R2KS: A Novel Measure for Comparing Gene Expression Based on Ranked Gene Lists

SHENGYU NI[1,2] and MARTIN VINGRON[2,1]

ABSTRACT

Bioinformatics analyses frequently yield results in the form of lists of genes sorted by, for example, sequence similarity to a query sequence or degree of differential expression of a gene upon a change of cellular condition. Comparison of such results may depend strongly on the particular scoring system throughout the entire list, although the crucial information resides in which genes are ranked at the top of the list. Here, we propose to reduce the lists to the mere ranking of the genes and to compare only the ranked lists. To this end, we introduce a measure of similarity between ranked lists. Our measure puts particular emphasis on finding the same items near the top of the list, while the genes further down should not have a strong influence. Our approach can be understood as a special version of a two-dimensional Kolmogorov-Smirnov statistic. We present a dynamic programming algorithm for its computation and study the distribution of the similarity values. The performance on simulated and on real biological data is studied in comparison to other available measures. Supplementary Material is available online (www.liebertonline.com/cmb).

Key words: algorithms, computational molecular biology, gene expression, sequence analysis, statistics.

1. INTRODUCTION

MANY BIOLOGICAL OR BIOINFORMATICS RESULTS COME IN THE FORM of ranked lists of genes. Typical examples comprise database searching with, for example, BLAST, which results in a list of similar sequences ranked by their p-value, or gene expression experiments, where one obtains a ranked list of genes differentially expressed under two conditions. A question, however, that is hardly ever asked is, how similar are two ranked lists. Yet, the ability to compare ranked lists would seem useful in many applications. In gene expression studies, for example, it would allow comparing different experiments. For database search output, it might be used to compare different search methods. In fact, different web search engines have been compared by comparing the ranked lists of the URLs which they identified (Jacso, 2005).

Phrasing the comparison problem on the ranking is tempting in particular when the values, based on which the ranking was performed, are not readily comparable between different lists. This situation arises,

[1]CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes of Biological Sciences, Shanghai, China.
[2]Max Planck Institute for Molecular Genetics, Berlin, Germany.

e.g, upon comparing gene expression experiments performed with different technologies. In this case, the hope is that comparing the experiments solely based on gene ranking might avoid introduction of artefacts due to normalization routines. This problem has become particularly pressing with the advent of RNA-sequencing (RNA-seq), or whole transcriptome shotgun sequencing. RNA-seq has become the leading technology for examining gene expression and has largely substituted for microarrays. However, the two methods have fundamentally different underlying statistical models. Consequently, it is a challenge to interpret the biological findings: how can we evaluate the similarity of experimental results across different platforms? Is it feasible to simply compare the ranked lists of genes?

In this paper we study the problem of comparing ranked lists, put forward a novel similarity measure for ranked lists, and compare it to existing approaches. The key feature of these ranked lists is that the interesting genes are accumulated at the top of the list or, for down-regulated genes, at the bottom. The order of the other genes is unimportant and yet it is generally hard to draw a cut-off, above which we hope to find the bulk of the interesting genes. This distinguishes the comparison of ranked lists from simply comparing two vectors of ranks, say by Spearman correlation coefficient or Kendall's tau. Both of these measures look for conservation of order from top to bottom of the list.

In fact, the GSEA program (Subramanian et al., 2005) captures this emphasis on the top of the list quite clearly, albeit relying on a discrete set of genes to which a given ranked list is compared. In as far as this discrete set may itself be derived by introducing a cut-off in a list sorted according to some criterion, GSEA answers a specialization of our question of comparing two ranked lists.

A number of methods for comparing ranked lists in the sense described have been proposed in the literature. One class of approaches is based on the intersection of the sets consisting of a fixed number k of top-ranking genes (Ein-Dor et al. 2006; Jurman et al., 2008; Qiu et al., 2006; Zucknick et al., 2008). The similarity score from Yang et al. (2006), termed ''overlapping score,'' drops the restriction to a fixed cutoff by integrating an exponentially weighted overlap count over all possible cutoffs k. An alternative approach assesses the overlap above a cutoff using a hypergeometric distribution while varying the cut-off k. The similarity then is determined by the the hypergeometric probability at the cut-off that gives the most significant overlap between the top ranking gene sets (Eden et al., 2007; Fury et al., 2006; Roider et al., 2009). The resulting hypergeometric p-value is not a p-value for the similarity of the two lists due to testing over many possible cutoffs.

GSEA is based on the observation that if a set is irrelevant with respect to a particular ranked list, then the genes from the set will be uniformly distributed over the list. This method then tests for violation of this assumption, essentially by a Kolmogorov-Smirnov test. Our approach to comparison of two ranked lists generalizes this: When two ranked lists have nothing to do with each other, the two-dimensional scatter plot of the rank in the first list versus the rank in the second list will show a two-dimensional uniform distribution. We propose a restricted version of a two-dimensional Kolmogorov-Smirnov statistic, which also captures our expectation that for two similar lists there should be rank conservation at the top of the lists. A special feature of our set-up is that per row or column of the scatter plot there is exactly one dot, such that, even in the random case, we are not really looking at a uniform distribution. We thus call our proposed similarity measure ''R2KS-score'' for Restricted 2-Dimensional Kolmogorov-Smirnov score.

Two-dimensional generalizations of the Kolmogorov-Smirnov statistic have been studied before (Peacock, 1983; Fasano and Franceschini, 1987). There, samples from astronomy are tested for two-dimensional uniformity. Our problem differs in two ways. First, we want to design a similarity measure rather than a statistical test. Testing is not in our focus, because in biological applications we usually have large numbers of data points. Second, the one-dot-per-row/column property introduced above implies an additional specification on the data.

This similarity measure we put forward is expected to give high similarity values for identical or near identical list. The lowest scores should be obtained for pairs of lists that are chosen by random process, while lists with more or less the same genes accumulating at the top should be in between. For R2KS, we will test for this behaviour, both on biological examples, and by systematically generating pairs of lists under a generative model that mimics differential gene lists. We will also use it in a biological application to interpret the score as a measure for the similarity of biological conditions. A simple dynamic programming algorithm for the computation of the score will be presented. For the un-weighted version of our similarity measure, we will also provide numerical evidence allowing to estimate the p-values for pairs of lists.

2. METHODS

2.1. Definition of R2KS-score

The comparison of two ranked lists of genes can be graphically illustrated in a scatter plot. If the two lists are completely unrelated to each other, the genes from one list should be uniformly distributed throughout the other list. Such an example is shown in Figure 1, yielding an approximately uniform distribution in two-dimensional space. Note, however, that in contrast to a sample drawn from a two-dimensional uniform distribution each row or column contains exactly one dot, because each gene occurs once per list. If the same genes tend to accumulate at the top of the two lists, in this depiction one would expect to see a concentration of dots in some lower left rectangle. Our measure aims to quantify this effect by comparing the observed density of genes in such a lower-left rectangle with the expected density. Like for the Kolmogorov-Smirnov statistic, we take the difference between the two densities at the point where this difference becomes maximal. We call this score R2KS for restricted two-dimensional Kolmogorov-Smirnov statistic. The term ''restricted'' refers to the fact that we are not really looking a two dimensional uniform distribution but allow only one dot per row or column.

Formally speaking, when two ranked lists $L1$ and $L2$ of length N each are under investigation, let $A_{i,j}$ be the rectangle $[1, \ldots, i] \times [1, \ldots, j]$, whose indices i and j are positions from $L1$ and $L2$, respectively. Let U be the whole space $[1, \ldots, N] \times [1, \ldots, N]$. See Fig. 1 for illustration. For each entry of $i \times j$, $R_{i,j}$ is defined as:

$$R_{i,j} = \frac{\text{The number of genes in area } A_{i,j}}{N} - \frac{i \times j}{N \times N} \tag{1}$$

The R2KS score R^* is then defined as:

$$R^* = \max_{i,j} (R_{i,j}) \tag{2}$$

A special situation arises in the application of ranked list comparison to gene expression lists, where both up- and down-regulation may be biologically interesting. In this case, we propose to calculate the R2KS score separately from top and bottom, where the score for the bottom of the list is computed simply by reversing the ranked lists and applying the same procedure. The final score is defined as the larger score of the two separate runs. This rule ensures that the absence of either up- or down-regulated genes would not compromise the overall score as would be the case when taking, for example, the average of the runs (Boulesteix and Slawski, 2009).

2.2. Distribution of R2KS-score

For the purpose of determining p-values for the R2KS-score, randomization is always an option. However, it is desirable to describe the distribution of the R2KS-score in a manner independent of the
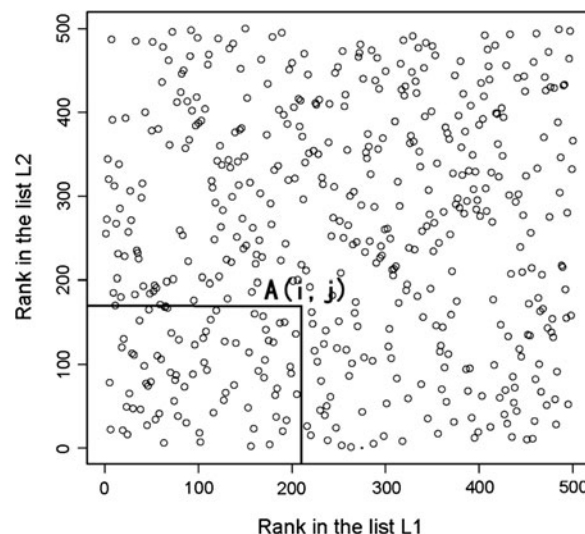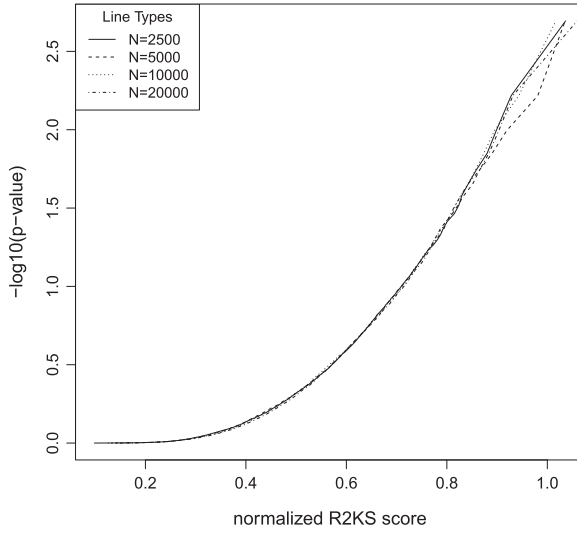

FIG. 1. A Scatter Plot of two unrelated ranked lists.

FIG. 2. Normalized R2KS score. $\sqrt{N} \cdot R$ of ranked lists with different length versus negative logarithm of empirical p-value from simulation.

length N of the lists. For the standard Kolmogorov-Smirnov statistic, the dependence of the significance level on the number of data points n goes with $1/\sqrt{n}$. The same behavior was observed for the two-dimensional Kolmogorov-Smirnov statistic in Peacock (1983). In our case, n is fixed to the list length N. As a hypothesis, we tested the correction with $1/\sqrt{N}$ also for our restricted two-dimensional case. To this end, we define the normalized score $\sqrt{N} \cdot R$ and compute its distribution for different lengths N. Figure 2 shows a plot relating normalized R2KS-scores from randomly generated lists of different length with the negative logarithm of the empirical p-values. Due to the normalization, the distributions have become nearly identical. This suggests that one can tabulate p-values from simulation and use the same length-normalization as for the Kolmogorov-Smirnov statistic.

2.3. Weighted version of R2KS

For practical applications we also employ a weighted version the R2KS-score. The motivation here is to inform the R2KS score about the user's expectation within which interval from the top the informative genes will be found. Each gene is assigned a weight w_k. We propose to use a pivot in the ranked list which conceptually divides important genes from the rest. We keep the number of important genes deliberately large by having it account for roughly 20% of the whole list. For genes above the reference point, the weight is defined as follows:

$$w_k = \frac{h \cdot (h+1)}{2} \tag{3}$$

where h is the distance from the pivot. For the rest of the genes, a w_k of 1 is assigned. When two ranked lists are compared, a gene k has a different w_k in each ranked list and the smaller one is chosen.

Thus, in the weighted version of R2KS, $R'_{i,j}$ is defined analogously to before:

$$R'_{i,j} = \frac{\sum_{k \in A_{i,j}} w_k}{\sum_{k \in U} w_k} - \frac{i \times j}{N \times N} \tag{4}$$

And again, the R2KS score is defined to be the maximum of $R'_{i,j}$.

2.4. Algorithm

The R2KS score is the maximum of equation 4. We give a simple dynamic programming algorithm that computes rectangles from lower left to upper right. Unfortunately, we cannot economize and restrict attention to the actually occuring dots for candidate top right corners of a rectangle. For example, let two genes fall at the points [2,7] and [9,3], respectively. The entries [2,7] and [9,3] are surely candidates, but [9,7] is a candidate too, because it is the smallest rectangle that covers both genes. Therefore, our algorithm

calculates every entry of $R'_{i,j}$ using dynamic programming in $O(N^2)$, which is the same order of the number of the candidates in the worst case:

Algorithm A simple dynamic programming algorithm that calculates the R2KS score

---

1. Let $R := 0$.
2. Calculate the sum of weight of all genes $C := \sum_{k \in U} w_k$ by simple iteration.
3. Maintain a matrix $M = [0, \ldots, N][0, \ldots, N]$ with initial value $\forall i, M[0][i] := M[i][0] := 0$.
4. For each entry of M, begin loop:
5. $M[i][j] := \begin{cases} M[i-1][j-1] + w_k & \text{if gene } k \text{ falls at point } (i,j) \\ M[i-1][j] + M[i][j-1] - M[i-1][j-1] & \text{Otherwise} \end{cases}$
6. If $R \; R < \frac{M[i][j]}{C} - \frac{i \times j}{N \times N}$, then $R := \frac{M[i][j]}{C} - \frac{i \times j}{N \times N}$
7. End of loop
8. Return R

In the algorithm, we assert $M[i][j]$ to be $\sum_{k \in A_{i,j}} w_k$, which can be proved by mathematical induction. The time complexity of the algorithm is dominated by calculating each entry of M. Since each step in the loop can be done in $O(1)$, the total time complexity is $O(N^2)$.

3. RESULTS AND DISCUSSION

3.1. Simulation scenario

We first evaluate the behavior of the R2KS-score by applying it to simulated data. Rather than perturbing permutations, we perturb the underlying experimental data. Consider a microarray experiment where two conditions are compared using several replicates per condition. We will simulate pairs of lists at varying degrees of perturbation of differential expression. Emphasis is placed on maintaining stability at the top of the list, while perturbing the values more strongly as one goes down the list. Then, these lists will be converted into ranked gene lists in order to study the effectiveness of comparing ranked lists.

We use an experiment from the literature (Lamb et al., 2006) (build02, batch 767, tanespimycin) as our source of real data. The degree of change of a gene's expression is usually measured using a t-statistic. The gene has an observed mean and variance under each of the two conditions and the replicates provide us with a variance. Let the difference between the means be m, and the estimate of the variance be σ^2. To obtain simulated data we want to draw the perturbed values for each gene from a t-distribution with mean m and standard deviation σ, and let df be the degree of freedom for the t-distribution. Since in some cases the standard deviation may be very small—which contradicts our intention to produce randomly perturbed measurements—we define a new standard deviation σ' for each gene as $\sqrt{\bar{\sigma}^2 + \sigma^2}$, where $\bar{\sigma}$ is the average over the standard deviations of all genes. An increased perturbation is then implemented by multiplying σ' with a fudge factor x. Then we sample the t-values of each gene from a t-distribution $(m, x \cdot \sigma', df)$, and order them to get a new ranked list. Ties are broken randomly. Such generated ranked lists have the desired similarity and will be used as standard to compete with different methods. However, the independent sampling of each gene loses certain information like co-expression of genes, and as a result the simulation data lack some biological depth.

To verify that ranked lists generated by this simulation scenario show the desired behavior, Figure 3 displays scatter plots for three pairs of ranked lists when the fudge factor x goes from 1, through 3 to 10.
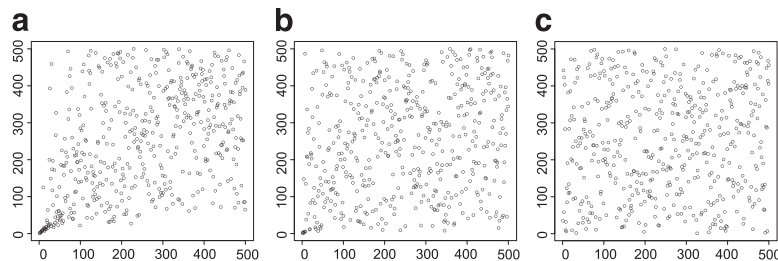
FIG. 3. Three examples of comparing randomly perturbed lists under fudge factors 1 (**a**), 3 (**b**), and 10 (**c**).

3.2. Comparing different measures

We now compare the R2KS score with previous methods of comparing ranked lists. The alternative approaches as given in the Introduction are the hypergeometric model for overlap probability (Eden et al., 2007; Fury et al., 2006; Roider et al., 2009) and the overlapping score due to Yang et al. (2006) implemented in the package *OrderedList* (Lottaz et al. 2006). Default parameter settings from this package are used. To include a method that is not based on ranking, the Euclidean distance between the original data vectors is also employed in the comparison. This is used in microarray analysis after normalization between experiments (Hughes al., 2000). We execute all comparisons among 100 simulated lists with varying degrees of randomness, leading to 4950 comparisons per setting. Table 1 gives the mean and standard deviation of each group of tests.

The results show that generally the measures are monotonic in the fudge factor, with the slight exception of the overlapping score for very high fudge factor. The hypergeometric score distinguishes well for low fudge factors but resolves badly for higher ones, especially in light of the high standard deviation there. Likewise, in the overlapping score the standard deviation for higher fudge factors blurs the distinction between the averages. R2KS both in the un-weighted and the weighted version has a low standard deviation and resolution is maintained up to a fudge factor of 20. The Euclidean distance performs as well as R2KS, albeit using the actual value of log fold change rather than ranks. It is reassuring that the information loss upon moving to ranks seems to be limited.

From amongst these three methods for ranked list comparison, the iterated hypergeometric test and the R2KS-score are more similar to each other in their design. Both R2KS and iterated hypergeometric score assess the amount of surprise in the overlap between top-ranking gene sets. One class of method does this using the hypergeomtric p-value, the other using the difference between observed and expected amount of overlap. Both approaches vary the cutoff to look for the most significant overlap. The overlapping score in contrast integrates over the overlap counts after assigning exponential weights to the counts.

3.3. Application to clustering of biological data

To test whether the R2KS-score is a reasonable measure on biological data, we compare microarray clustering results produced produced with R2KS-score to a published clustering from the literature. Since it is so well studied we utilize the yeast compendium data (Hughes et al., 2000). A hierarchical clustering based on Euclidean distance among 127 experiments is published in the supplementary material of the original article. For comparison, we have computed a hierarchical clustering based on the pair-wise R2KS score of the same set of experiments. Our tree is shown as Supplementary Figure 1 (see online Supplementary Material at www.liebertonline.com/cmb), and the original tree can be found in the supplementary material of Hughes et al. (2000). Although it is notoriously difficult to compare two hierarchical trees, visual inspection of the two trees shows considerable agreement. For example, spotting hybridization *aep2* in the Hughes tree shows its closest neighbors as *ymr293c, msul, yer050c,* and *rml2(**13)*. In our R2KS based tree, its neighbors are exactly the same. Although only one anecdote, this indicates that in principle rank based clustering can obtain similar results to accepted, score-based clustering.

TABLE 1. COMPARISON AMONG METHODS ON SIMULATION DATA

	$\sqrt{N} \cdot R2KS$		$\sqrt{N} \cdot R2KS(W)$		Hypergeometric		Overlappingscore		Euclidean	
x	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	12.813	0.199	69.868	0.757	565.831	23.337	52129.58	680.448	24.492	0.222
2	6.474	0.220	40.508	0.864	129.186	10.504	23646.36	700.536	35.579	0.325
5	2.164	0.202	19.316	0.816	10.109	3.011	5851.804	540.414	43.459	0.225
10	1.025	0.173	14.693	0.664	5.775	2.452	3598.595	473.833	44.770	0.216
20	0.709	0.145	13.742	0.628	5.547	2.382	3313.147	460.020	45.059	0.208
50	0.619	0.134	13.561	0.632	5.597	2.495	3311.441	462.963	45.129	0.214
100	0.611	0.133	13.552	0.625	5.664	2.413	3322.183	471.854	45.134	0.216

Next we turn to comparing across different gene expression technologies. With the recent improvements in the efficiency, quality and cost of massively parallel sequencing, RNA-seq is now an alternative way for high-throughput studies of gene expression. However, the statistical model for RNA-seq is very different from that for microarray experiments making the direct comparison between RNA-seq and microarray results difficult. Ranked gene lists are a possible bridge to connect these two different technologies and we exemplify such a connection by investigating tissue-specific expression experiments from both technologies: Microarray experiments from the Novartis gene atlas done with Affymetrix Human Genome U133A arrays (Su et al., 2004) and RNA-seq experiments done with Illumina Genome Analyzer (Wang et al., 2008). The RNA-seq data is mapped to Affymetrix probe sets in the U133A array and RPKM normalization is used. Additionally a pseudo count is added to the expression levels to mimic the background noise in microarray experiments.

The benefit of tissue-specific expression experiments is that, independent of the technological platt-form, one expects the same genes to mark a certain tissue in both experiments. This is why we use such a data set as a benchmark to evaluate the comparison measure. However, with this design, we lack the control group to determine the differentially expressed genes that best characterize the tissue that they are supposed to represent. To compromise, an artificial control tissue is defined by taking the median expression of each gene within one platform. The ranked lists are then ordered by the fold change between tissue-specific expression and expression in the artificial control tissue. Ties are broken randomly. We present the result for weighted R2KS, while the un-weighted version is shown in Supplementary Figure 2.

The lists for each of the two technologies show a strong within-plattform similarity. Therefore, we concentrate on the between-technology comparisons. Figure 4 shows the similarities between Illumina (RNA-seq) experiments on the horizontal axis and Affymetrix experiments in the rows. Rows and columns are each arranged according to a clustering, such that similar rows and similar columns, respectively, are displayed close to each other. On the bottom left of the matrix there is a dark cluster linking two Illumina brain profiles with a number of Affy brain experiments. One column further to the right the Illumina Lymph node experiment shows strong similartiy to a group including Affy lymphnode together with bone marrow, salivary gland, thymus and tonsils. Further on to the right, the Illumina testis profile is matched up with a number of testis related tissues generated with Affymetrix arrays. The next adjacent column to the right contains a liver cluster. The following column is annotated MAQC UHR which refers to "universal human control." It happens to show similarity to liver and lung. Next, skeletal muscle and heart rows and columns match up correctly. Adipose tissue and colon from Illumina show little similaritiy to any of the Affy lists. Lastly, on the top right there is a large cluster linking healthy and tumour breast tissue and cell lines from Illumina with a number of cancer cell lines and immune system cells from Affymetrix. The plot generated with the unweighted version of R2KS shown in the Supplementary Material contains essentially the same results, although it is less clear and harder to interpret. Overall, however, this analysis shows clearly that even after the gene lists from different technologies are reduced to their mere ranks, the R2KS-score is still capable of recovering their similarity.

4. CONCLUSION

We have designed the R2KS-score in order to measure the similarity between ranked lists of genes, where we expect to see an accumulation of the same genes at the top of each list. The order of the genes further down is unimportant to the comparison. Since *a priori* no cutoff is known, we take an approach that tests for a good cutoff. We have tried to show that our measure captures this intuition. By applying it to a comparison between microarray generated expression data and RNA-seq data, we have given an example that the rank-based comparison is useful.

The R2KS-score draws on the intuition of a two-dimensional Kolmogorov-Smirnov statistic and constitutes a similarity score between ranked lists. In the un-weighted case, the simulation shows that the distribution of the R2KS-score for lists of different length is almost the same after normalization by multiplying by $\sqrt{N}$ (Fig. 2), which is exactly the same property as in the Kolmogorov-Smirnov statistic. Unfortunately, for the weighted case we do not see a way to make a similar statement, since the influence of the exact weighting scheme would need to be accounted for. Therefore, one needs to resort to separate simulations for each length in order to determine statistical significance of the similarity of two ranked lists

FIG. 4. Heatmap of ranked lists from different platforms.

under the weighted R2KS-score. Nevertheless, introducing a weighting helps to interpret the ranked list correctly, because the design of biological experiments cares only for the most differentially expressed genes, which amount to 5–10% of total genes, and other genes are often not truly randomly distributed at the bottom or even in the middle of the ranked lists.

Ranked lists are not unique to gene expression. Indeed, many biological results come in the form of ranked lists of genes. A typical example is database searching with BLAST, which results in a list of similar sequences ranked by their p-value. Moreover, different web search engines have also been compared through the ranked lists of the URLs they identified (Jacso, 2005). The common feature of ranked lists is that the items in the list become less and less important and degenerate into random noise. R2KS appropriately models this behaviour and we show that it can be a useful tool to compare ranked lists across different technological platforms or sources.

ACKNOWLEDGMENTS

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Boulesteix, A.L., and Slawski, M. 2009. Stability and aggregation of ranked gene lists. *Brief Bioinform.* 10, 556.

Eden, E., Lipson, D., Yogev, S., et al. 2007. Discovering motifs in ranked lists of DNA sequences. *PLoS Comp. Biol.* 3, e39.

Ein-Dor, L., Zuk, O., and Domany, E., 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA* 103, 5923.

Fasano, G., and Franceschini, A. 1987. A multidimensional version of the Kolmogorov-Smirnov test. *Mon. Not. Roy. Astron. Soc.* 225, 155–170.

Fury, W., Batliwalla, F., Gregersen, P.K., et al. 2006. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*

Hughes, T.R., Marton, M.J., Jones, A.R., et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126.

Jacso, P. 2005. Visualizing overlap and rank differences among web-wide search engines. *Online Inform. Rev.* 29, 554–560.

Jurman, G., Merler, S., Barla, A., et al. 2008. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics* 24, 258.

Lamb, J., Crawford, E.D., Peck, D., et al. 2006. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935.

Lottaz, C., Yang, X., Scheid, S., et al. 2006. OrderedList—a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics* 22, 2315.

Peacock, JA. 1983. Two-dimensional goodness-of-fit testing in astronomy. *Mon. Not. Roy. Astron. Soc.* 202, 615–627.

Qiu, X., Xiao, Y., Gordon, A., et al. 2006. Assessing stability of gene selection in microarray data analysis. *BMC Bioinform.* 7, 50.

Roider, H.G., Manke, T., O'Keeffe, S., et al. 2009. PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* 25, 435.

Su, A.I., Wiltshire, T., Batalov, S., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101, 6062.

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

Wang, E.T., Sandberg, R., Luo, S., et al. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456 470–476.

Yang, X., Bentink, S., Scheid, S., et al. 2006. Similarities of ordered gene lists. *J. Bioinform. Comput. Biol.* 4, 693.

Zucknick, M., Richardson, S., and Stronach, E.A. 2008. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Stat. Appl. Genet. Mol. Biol.*, 7, article 7.

Address correspondence to:
Dr. Shengyu Ni
CAS-MPG Partner Institute for Computational Biology
Shanghai Institutes of Biological Sciences
Yueyang Road 320
200031 Shanghai, China

E-mail: sendru@picb.ac.cn