean(\overrightarrow{y})$.

Note that the correlation and mean-expression distances are not metrics in a strict sense (e.g., do not satisfy triangular inequality) but introduced here because they are commonly used in practice for microarray analysis. Although Euclidian distance is a common method to represent the similarity or dissimilarity between two vectors, it does not take into account the natural bias of expression level of different types of genes. Some house-keeping genes may naturally express highly while some other genes may always express at a low level. Thus, the distance measure may appear larger for these two types of genes. If we are interested in the fluctuation of the expression levels rather than the absolute gene expression values, then the Euclidian distance measure may not be proper to use. In this case, the correlation metrics could be used.

The mean-expression distance is somewhat simplistic but popular in practice because it gives a natural interpretation of the expression level differences, and can be applicable to the gene vectors with different dimensions. In reality, gene vectors (rows) from different microarray tables almost always have different dimensions (e.g., one table has columns of lymphoblastic leukemia samples and

the other has myeloid leukemia samples; number of columns also may differ.) The first three metrics will not work for such comparison. In contrast, all four distance metrics can be used with our model, after transforming the original gene vectors into the corresponding signature vectors.

Now, consider the example in Figure 2. On the left, it shows an input microarray data table M. Suppose the user selected g_1 and g_3 as the two landmarks, l_1 and l_2, respectively. **SigCalc** transforms the original table into a 7×2 table whose rows represent the signature vectors of the corresponding genes in the original table. In this example we used the correlation distance $(0.5 \times (1 - cor(\overrightarrow{x}, \overrightarrow{y}))$ to calculate the signatures. For example, consider $\overrightarrow{sig}(g_7)$ in the signature vector table. It has two entries [0.91, 0.94] representing correlation distances of gene g_7 to the two landmark genes, g_1 and g_3, respectively.

How do we interpret the distance to the landmarks from a gene? What does it exactly mean that the distance is 0.91 or 0.19? The correlation distance ranges from [0, 1], and a distance close to zero implies the two vectors are correlated and a distance close to one implies the two vectors are inversely correlated. If it is 0.5 it means there is no correlation. Now, let us assume that the two landmark genes, l_1 and l_2, are known regulator genes with similar functions. In this example, if a gene's signature vector contains close-to-zero values, it may mean that the gene is *activated* by the two regulator genes. The opposite also holds. The third table from the left of Figure 2 shows the heat map visualizing the activation/repression relations. In our example, g_7 is repressed while g_4 is activated ($\overrightarrow{sig}(g_1)$ and $\overrightarrow{sig}(g_3)$ are also low but they are the landmark genes, and thus ignored.)

A critical precondition that needs to hold to make the proposed approach work is that some genome-wide dependency relations between genes exist and that the relations are conserved across the different experiments, samples, organs, or even across different organisms. In fact, this is a general belief in the biology community. Genes do not act alone: one gene's expression triggers another gene's expression. While most of the dependency relation will remain unchanged, some statistically meaningful changes may be detected from a comparison like *normal cells* vs. *cancerous counterparts*.

One of the main strengths of our approach is the flexibility in landmark selection. The signatures can be further tuned for a specific analysis by choosing landmarks from only the genes that are relevant to the current analysis. For example, suppose one tries to identify how genes behave differently in two sets of cancer samples (e.g., Leukemia and B-cell lymphoma), with respect to only the genes of certain functions (e.g., cell cycle or metabolism). Using our approach, such comparisons become straightforward; we just need to choose landmarks from the genes with cell cycle or metabolism functions.

Our approach also allows flexible cross-validation and analysis. Virtually any expression data sets can be compared provided that the signatures are generated over the common landmarks. One can compare the properties of genes across different tissues (e.g., skin, liver, blood etc.), different clinical stages of cancers (e.g., metastasis vs. primary, recurrent vs. non-recurrent etc.), or can compare

across even different organisms (e.g., mouse vs. human; mice and men share 99% of genes [58]).

In order to validate the model, we need to answer the following two important questions: 1) *How much information is captured and how much is lost during the signature projection?* and 2) *Are the projected signatures really comparable across datasets if common landmarks are chosen?* In what follows, we present the results of our empirical validation addressing these questions using the real microarray data sets [9, 55].

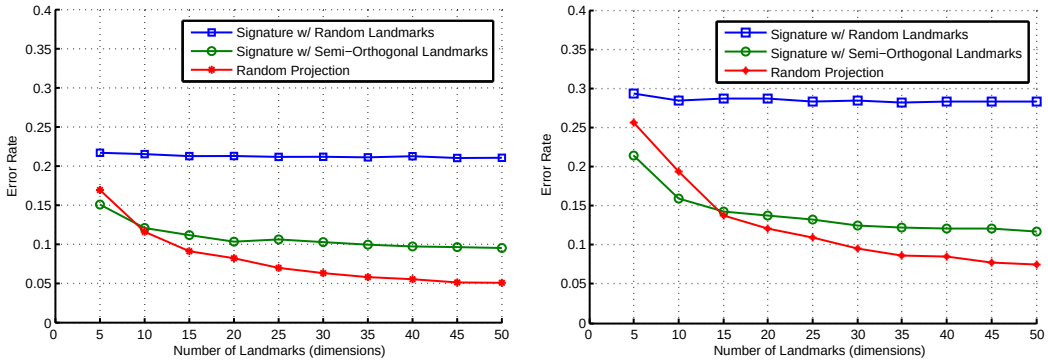
3 Measuring Information Loss

When the raw microarray data is transformed (or projected) into a set of gene signatures, certain information would be lost. The amount of the information loss during the transformation is crucial to the success of the gene signature method. To quantify the information loss, we measured how well the pairwise distances between all pairs of genes are preserved after transformation. The Following is the metric that we used:

> **(Global Distortion Rate)** Let $||.||_F = \sqrt{\sum a_{ij}^2}$ be the Frobenius norm of a matrix, A be a distance matrix whose entries $a_{ij} = dist(\overrightarrow{g_i}, \overrightarrow{g_j})$ where $\overrightarrow{g_i}$ and $\overrightarrow{g_j}$ are two gene vectors in the original space, and B be a distance matrix with $b_{ij} = dist(f(\overrightarrow{g_i}), f(\overrightarrow{g_j}))$ where f is a signature projection. Then, the distortion rate is defined as: $\frac{||A-B||_F}{|||A|+|B|||_F}$.

In other words, A is an $n \times n$ distance matrix that contains all pairwise distance of genes, $\overrightarrow{g_i}$ and $\overrightarrow{g_j}$ $(1 \leq i, j \leq n)$, in the original space, and B is a corresponding $n \times n$ distance matrix containing all pairwise distances of the same gene pairs in the projected space. We compute Frobenius norm of the differential of the two matrices, normalized by the Frobenius norm of the absolute sum of the two. We used the cosine correlation ($cos(\overrightarrow{g_i}, \overrightarrow{g_j})$) as the distance function. Unlike Euclidean distance, the cosine correlation measures the similarity of the two vectors and it is invariant to the magnitudes of the input vectors. In Microarray data analysis, it is often more important to preserve the pattern similarity among genes than their expression magnitudes.

Figure 3 shows the result of our preliminary investigation. Figure 3(a) compares the error (distortion) produced by three methods: Signature with Random Landmarks (SR), Signature with Semi-Orthogonal Landmarks (SO), and Random Projection (RP). Suppose we are projecting gene vectors in the original n-dimensional space onto a k-dimensional subspace. For SR, we randomly select k landmark genes, $l_1, \ldots, l_k$, from the original space. Each gene g_i is then transformed in to a k-dimensional signature vector $\overrightarrow{sig}(g_i) = [dist(g_i, l_1), dist(g_i, l_2), \ldots, dist(g_i, l_k)]$. For SO, we do the same except the landmark genes are carefully chosen from a set of semi-orthogonal genes. Orthogonal gene vectors are orthogonal to each other in the original n-dimensional space. Semi-orthogonal genes represent a group of genes in which all pairwise cosine similarities are less than a certain threshold θ (we used $\theta = 0.2$ in the experiments).



(a) Leukemia Patients (120 Samples) [9] (b) Myeloma Patients (173 Samples) [55]

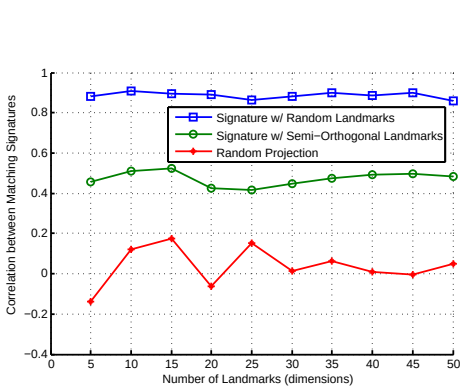
Fig. 3. The error produced by RP, SR, and SO on two microarray datasets. Test performed on a sample of 2000 genes. For each landmark size, iterated 20 times and averaged the result.

Random projection and correlation signatures: RP is a popular dimensionality reduction method proven to be useful in many application areas including text retrieval [3, 45], image processing [3], clustering [14, 29, 46], motif discovery in bioinformatics [6], multimedia indexing [37], just to name a few. Our signature projection method has strong similarity with RP-based approaches. In fact, the signature projection is reduced to an RP problem if the cosine similarity is used as the distance metric (for both signature and global distortion computation), with only difference being that RP projects the original high-dimensional space onto a random subspace while the correlation signatures project the original space onto a subspace whose coordinates correspond to the landmark genes. We make this explicit in the following definition:

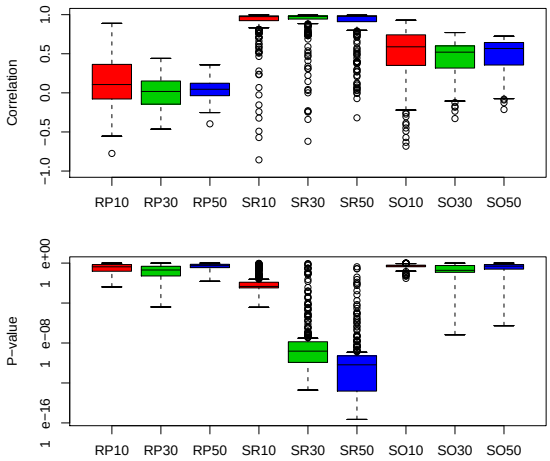
> **(Random Projection)** Let $X_{m\times n}$ be an $m \times n$ matrix whose rows are vectors and columns are dimensions, and $R_{n\times k}$ be an $n \times k$ random matrix whose columns have unit lengths. Then, $X^{RP}_{m\times k} = X_{m\times n}P_{n\times k}$ is a random projection of $X_{m\times n}$ using a projection matrix $R_{n\times k}$.

Strictly speaking, RP is not a projection because the projection matrix $R_{n\times k}$ is rarely orthogonal. However, in a sufficiently high dimension space, vectors with random directions are likely to be close to orthogonal [28], thus making $R_{n\times k}$ an approximate orthogonal matrix. Although RP is known to be generally effective in embedding high-dimensional data into a row-dimensional subspace, it may not solve our problem because it projects the original data into a random subspace, and as a result, the projected subspaces from different datasets are not generally comparable. We use RP as guideline to compare against the performance of our approach. The following lemma gives a bound on the distortion that RP may produce during the projection.

> **(Johnson-Lindenstrauss Lemma)** [32] Given $\epsilon > 0$ and a projection f, if a set of n points $p_1, ..., p_n \in S$ in Euclidean space are projected onto



(a) Correlation between matching signatures



(b) Distribution of correlation and P-values

Fig. 4. Two sets of signatures are generated from two disjoint subsets of the leukemia samples. Signatures are aligned based on their gene ids and then compared. (RP - Random Projection, SR - Signature w/ Random Landmarks, SO - Signature w/ Semi-Orthogonal Landmarks, with 10, 30, 50 dimensions.)

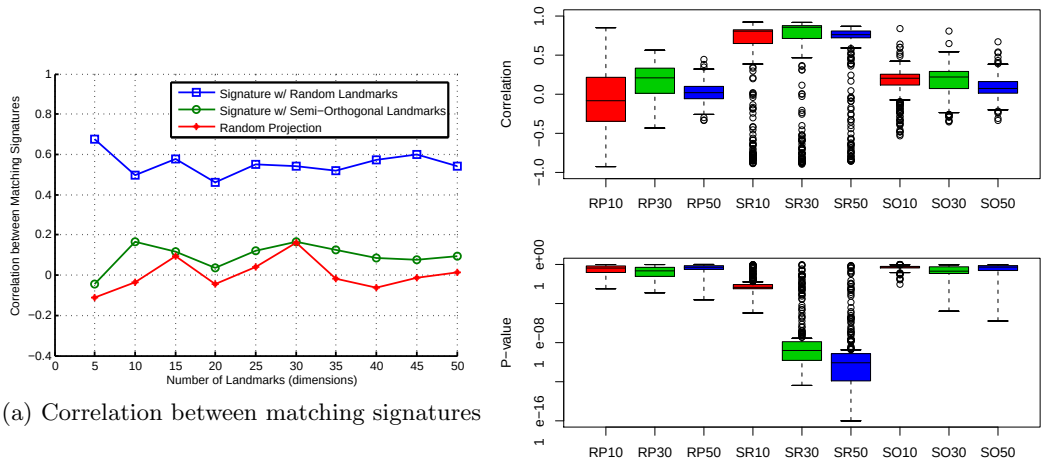
a random k-dimensional space where $k = O((\log n)/\epsilon^2)$, then with high probability, for all $p_i, p_j \in S$

$$(1-\epsilon)||p_i - p_j||^2 \leq ||f(p_i) - f(p_j)||^2 \leq ||p_i - p_j||^2(1+\epsilon)$$

The above lemma states that if points in a vector space are projected onto a random subspace with sufficiently large dimensions, then all pairwise distances between the points are approximately preserved. Unlike RP, in SR and SO, the projection matrix is an $n \times k$ matrix whose columns are landmark gene vectors either selected randomly (SR) or selected from a set of semi-orthogonal gene vectors (SO). Although we draw a small number of genes (typically 10-50) randomly from a large pool of genes (>10K), it is quite possible that some genes in the landmarks have some correlations. This will introduce some distortion in the projected subspace. In order to avoid that, in SO, we preprocessed the input tables to find a maximal set of genes (Q) that are semi-orthogonal (i.e., for all $\overrightarrow{g_i}, \overrightarrow{g_j} \in Q$, $cos(\overrightarrow{g_i}, \overrightarrow{g_j}) < \theta$), and chose the landmarks only from the set. In order to find the semi-orthogonal gene set, we reduced the problem to a *maximal clique* problem as follows: (1) compute a distance matrix D whose entry $d_{ij} = cos(\overrightarrow{g_i}, \overrightarrow{g_j})$, (2) set all $d_{ij} \geq \theta$ to 0 and all $d_{ij} < \theta$ to 1, and (3) solve the maximal clique problem using D as a connectivity graph. There exist some polynomial time approximation algorithms for this problem [4, 13] (The maximal clique problem is NP-hard). However, the size of the graph (#of nodes $>$ 10K) may still be problematic even for a good approximation algorithm. In order to address this problem, we employed the following simple strategy: (1)

select a random subgraph D' of k nodes ($k \ll 10K$) from D, (2) compute the maximal clique $Q_{D'}$ for D', and (3) perform monotonic greedy search using $Q_{D'}$ as a starting point.

Turning back to Figure 3, the result confirmed our expectation. In both datasets, RP performed the best, followed by SO and SR. RP and SO improved as more numbers of landmarks were used, while SR didn't show any significant changes over the range of different landmarks. It appears that, unlike SR, the projection matrices of RP and SO were close enough to orthogonal and adding more dimensions to the projection matrix, following Johnson-Lindenstrauss Lemma, improved the result. RP achieved 5% and 8% distortion rate in leukemia and myeloma datasets, respectively, while SO achieved 10% and 12%, respectively.



(a) Correlation between matching signatures

(b) Distribution of correlation and P-values

Fig. 5. Two sets of signatures are generated from two independent data sets (leukemia and myeloma). Signatures are aligned based on their gene ids and then compared. (RP - Random Projection, SR - Signature w/ Random Landmarks, SO - Signature w/ Semi-Orthogonal Landmarks, with 10, 30, 50 dimensions are compared.)

4 Comparing Signatures across Datasets

As shown in Figure 3, RP clearly outperformed the signature methods with respect to preserving the pairwise distances. However, RP can not be used for comparing or integrating microarray data from different sources because RP projects onto a random subspace which can not be shared by multiple datasets. On the other hand, signature methods choose a set of landmark genes from its own dataset and use them as the coordinates of the subspace on which all gene vectors of the dataset will be projected. As a result, if the same set of genes are

selected as landmarks, the projected subspaces, even if they are from different datasets, are comparable. In order to test this, we performed the following two experiments:

- *Comparing Disjoint Subsets*: (1) We split a dataset A vertically into two disjoint subsets, A_1 and A_2, so that two sets have disjoint samples (columns), (2) select k landmark genes $l_1, ..., l_k$, let the set of gene vectors corresponding to the chosen landmarks in A_1 and A_2 be L_1 and L_2, respectively, (3) compute signatures for all genes in each dataset independently using the corresponding sets of landmark vectors, L_1 and L_2, and then finally (4) compute the correlations between all pairs of matching signatures.
- *Comparing Different Datasets*: We do the same as above except that here we use two independent datasets, A and B, instead.

Figure 4 shows the result of the first test using the leukemia dataset (10K genes, 120 samples). In the test the original dataset is split into two 60 column tables and then compared. If the model really captures the information, the signature vectors of corresponding genes across the two sets should be very similar because they are generated from the same type of samples. Figure 4(a) shows the average correlation of all matching signatures across the two sets using different sizes of landmarks (iterated 20 times for each data point and averaged result). Interestingly, SR outperformed SO with a big margin throughout the entire test range. It struck us as a surprise. In the previous experiment shown in Figure 3, SO preserved much more information than SR (more than 100% better in both datasets). We can think of two possible reasons: 1) First, the set of semi-orthogonal gene vectors from which the SO landmarks are chosen, may not have been a representative set covering broad range of gene functions. If the set contains genes that represent only a small fraction of biological functions, the signatures computed against them will be skewed toward that functions represented in the set. 2) Second, SO may be overfitting. Microarray data is very noisy. If SO preserved too much information it might fail to generalize the real signal. We are leaning toward the first explanation but it is not conclusive at this stage with currently available evidences.

Figure 4(b) shows two boxplots illustrating the distributions of correlations and p-values between all matching pairs of signatures. It shows that the distributions of correlations for SRs (10,30,50 landmarks) are highly skewed toward 1.0 (perfect correlation), while, as expected, that of RPs are close to zero (no correlation). The second boxplot shows the p-value distributions. In our context, the p-value states the probability of observing a correlation between two signature vectors *by chance* at the level greater than or equal to the observed correlation. If the pair's p-value is low we can assume that the correlation value between the pair is statistically significant. On the other hand, a high p-value may suggest that no statistically significant correlation exists between the two signature vectors. As shown in Figure 4(b), the p-values of SR gradually improves (become smaller) as more numbers of landmarks are used, while those of RP and SO did not change. This implies that more numbers of landmarks, although did not improve the correlation, helped strengthen the statistical confidence of the

measured correlations. The median correlation of SR with 50 landmarks was 9.8 (in Figure 4(b) top) and its median p-value was less than 10^{-10} (in Figure 4(b) bottom).

Figure 5 shows a similar result using the two different datasets (leukemia and myeloma [9, 55]). The average of all pairwise correlations for SR was less than the previous test. It ranged from 0.5 to 0.6 throughout the test. The median correlation for SR with 50 landmarks was 0.76 and median p-value was 1.17×10^{-10}. In both tests, the result showed the clear differences in the correlations and p-values between SR and RP, suggesting that the signature models indeed capture some statistically significant information about each individual gene and that the signatures can be compared across the disjoint datasets.

5 Conclusion

In this work, we introduced a novel data model, *correlation signature model*, for integrating heterogeneous microarray gene expression profile data. This model allows the coherent interpretation of the independently generated heterogeneous microarray experimental data. The proposed model exploits the dependency structure among data elements within a table to make the comparison.

References

1. Dimitris Achlioptas. Database-friendly random projections. In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, New York, NY, USA, 2001. ACM Press.
2. D.J. Allocco, I.S. Kohane, and A.J. Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5, 2004.
3. Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, New York, NY, USA, 2001. ACM Press.
4. Vincent Bouchitte and Ioan Todinca. Listing all potential maximal cliques of a graph. *Theor. Comput. Sci.*, 276(1-2):17–32, 2002.
5. A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet.*, 29(4):373, 2001.
6. Jeremy Buhler and Martin Tompa. Finding motifs using random projections. In *RECOMB '01: Proceedings of the fifth annual international conference on Computational biology*, pages 69–76, New York, NY, USA, 2001. ACM Press.
7. J. Burke, H. Wang, W. Hide, and D.B. Davison. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.*, 8(3):276–90, 1998.

8. Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 93–103, San Diego, CA, 2000. (data sets are available at http://arep.med.harvard.edu/biclustering/).
9. MH Cheok, W Yang, CH Pui, JR Downing, C Cheng, CW Naeve, MV Relling, and WE Evans. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nature Gen.*, 2003.
10. M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Pacific Symposium on Biocomputing 2003*, 2003.
11. V. Detours, J.E. Dumont, H. Bersini, and C. Maenhaut. Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Lett*, 546:98–102, 2003.
12. A. Drawid and M. Gerstein. A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol.*, 301:1059–75, 2000.
13. Uriel Feige. Approximating maximum clique by removing subgraphs. *SIAM Journal on Discrete Mathematics*, 18(2), 2004.
14. X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *In Machine Learning, Proceedings of the International Conference on*, 2003.
15. Mary Fernandez, Daniela Florescu, Jaewoo Kang, Alon Levy, and Dan Suciu. Strudel: a web site management system. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 549–552. ACM Press, 1997.
16. Mary Fernandez, Daniela Florescu, Jaewoo Kang, Alon Levy, and Dan Suciu. Catching the boat with strudel: experiences with a web-site management system. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 414–425. ACM Press, 1998.
17. Mary Fernandez, Daniela Florescu, Jaewoo Kang, Alon Levy, and Dan Suciu. Overview of strudel - a web-site management system. *Networking and Information Systems Journal*, 1:115–140, 1998.
18. H. Ge, Z. Liu, G.M. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet.*, 29:482–6, 2001.
19. D.L. Gerhold, R.V. Jensen, and S.R. Gullans. Better therapeutics through microarrays. *Nature Genetics*, 32:547–551, 2002.
20. Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *J. Data Mining and Knowledge Discovery*, 1(1):29–53, 1997.
21. A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Res.*, 29:3513–9, 2001.
22. Anupam Gupta. Embedding tree metrics into low dimensional euclidean spaces. In *STOC '99: Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 694–700, New York, NY, USA, 1999. ACM Press.
23. S. Heber. *Algorithms fir Physical Mapping.* PhD thesis, Rupprecht-Karls-Universitaet Heidelberg, 2001.
24. S. Heber, M. Alekseyev, S.H. Sze, H. Tang, and P.A. Pevzner. Splicing graphs and est assembly problem. *Bioinformatics*, 18 Suppl. 1:181–188, 2002.

25. S. Heber and C. Savage. Common intervals of trees. *IPL*, pages 69–74, 2004.
26. S. Heber and J. Stoye. Algorithms for finding gene clusters. *Lecture Notes in Computer Science*, 2149:252–263, 2001.
27. S. Heber and J. Stoye. Finding all common intervals of k permutations. *Lecture Notes in Computer Science*, 2089:207–218, 2001.
28. R. Hecht-Nielsen. Context vectors: general purpose approximate meaning representations self-organized from raw data. *Computational Intelligence: Imitating Life*, 1994.
29. Alexander Hinneburg and Daniel A. Keim. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*, pages 506–517, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.
30. I. Holmes and W.J. Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. In *Proc Int Conf Intell Syst Mol Biol. 2000*, pages 202–10, 2000.
31. Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, New York, NY, USA, 1998. ACM Press.
32. William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Amer. Math. Soc.*, 26:189–206, 1984.
33. J. Kang, T. S. Han, D. Lee, and P. Mitra. Establishing value mappings using statistical models and user feedback. Submitted for Publication (available at http://www.csc.ncsu.edu/faculty/kang/pubs/khlm05.pdf), 2005.
34. J. Kang and J. F. Naughton. "On Schema Matching with Opaque Column Names and Data Values". In *ACM SIGMOD*, San Diego, CA, Jun. 2003.
35. Jaewoo Kang. *Toward the Scalable Integration of Internet Information Sources*. PhD thesis, U. Wisconsin-Madison, 2003.
36. Jaewoo Kang, Jiong Yang, Wanhong Xu, and Pankaj Chopra. Integrating heterogeneous microarray data sources using correlation signatures. In *International Workshop on Data Integration in the Life Sciences (DILS)*, 2005.
37. Mikko Kurimo. Indexing audio documents by using latent semantic analysis and som. *In Erkki Oja and Samuel Kaski, editors, Kohonen Maps*, pages 363–374, 1999.
38. G.R.G. Lanckriet, N. Cristianini, M.I. Jordan, and W.S. Noble. *Kernel Methods in Computational Biology*, chapter 11. Kernel-Based Integration of Genomic Data Using Semidefinite Programming. MIT Press, 2004.
39. J. Leipzig, P. Pevzner, and S. Heber. The Alternative Splicing Gallery (ASG): Bridging the Gap Between Genome and Transcriptome. *Nucleic Acids Res.*, 32(13):3977–83, 2004.
40. Jinze Liu, Jiong Yang, and Wei Wang. Gene ontology friendly biclustering of expression profiles. In *Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB)*, 2004.
41. D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.
42. E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–6, 1999.

43. R. Mrowka, W. Liebermeister, and D. Holste. Does mapping reveal correlation between gene expression and protein-protein interaction? *Nat Genet.*, 33:15–6, 2003.
44. A. Nakaya, S. Goto, and M. Kanehisa. Extraction of correlated gene clusters by multiple graph comparison. *Genome Inform Ser Workshop Genome Inform.*, 12:44–53, 2001.
45. Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: a probabilistic analysis. In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168, New York, NY, USA, 1998. ACM Press.
46. Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl.*, 6(1):90–105, 2004.
47. M. Schena, D. Shalon, R.W. Davis, and Brown.P.O. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.
48. A.O. Schmitt, T. Specht, G. Beckmann, E. Dahl, C.P. Pilarsky, B. Hinzmann, and A. Rosenthal. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.*, 27(21):4251–60, 1999.
49. E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nat Genet.*, 36:1090–8, 2004.
50. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet.*, 34:166–76, 2003.
51. P.T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, G. Bernhart, D.and Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B.J. Aronow, A. Robinson, D. Bassett, Jr.C.J. Stoeckert, and A. Brazma. Design and implementation of microarray gene expression markup language (mage-ml). *Genome Biol.*, 3, 2002.
52. S. Steiner and N.L. Anderson. Expression profiling in toxicology–potentials and limitations. *Toxicol Lett.*, pages 467–71, 2000.
53. J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003.
54. A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:136–44, 2002.
55. E Tian, F Zhan, R Walker, E Rasmussen, Y Ma, B Barlogie, and JD Shaughnessy. The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. *N Engl J Med*, 2003.
56. V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.
57. B. Vogelstein and K. Kinzler. Cancer genes and the pathways they control. *Nature Med.*, 10, 2004.
58. Marsha Walton. Mice, men share 99 percent of genes. CNN Science (http://archives.cnn.com/2002/TECH/science/12/04/coolsc.coolsc.mousegenome/), 2002.
59. Haixun Wang, Wei Wang, Jiong Yang, and Philip Yu. Clustering by pattern similarity in large data sets. In *sigmod*, 2002.
60. Eric P. Xing, Michael I. Jordan, and Richard M. Karp. Feature selection for high-dimensional genomic microarray data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 601–608, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

61. Zhou XJ, Kao MJ, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, and Wong WH. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nature Biotechnology*, 23(2), 2005.
62. Lei Yu and Huan Liu. Redundancy based feature selection for microarray data. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–742, New York, NY, USA, 2004. ACM Press.