RECIPROCAL RANK-BASED COMPARISON OF ORDERED GENE LISTS

Ronald K. Pearson

ProSanos Corporation
Harrisburg, PA, USA
ronald.pearson@prosanos.com

ABSTRACT

There is growing interest in using rank-ordered gene lists to avoid excessive dependence on measured gene expression levels, which can vary strongly across experiments, platforms, or analysis methods. As a useful tool for working with these lists, this paper describes two extensions of an ordered list comparison measure, recently proposed for comparing Internet search engines: the use of random permutations to assess the significance of differences between ordered lists, and a graphical extension that highlights the items responsible for the main differences between two lists. The method is illustrated for a prostate cancer example from the genomics literature.

1. INTRODUCTION

Gene expression results obtained by microarray analyses can vary widely as a function of platform [8], data analysis method [2], and investigator [3]. This variability motivates the use of rank-ordered gene lists [2, 7], which are invariant to monotone intensity transformations (e.g., logarithmic transformations) and reduce the influence of outliers [6, Sec. 2.2.4]. As an aid to working with rank-ordered gene lists, this paper describes a simple comparison tool for ordered lists, based on recent results of Bar-Ilan *et al.* [1]. Specifically, the following sections briefly describe their approach (Sec. 2), along with two extensions (permutation-based significance assessments, in Sec. 3, and a graphical diagnostic for identifying the sources of discrepancies between lists, in Sec. 4) that we have found useful in comparing drug safety data mining results. Here, these methods are used to compare gene expression results from different investigators, using different platforms.

2. THE RECIPROCAL-RANK MEASURE

Many methods exist for comparing ordered lists [4, 5], but a key issue in comparing gene lists is that the genes included in the two lists are frequently not identical, leading to a *partial ranking problem*. As Marden notes, some standard methods "adapt well to some types of partial rankings, but none adapt well to all types, and some to no types" [5, p. 4]. Bar-Ilan *et al.* consider four methods for comparing the results of Internet search engines [1], an application that poses a similar partial ranking problem. In particular, two different Internet search algorithms typically return different ordered lists of search results in response to the same query. Since these lists are frequently quite long (e.g., a Google query on "comparing rankings" finds 3,710,000 matches), users are most interested in a subset of the highest-ranked responses, which *should be* the "most relevant." Unfortunately, as in the microarray interpretation problem [2], the Internet page ranking problem is one for which there is no "gold standard" for assessing accuracy (i.e., "true relevance"). Consequently, rather than attempting to address the question of which result is "better," the authors address the related but simpler question of "how different are two results?"

Specifically, Bar-Ilan *et al.* consider four measures of similarity or dissimilarity between two ordered lists, $\mathcal{L}_A$ and $\mathcal{L}_B$, each representing the K highest-ranked URL's returned by one of the search engines compared. The first of these measures is the fractional overlap, defined as

$$O_{ab} = \frac{|\mathcal{L}_A \cap \mathcal{L}_B|}{K}, \tag{1}$$

where $|\mathcal{S}|$ denotes the number of elements in the set $\mathcal{S}$. Note that O_{ab} varies between 0, when the two lists are completely disjoint, and 1, when the two lists rank the same K objects. As the authors note, a key limitation of this comparison measure is that it makes no use of the rank data: two lists of the same K objects exhibit $O_{ab} = 1$, regardless of whether these objects are ranked in the same order, the opposite order, or in any other order. The other three approaches considered by Bar-Ilan *et al.* do make use of rank information, and the one that the authors ultimately recommend is based on the following extended rank idea of Fagin *et al.* [4]. For every object in the union of the two ranked lists, a modified rank with respect to list $\mathcal{L}_A$ is defined as:

$$\tilde{R}_A(i) = \begin{cases} R_A(i) & \text{if } R_A(i) \leq K \\ K+1 & \text{if } R_A(i) > K \\ & \text{or } R_A(i) \text{ is undefined,} \end{cases} \tag{2}$$

with the analogous definition for the modified rank $\tilde{R}_B(i)$ with respect to list $\mathcal{L}_B$. Given these modified ranks, Bar-Ilan *et al.* define the *reciprocal-rank dissimilarity* between lists $\mathcal{L}_A$ and $\mathcal{L}_B$ as:

$$\Delta_{ab} = \sum_{i=1}^{M} \left| \frac{1}{\tilde{R}_A(i)} - \frac{1}{\tilde{R}_B(i)} \right|, \tag{3}$$

1-4244-0999-3/07/$25.00 ©2007 IEEE.

where M is the total number of elements in the union of the two lists. The advantage of reciprocal ranks over the ranks themselves in this measure is that differences between upper ranks (e.g., 1 vs. 4) are given more weight than the same absolute difference between lower ranks (e.g., 7 vs. 10).

3. ASSESSING SIGNIFICANCE

Since similarities between rankings are easier to interpret than dissimilarities, the results presented here are based on the following normalized similarity measure:

$$S_{ab} = 1 - \Delta_{ab}/\Delta_{ab}^{+}, \tag{4}$$

where Δ_{ab}^{+} is the maximum possible value for Δ_{ab}, achieved for two disjoint top-K lists and given by [1]:

$$\Delta_{ab}^{+} = 2\sum_{i=1}^{K}\left(\frac{1}{i} - \frac{1}{K+1}\right). \tag{5}$$

For two top-K lists, this similarity measure lies between a minimum value of 0, if and only if the two lists are disjoint, and a maximum value of 1, if and only if they are identical.

Despite the inherent advantages of a normalized measure, the question of interpretation remains: does a given value of S_{ab}, say 0.40, suggest weak, moderate, or strong similarity? To address this question, the following randomization strategy is proposed: given the modified rank vectors $\{\tilde{R}_A(i)\}$ and $\{\tilde{R}_B(i)\}$, compute the similarity measure $S_{a\pi b}$ between $\{\tilde{R}_A(i)\}$ and a random permutation π of $\{\tilde{R}_B(i)\}$. If the original modified ranks exhibit significant similarity, this should be destroyed by the permutation, giving a substantially smaller value for $S_{a\pi b}$. Repeating this procedure Q times, for a collection $\{\pi_j\}$ of Q independent random permutations, then provides a range of reference values $\{S_{a\pi_j b}\}$ that can be used to interpret the original similarity measure S_{ab}. In particular, one informal significance measure of the result is the z-score:

$$z_{ab} = \frac{S_{ab}^{0} - \bar{S}_{ab}}{\sigma_{ab}}, \tag{6}$$

where S_{ab}^{0} is the original similarity result, $\bar{S}_{ab}$ is the mean of the Q random permutations, and σ_{ab} is their standard deviation. In general, the larger the magnitude of the z-score, the more significant the similarity value S_{ab}^{0}, relative to two randomly re-ordered lists. A more direct significance measure is the empirical probability E_{ab}, defined as

$$E_{ab} = \begin{cases} \frac{N_{>}+1}{Q+1} & \text{if } z_{ab} \geq 0, \\ \frac{Q-N_{>}}{Q+1} & \text{if } z_{ab} < 0, \end{cases} \tag{7}$$

where $N_{>}$ is the number of random permutation values $S_{a\pi_j b}$ that exceed the original similarity value S_{ab}^{0}.

To illustrate these ideas, consider the following example. DeConde *et al.* [3] consider published microarray results from five different research groups, each comparing normal to cancerous prostate samples. Three of these studies used custom-built spotted cDNA arrays and the other two used commercial oligonucleotide arrays. The authors present the top 25 up-regulated gene lists from each study, noting that "of the 89 genes that appear in the top-25 up-regulated genes in at least one list, only 23 appear in more than one list and only one gene, hepsin, appears in all five lists." Table 1 gives the results obtained from the reciprocal rank-based comparison method described here, for each of the ten possible pairs of these five top-25 lists, with each result based on 1000 random permutations. Note that the overlap values O_{ab} range from 4% (i.e., one gene of 25 common to both lists) to 44%, while the similarity values S_{ab} vary from 0.164 to 0.510. Further, it is clear that these two measures are only weakly related: both the smallest S_{ab} value (0.164) and the third-largest (0.413) correspond to the same moderate overlap value (0.32), while the first, sixth and ninth-ranked of the ten S_{ab} values (0.510, 0.335, and 0.247) all correspond to the smaller overlap $O_{ab} = 0.12$, representing three common genes out of 25.

Pair	O_{ab}	S_{ab}	z_{ab}	E_{ab}
1,2	0.12	0.510	4.396	0.004
1,3	0.12	0.247	1.086	0.138
1,4	0.12	0.335	2.055	0.036
1,5	0.04	0.337	2.255	0.034
2,3	0.44	0.426	3.083	0.020
2,4	0.20	0.397	2.900	0.024
2,5	0.32	0.413	2.798	0.031
3,4	0.20	0.272	1.319	0.090
3,5	0.32	0.164	-0.250	0.498
4,5	0.28	0.274	1.185	0.105

Table 1. Comparison of the top 25 gene lists from five prostate cancer studies, from [3].

4. A GRAPHICAL DIAGNOSTIC TOOL

To understand results of this kind—i.e., to explain either why two lists with large overlaps appear strongly dissimilar or why two lists with small overlaps appear strongly similar—consider the following diagnostic measure, defined for each object i in the union of the two top-K lists:

$$\Delta_{ab}(i) = \left(\frac{K+1}{K}\right)\left[\frac{1}{\tilde{R}_A(i)} - \frac{1}{\tilde{R}_B(i)}\right], \tag{8}$$

which varies between -1 (if object i has rank 1 in list $\mathcal{L}_B$ but is not in the top K elements of list $\mathcal{L}_A$) and $+1$ (if object i has rank 1 in list $\mathcal{L}_A$ but is not in the top K elements of list $\mathcal{L}_B$). Note that if object i has the same rank in both lists, $\Delta_{ab}(i) = 0$.

The main diagnostic tool proposed here (illustrated in Fig. 1) is a plot of $\Delta_{ab}(i)$ against i, where the genes have been ordered so that first, $R_A(i) = i$ for $i = 1, 2, \ldots, K$

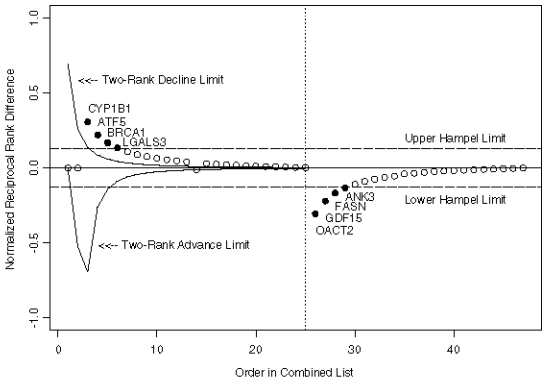

Figure 1. Reciprocal rank-based graphical diagnostic plot comparing the most similar pair of top 25 up-regulated gene lists for the prostate cancer example.

and second, if $M > K$, then $R_B(i)$ is in ascending order for $i = K+1, \ldots, M$. Under this ordering scheme, it is possible to construct two useful reference lines. The *two rank decline limit* is defined by the $\Delta_{ab}(i)$ values that arise when the ranks $R_A(i) = i$ decline by two units, to $R_B(i) = i + 2$. Similarly, the *two rank advance limit* is defined by the $\Delta_{ab}(i)$ values that arise when the ranks $R_A(i) = i$ advance by two units, to ranks $R_B(i) = i - 2$. Since no advance is possible when $R_A(i) = 1$, the two rank advance value for $i = 1$ is defined to be zero; similarly, since only a one unit advance is possible when $R_A(i) = 2$, the two rank advance value for $i = 2$ corresponds to the change from $R_A(2) = 2$ to $R_B(2) = 1$. The motivation for defining these limits is that identifying genes ranked first in list $\mathcal{L}_A$ but third in list $\mathcal{L}_B$ is generally of less interest than identifying those that list $\mathcal{L}_A$ ranks first but list $\mathcal{L}_B$ ranks much lower.

The diagnostic plots proposed here also include three other reference lines. One is a vertical line at $i = K$, separating those adverse events that are ranked in the top K in list $\mathcal{L}_A$ (corresponding to points lying to the left of the reference line) from those that are not (those points lying to the right of the reference line). The other two reference lines are outlier limits for the $\Delta_{ab}(i)$ values, corresponding to the upper and lower detection limits of the Hampel identifier with a threshold value of $t = 3$ [6, Sec. 1.4.2]. Points falling outside these detection limits are large enough in magnitude to be deemed "unusual" relative to the majority of the data points and are represented as solid circles, while points lying within these outlier detection limits are represented as open circles.

As a specific example, Fig. 1 shows the results obtained when the graphical diagnostic tool just described is applied to the two most similar of the five prostate cancer gene lists discussed in Sec. 3. As noted earlier, the small overlap value for these two lists ($O_{ab} = 0.12$) implies that these lists share only three of 25 genes in common, but the first and second ranked genes are the same in both lists, as indicated by the fact that the left-most two points in Fig. 1 fall on the $\Delta_{ab}(i) = 0$ line. In addition, another point falls very near this line, corresponding to a gene that is ranked 14^{th} in List 1 and ranked 12^{th} in List 2. The reason the similarity value S_{ab} is not higher than 0.501 for this case is that none of the other genes are common to both lists; in particular, the solid points in Fig. 1 correspond to genes ranked 3^{rd} through 6^{th} in each list, but absent from the other list. The statistical significance of this S_{ab} value, despite the small overlap between the two lists, reflects the basis for the rank product method for microarray data analysis advocated by Breitling *et al.* [2]: the same genes are unlikely to appear highly ranked in different lists by random chance alone.

5. SUMMARY

This paper has described a method for comparing ordered gene lists, based on a method proposed for comparing Internet search engines [1], with extensions to assess the significance of these similarity values (discussed in Sec. 3) and to identify specific genes responsible for the differences between two lists (discussed in Sec. 4).

6. REFERENCES

[1] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Computer Networks*, v. 50, 2006, pp. 1448–1463.

[2] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Letters*, v. 573, 2004, pp. 83–92.

[3] R.P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, and R. Etzioni, "Combining Results of Microarray Experiments: A Rank Aggregation Approach," *Statistical Applications in Genetics and Molecular Biology*, v. 5, 2006, article 15 (electronic journal).

[4] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing Top k Lists," *SIAM J. Discrete Math.*, v. 17, 2003, pp. 134–160.

[5] J.I. Marden, *Analyzing and Modeling Rank Data*, Chapman and Hall, 1995.

[6] R.K. Pearson, *Mining Imperfect Data*, SIAM, Philadelphia, 2005.

[7] J.D. Pylatuik and P.R. Fobert, "Comparison of Transcript Profiling on *Arabidopsis* Microarray Platform Technologies," *Plant Molecular Biology*, v. 58, 2005, pp. 609–624.

[8] C.L. Yauk, M.L. Berndt, A. Williams, and G.R. Douglas, "Comprehensive comparison of six microarray technologies," *Nucleic Acids Research*, v. 32, 2004, paper e124.