A framework for list representation, enabling list stabilization through incorporation of gene exchangeabilities

CHARLOTTE SONESON∗, MAGNUS FONTES
Centre for Mathematical Sciences, Lund University,
Box 118, SE-221 00 Lund, Sweden
lottas@maths.lth.se

SUMMARY

Analysis of multivariate data sets from, for example, microarray studies frequently results in lists of genes which are associated with some response of interest. The biological interpretation is often complicated by the statistical instability of the obtained gene lists, which may partly be due to the functional redundancy among genes, implying that multiple genes can play exchangeable roles in the cell. In this paper, we use the concept of exchangeability of random variables to model this functional redundancy and thereby account for the instability. We present a flexible framework to incorporate the exchangeability into the representation of lists. The proposed framework supports straightforward comparison between any 2 lists. It can also be used to generate new more stable gene rankings incorporating more information from the experimental data. Using 2 microarray data sets, we show that the proposed method provides more robust gene rankings than existing methods with respect to sampling variations, without compromising the biological significance of the rankings.

Keywords: Exchangeability; Gene ranking; List; Microarray.

1. INTRODUCTION

Since the advent of the microarray technology, high-throughput experiments generating vast amounts of data have been ubiquitous for studying, for example, genome-wide patterns of gene expression and copy number alterations. The output of univariate analysis of such high-throughput experiments is often a “gene list,” consisting of genes related to some response of interest. The gene list can be ordered or unordered and it can contain all studied genes or only a subset. The challenge is then to interpret the obtained list in a biological context to understand the underlying processes and generate biologically valid hypotheses. An inherent problem compromising the interpretability of the observed gene lists is that they are often highly unstable, both with regards to small changes in the underlying data set and with regards to changes in the ranking method (Fortunel and others, 2003; Irizarry and others, 2005; Michiels and others, 2005; Ein-Dor and others, 2006; Fan and others, 2006; Boulesteix and Slawski, 2009; Abraham and others, 2010). One reason for this instability can be that many genes have similar functions in the cell and are

*To whom correspondence should be addressed.
thereby exchangeable in a given experimental list. Thus, the observed gene list depends on the selection of samples in the data set. This means that the functional overlap between 2 lists may be substantial even if the actual gene overlap is small. Other causes of the instability can be measurement noise and small sample sizes (Ein-Dor and others, 2006; He and Yu, 2010).

In this paper, we propose a method for stabilization of observed gene rankings, using information extracted from the experimental data. We employ the concept of exchangeability of random variables to quantify the functional redundancy among the genes, and we propose a general framework for incorporating exchangeability into the representation of gene lists. In this framework, each list is represented as a vector in $\mathbb{R}^M$, where $M$ is the number of genes in some universal set, typically the genes measured by a microarray chip. Each entry of the list vector quantifies the contribution from one of the genes to the list. This representation allows straightforward comparison of any 2 gene lists by means of any of the conventional measures of similarity or dissimilarity defined on $\mathbb{R}^M \times \mathbb{R}^M$. This is in contrast to previously proposed methods for list comparison, which are tailored to compare specific types of lists. We show that using the proposed method, we obtain gene rankings that are more robust than the original lists against sampling variations in the underlying data, without compromising the relevance to the response.

2. RELATED WORK

The stabilization of gene rankings has attracted considerable interest during the last decade. Some authors have addressed the ranking method directly and proposed methods providing more robust and accurate ranking results and differential expression detection for the “large p, small N” situation, which is standard in biomedical applications (e.g. Tusher and others, 2001).

More stable rankings can potentially also be obtained by combining the information from several different rankings (e.g. Rhodes and others, 2002, 2004; Breitling and others, 2004; DeConde and others, 2006; Hong and Breitling, 2008; Abeel and others, 2010). An overview of the most well-known aggregation methods is given by Boulesteix and Slawski (2009). The most straightforward methods involve computing a univariate statistic for each gene from the set of rankings and reordering the genes by their value of this statistic. The statistic can be, for example, the mean of the positions for the gene (e.g. Jurman and others, 2008), a rank product of the positions (Breitling and others, 2004), or the fraction of the rankings where the gene is among the top-$k$ genes for some $k$ (e.g. Pepe and others, 2003; Jurman and others, 2008). More complex aggregation methods exist for extracting an optimal top-$k$ list, based on, for example, Markov chains (DeConde and others, 2006).

Comparison of gene lists is an essential part of many algorithms, for example, enrichment analysis of gene sets (e.g. Drăghici and others, 2003; Subramanian and others, 2005; Ein-Dor and others, 2006; Ackermann and Strimmer, 2009) and assessment of the stability of gene rankings (e.g. Jurman and others, 2008; Jurman and others, 2010; Abraham and others, 2010). In general, each existing method is adapted to compare specific types of lists, for example, 2 ordered lists with the same number of genes. The supplementary material available at Biostatistics online provides a brief overview of the most well-known list comparison methods and show that many of them can be cast in the framework proposed here.

3. METHODS

3.1 Exchangeability of random variables

Consider a probability triple $(\Omega, \mathcal{F}, P)$ and let $X_1, \ldots, X_m$ denote random variables on $\Omega$, taking values in some space $\mathbb{M}$. Given $X_1, \ldots, X_m$, we define the random variable

$$X_1 \times \ldots \times X_m : \Omega \rightarrow \mathbb{M} \times \ldots \times \mathbb{M}.$$
by $X_1 \times \ldots \times X_m(\omega) = (X_1(\omega), \ldots, X_m(\omega))$. To each random variable $X_1 \times \ldots \times X_m$, there is an associated measure $\Pr_{X_1 \times \ldots \times X_m}$ defined by

$$\Pr_{X_1 \times \ldots \times X_m}(A) = P((\omega \in \Omega; \ X_1 \times \ldots \times X_m(\omega) \in A)),$$

for all measurable subsets $A \subseteq \mathbb{M} \times \ldots \times \mathbb{M}$. Conventionally, a finite sequence $(X_1, \ldots, X_m)$ of random variables is called “exchangeable” if their joint distribution is invariant under permutation of $X_1, \ldots, X_m$, that is, if $\Pr_{X_1 \times \ldots \times X_m} = \Pr_{X_{\pi(1)} \times \ldots \times X_{\pi(m)}}$ for each $\pi \in S_m$ (the group of permutations of $\{1, \ldots, m\}$). This means that from a statistical point of view, the order of the variables in the product is completely irrelevant. From this definition, it is clear that any sequence of independent and identically distributed random variables is exchangeable but the reverse implication is false. For overviews on exchangeability, see for example, Kingman (1978) and Aldous (1985). We introduce a much weaker notion of exchangeability as follows:

**Definition 1** The finite sequence of random variables $(X_1, \ldots, X_m)$ is weakly exchangeable if the null sets of $\Pr_{X_1 \times \ldots \times X_m}$ are invariant under permutations, that is, if

$$\Pr_{X_{\pi(1)} \times \ldots \times X_{\pi(m)}} \ll \Pr_{X_{\tau(1)} \times \ldots \times X_{\tau(m)}}$$

for all $\pi, \tau \in S_m$. Here, $\mu \ll \nu$ denotes that the positive measure $\mu$ is absolutely continuous with respect to the positive measure $\nu$.

It is clear that a finite sequence of random variables $(X_1, \ldots, X_m)$ that is exchangeable is weakly exchangeable, but the opposite implication is false in general.

### 3.2 The exchangeability of genes

We will now use the general concept of exchangeability to define the exchangeability of a set of genes in a given experiment. We assume that we are given a universal set of $M$ genes, denoted $\mathcal{G} = \{g_1, \ldots, g_M\}$. We use the term “experiment” to denote a pair consisting of a set of experimental conditions (defining the population and the employed measurement methods) and a variable ranking method (e.g. a $t$-test). The sample space $\Omega$ consists of all possible rankings of the $M$ genes, and the random variables $X_1, \ldots, X_M: \Omega \to \{1, \ldots, M\} =: \mathbb{M}$ represent the ranking positions of the genes in $\mathcal{G}$. A finite set of genes $\{g_{i_1}, \ldots, g_{i_m}\}$ is said to be (weakly) exchangeable if and only if the corresponding sequence of random variables $(X_{i_1}, \ldots, X_{i_m})$ is (weakly) exchangeable.

Using the weak exchangeability concept, we focus only on the possible ranking positions for the genes, not on the probabilities of obtaining the different positions. In a practical situation, the exchangeabilities of genes will be estimated from finite (possibly small) collections of gene list replicates. Under these conditions, we will not be able to reliably estimate the probability distributions of $X_1, \ldots, X_M$. Therefore, we instead estimate the supports of the distributions.

### 3.3 Quantifying the degree of exchangeability

We will now discuss how to define the degree of weak exchangeability of a set of genes. A more elaborate discussion can be found in the supplementary material available at *Biostatistics* online. With notation as in Section 3.2, we have a discrete probability space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F} = 2^\Omega$ is a $\sigma$-algebra and $P: \mathcal{F} \to [0, 1]$ is a probability measure. The support of the random variable $X_1 \times \ldots \times X_m$ is defined by

$$\text{supp } X_1 \times \ldots \times X_m := \{(q_1, \ldots, q_m) \in \mathbb{M} \times \ldots \times \mathbb{M}; \ \Pr_{X_1 \times \ldots \times X_m}((q_1, \ldots, q_m)) > 0\}.$$
For discrete random variables, Definition 1 implies that \((X_1, \ldots, X_m)\) is weakly exchangeable if and only if \(\text{supp}(X_1 \times \ldots \times X_m) = \text{supp}(X_{\pi(1)} \times \ldots \times X_{\pi(m)})\) for all \(\pi \in S_m\). Therefore, the degree of weak exchangeability of a finite set of discrete random variables will be defined by comparing the supports of the joint distributions.

We will mainly focus on the exchangeabilities of gene pairs, for computational reasons and since this will allow us to represent the exchangeabilities in a matrix. Thus, for a given gene pair \((g_1, g_2)\), we define the distance between 2 sets \(A \times \{g_1, g_2\}\) and \(A \times \{g_2, g_1\}\) by \(\text{dist}_\rho(A, B) := \min_{a \in A, b \in B} \rho(a, b)\). Intuitively, we may argue that the relative rank order of a pair of genes that is highly exchangeable should be arbitrary. In order to stress this, we will define a “one-sided” exchangeability score for a gene pair. For a pair of random variables \((X_1, X_2)\), we first define a set-valued random variable on \(\Omega\) by

\[
R(X_1 \times X_2)(\omega) := \{(x, y) \in \mathcal{M} \times \mathcal{M}; \ \text{sign}(x - y) = \text{sign}(X_1(\omega) - X_2(\omega))\}.
\]

**Definition 2** The one-sided mean exchangeability distance for a pair of discrete random variables \((X_1, X_2)\) is defined by

\[
oED_{X_1 \times X_2}^{\text{mean}} := \frac{\mathbb{E}_{X_1 \times X_2}[\text{dist}_\rho(X_1 \times X_2, \text{supp} X_2 \times X_1 \cap R(X_1 \times X_2))]}{\rho((1, 2), (M - 1, M))},
\]

if \(\text{supp} X_2 \times X_1 \cap R(X_1 \times X_2)(\omega) \neq \emptyset\) for all \(\omega \in \Omega\) with \(\Pr_{X_1 \times X_2}(X_1 \times X_2(\omega)) > 0\) and \(\noED_{X_1 \times X_2}^{\text{mean}} = 1\) otherwise.

We note that the one-sided mean exchangeability distance attains only values in [0, 1]. This allows us to define a similarity measure (an “exchangeability score”) for a pair of genes by

\[
oES_{X_1 \times X_2}^{\text{mean}} = 1 - \noED_{X_1 \times X_2}^{\text{mean}}.
\]

Finally, we define a normalized exchangeability score by comparing \(\noES_{X_1 \times X_2}^{\text{mean}}\) to the corresponding value for pairs of random variables with some prespecified distribution representing a null hypothesis of random ranking. In this paper, the main focus is on weak exchangeability of pairs of discrete random variables, in which case it is natural to compare to a random variable \(Y_1 \times Y_2\) uniformly distributed on a set \(S \subseteq \mathcal{M} \times \mathcal{M}\) with cardinality equal to that of \(\text{supp} X_1 \times X_2\). The normalized exchangeability score is defined as follows:

**Definition 3** The normalized one-sided mean exchangeability score for a pair of discrete random variables \((X_1, X_2)\) is defined by

\[
\noES_{X_1 \times X_2}^{\text{mean}} = \left(\frac{\noES_{X_1 \times X_2}^{\text{mean}} - \noES_{Y_1 \times Y_2}^{\text{mean}}}{1 - \noES_{Y_1 \times Y_2}^{\text{mean}}}\right)_+,
\]

where \(Y_1 \times Y_2\) is a random variable uniformly distributed on a set \(S \subseteq \mathcal{M} \times \mathcal{M}\) with \(|S| = |\text{supp} X_1 \times X_2|\) and \((a)_+ = \max(a, 0)\).

### 3.4 Estimation of the exchangeability score

To estimate the exchangeability score for each pair of genes in a data set with respect to a given experiment, we consider \(B\) gene rankings obtained under the given experimental conditions. For example, to estimate the exchangeability scores with respect to sampling variations under otherwise similar experimental conditions, we use resampling methods to generate a collection of \(B\) data sets for which we compute gene rankings. From the \(B\) gene rankings, we construct a position vector \(S_i\) for each gene \(g_i\),
by collecting all positions of the gene in the B rankings. The elements of a position vector $S_i$ are then samples of the random variable $X_i$. Combining 2 position vectors $S_i$ and $S_j$ gives samples of the variables $X_i \times X_j$ and $X_j \times X_i$ that can be used to estimate $E[D_{X_i \times X_j}^{\text{mean}}]$. This estimation is discussed further in the supplementary material available at Biostatistics online.

4. A GENERAL FRAMEWORK FOR LIST REPRESENTATION AND COMPARISON

In this section, we present a general framework for list representation and comparison. The lists are represented as vectors in $\mathbb{R}^M$, where the entry in the $i$th position gives the contribution of gene $g_i$. The vector representation allows us to compare ordered and unordered gene lists within the same framework, using one of the many similarity or dissimilarity measures available to compare vectors in $\mathbb{R}^M$. This is an advantage compared to existing methods for list comparison, which are specifically designed to compare certain types of lists. The proposed framework also provides a way to determine which genes are most important for explaining the similarity between 2 lists represented by vectors $x$ and $y$, by comparing the values of $x_i y_i$ (see the supplementary material available at Biostatistics online). Finally, ordering the genes by their contributions to the vector defines a new gene ranking.

An ordered (unordered) gene list is an ordered (unordered) subset of the universal set of genes $G$. By defining a function $f : (\text{positions}, \text{exchangeabilities}, \text{reliability}, \ldots) \mapsto l_{\ell} \in \mathbb{R}^M$,

we use various characteristics of the list to create a vector representation.

4.1 General idea

Let $\ell \subseteq G$ denote a list. If $\ell$ is ordered and if gene $g_i$ is contained in $\ell$, we denote its position by $\pi_{\ell}(i)$. If $g_i \notin \ell$, we define $\pi_{\ell}(i) = 0$. For an unordered list, we let $\pi_{\ell}(i) := \chi_{\ell}(g_i)$, where $\chi_{\ell}$ is the characteristic function of the set $\ell$. Given a list $\ell$, we define a corresponding “list matrix” $G_{\ell}$ as the product of 3 basic $M \times M$ matrices:

$$G_{\ell} := A_{\ell} V_{\ell} W_{\ell}.$$  

The 3 basic matrices are designed to represent different characteristics of $\ell$. We call $A_{\ell}$ the “position matrix,” $V_{\ell}$ the “exchangeability matrix,” and $W_{\ell}$ the “global weight matrix.” From the list matrix, we form a vector representation $l_{\ell} := ((l_{\ell})_1, \ldots, (l_{\ell})_M)$ of $\ell$ by letting $(l_{\ell})_i := h((G_{\ell})_i)$, where $(G_{\ell})_i$ denotes the $i$th column of $G_{\ell}$ and $h : \mathbb{R}^M \to \mathbb{R}$ is a summarization function, for example, a norm. Once all lists of interest are represented by vectors in $\mathbb{R}^M$, we can easily define the similarity between them, for example, as the cosine of the angle between the corresponding list vectors, that is,

$$s(\ell_1, \ell_2) = \frac{l_{\ell_1} \cdot l_{\ell_2}}{\|l_{\ell_1}\|_2 \|l_{\ell_2}\|_2},$$

where $\cdot$ denotes the inner product in $\mathbb{R}^M$, and we can define a dissimilarity coefficient by $d(\ell_1, \ell_2) = 1 - s(\ell_1, \ell_2)$.

Choosing $A_{\ell}$, $V_{\ell}$, $W_{\ell}$, and the (dis)similarity measure on $\mathbb{R}^M$ suitably, most methods currently available for list comparison fit into this general framework. In the supplementary material available at Biostatistics online, we show how this can be done for a collection of well-known methods.
4.2 The position matrix $A_\ell$

The position matrix $A_\ell$ is defined as a diagonal matrix that contains information about the type of list (ordered or unordered) and the positions of the genes within the list. We define the diagonal element $(A_\ell)_{ii}$ (the position value of gene $g_i$) via a monotone transformation of the ranking statistic of the gene. This will give the diagonal elements corresponding to genes in the top of the list $\ell$ high values, while those corresponding to genes further down in the list obtain lower values. All genes not in the list are given position value 0. In some cases, other choices of position values may be better suited for unordered lists, where it may be desirable to give the genes different weights, for example, depending on some external criterion, even though there is no specified ordering.

4.3 The exchangeability matrix $V_\ell$

The exchangeability matrix $V_\ell$ carries information about the exchangeability of the genes in $G$ in the specific experiment giving the list $\ell$. In this paper, we define the entry $(V_\ell)_{ij}$ to be the estimated normalized one-sided mean exchangeability score of $g_i$ and $g_j$ (i.e. $\hat{\text{noES}}_{X_i \times X_j}$), so the diagonal elements are always 1. If $V_\ell$ is diagonal, that is, $V_\ell = I_M$, then the only nonzero elements in the list vector are those corresponding to genes that are actually contained in $\ell$, and consequently, only the genes that are present in the list affect its vector representation. However, if $V_\ell$ is not diagonal, there is a possibility that the vector representation of the list is extended, meaning that it contains nonzero entries for genes that are not themselves present in the list but are exchangeable with some gene in the list. The (absolute) weight of a gene in the list can also be increased if it is highly exchangeable with a gene with a higher (absolute) position value since the high exchangeability indicates that under the given experimental conditions, the genes could as well have switched positions in the list.

The general framework for list representation supports any matrix of gene similarities in the place of $V_\ell$. One option that we will study is to use the positive part of the correlation matrix in place of $V_\ell$ since a high correlation between the expression levels of 2 genes is often considered to indicate similar biological functions of the genes.

4.4 The global weight matrix $W_\ell$

The global weight matrix $W_\ell$ is a diagonal matrix that permits weighting the influence of the genes differently, depending on some informativeness or reliability estimate. For example, we may wish to downweight the influence of a gene that has a high probability to be present in an arbitrarily chosen list.

5. Applications

5.1 Data sets

We illustrate the proposed methods by applying them to 2 microarray data sets.

The Boston lung cancer data (Bhattacharjee and others, 2001) contains expression measurements of 5217 genes in 62 lung cancer patients, classified according to outcome (good or poor) with 31 patients in each group. The data were downloaded from http://www.broadinstitute.org/gsea/datasets.jsp. The downloaded data set has already been preprocessed by replacing the original probe set IDs with gene symbols and summarizing all probe sets mapping to the same gene by the largest value for each sample (Subramanian and others, 2005).

The Hedenfalk breast cancer data (Hedenfalk and others, 2001) contains expression measurements of 3226 genes in 15 breast cancer patients, with either a BRCA1 mutation ($N = 7$) or a BRCA2 mutation ($N = 8$). The downloaded data were log transformed before the analysis.
5.2 Stabilization of ranked gene lists

The main objective for introducing exchangeabilities into the list representation is to increase the robustness of the resulting gene list. In this section, we evaluate the stability of the extended gene lists with respect to sampling variations. We apply the following experimental design to each of the 2 data sets:

1. We generate 10 modified data sets by bootstrapping samples from the original data set, taking the class labels into account.

2. For each modified data set, we rank the genes using 6 different methods. The basic method ranks the genes by their signal-to-noise ratios (SNR) for comparing the 2 sample groups in the data set. The resulting ranking is compared to rankings from 2 methods developed to give more robust rankings in the small sample context, a rank aggregation method, and the method proposed in this paper, with 2 different choices of the exchangeability matrix $V_\ell$:

   a) (SNR) Rank the genes by their SNR. The SNR of the $i$th gene is defined by SNR($i$) = $(m_1^j - m_2^j)/(\sigma_1^j + \sigma_2^j)$, where $m_1^j$ and $\sigma_1^j$ denote the mean value and standard deviation of the expression values of gene $i$ in the $j$th sample group.

   b) (SAM) Rank the genes by their values of the Significance Analysis of Microarrays (SAM) statistic (Tusher and others, 2001), computed via the “st” R package.

   c) (Efron-t) Rank the genes by their values of the modified $t$-statistic proposed by Efron and others (2001), computed via the “st” R package.

   d) (MedianAggregated) Subsample the data set 100 times (keeping two-third of the samples from each group in each subsampled data set), rank the genes in each subsampled data set by their SNR values, and define the final ranking of the genes by ordering them according to their median position across the 100 rankings.

   e) (ExtendedSNR) Compute the extended list vector as described in Section 4 and rank the variables according to their contribution to the list vector. The position matrix is derived from the SNR-based ranking of the genes, by letting

   $$(A_\ell)_{ii} = \begin{cases} b^2 & \text{if } \text{SNR}(i) \geq 0 \\ \frac{b^2}{(M - \pi_\ell(i))^2 + b^2} & \text{if } \text{SNR}(i) < 0 \end{cases}$$

   where $b^2 = 350$ and $M$ is the number of genes in the data set. We comment on the selection of this function and the parameter values in the supplementary material available at Biostatistics online. The normalized one-sided exchangeability scores are derived from position vectors computed by subsampling the original data set $B = 20$ times (each time keeping two-third of the samples from each group) and ranking the variables by their SNR values. We take the global weight matrix $W_\ell = I_M$. To create the vector representation from the list matrix, we define the $i$th entry of the list vector as the element with the largest magnitude in the $i$th column of the list matrix $G_\ell$. This means that a gene which is strongly exchangeable with a gene with a highly negative position value can be moved downwards in the list, so the 2 extreme ends are treated symmetrically.

   f) (CorrelationExtendedSNR) Compute the extended list vector as described in Step 2e but use the positive part of the correlation matrix of the original data set in place of the exchangeability matrix.

3. The correspondence between the lists from different bootstrap replicates are visualized through concordance plots. For each of the 6 ranking methods described in Step 2, let $f_k$ denote the number of genes that are among the top-$k$ in the resulting lists from all bootstrap replicates. The concordance
plot depicts $f_k$ as a function of $k$ for $k \in \{1, \ldots, M\}$. If the lists are highly reproducible with respect to sampling variation in the underlying data, we get $f_k \approx k$ for all $k$. We also construct concordance plots for the reversed lists, that is, letting $f_k$ be the number of genes that are among the bottom-$k$ in all lists. Additional estimates of the ranking stability with the different methods can be found in the supplementary material available at *Biostatistics* online.

The top rows of Figures 1 and 2 show the concordance plots obtained from the 2 data sets. It is clear that the exchangeability extended lists are more stable than the lists obtained by the other methods with

---

**Fig. 1.** Concordance plots for gene lists obtained by the 6 methods described in Section 5.2, applied to the Boston data set. The top row shows concordance plots for the observed data (left panel: top genes, right panel: bottom genes), and the bottom row shows corresponding plots for data where the class labels have been randomly permuted, so that the gene rankings from different bootstrapped data set are unrelated. The curves corresponding to the SNR ranking and the MedianAggregated ranking coincide almost completely in all cases, indicating that the ranking obtained by aggregation is not more stable than the original rankings with respect to sampling variations in the underlying data set. The extended lists provide a more stable gene ranking for the observed data. Interestingly, using the positive part of the correlation matrix to extend the gene lists gives less stable rankings. The modified $t$-statistics (SAM and Efron-t) provide slightly more stable gene rankings than the SNR.
Fig. 2. Concordance plots for gene lists obtained by the 6 methods described in Section 5.2, applied to the Hedenfalk data. The top row shows concordance plots for the observed data (left panel: top genes, right panel: bottom genes), and the bottom row shows corresponding plots for data where the class labels have been randomly permuted, so that the gene rankings from different bootstrapped data set are unrelated. The curves corresponding to the SNR, SAM, Efron-t, and MedianAggregated methods coincide almost completely. The extended lists provide a more stable gene ranking for the observed data. In this case, using the positive part of the correlation matrix to extend the gene lists gives a stabilization effect, but it is smaller than for the exchangeability extended lists.

respect to sampling variations in the underlying data set. Notably, the correlation-extended lists are less stable than the exchangeability extended lists for both data sets, indicating that the correlations in this case may not capture the relevant characteristics of the data. The bottom rows in Figures 1 and 2 show corresponding concordance plots for gene lists extracted from bootstrap replicates where the sample labels have been randomly permuted. These figures show that the stability of the extended lists that was noted in the top row is clearly dependent on that the gene lists actually share some information. Hence, the stabilization effect is not due to spurious features unrelated to the discrimination between patients from the different groups.
5.3 Relevance of ranked gene lists

Although stability of gene rankings is an important and desirable property, it is not the only thing that is of interest. For example, if we define a ranking method which always assigns a gene the same predefined position, the ranking would be extremely stable but most likely useless. We therefore study the informativeness of the rankings obtained as described above by examining the ability of the top-ranked genes in each list to discriminate between the 2 patient groups in the respective data sets. We use cross-validation (10-fold for the Boston data set, 5-fold for the smaller Hedenfalk data set) to assess the performance of the classifiers. For each training/test set split, we compute the 6 gene rankings as described in Section 5.2 for the training set and extract the top- and bottom-$k$ genes from each ranking. The expression levels for these genes are centered and standardized based on their mean value and standard deviation in the training set. The standardized expression levels of the selected genes are then used as features in a centroid classifier (Schölkopf and Smola, 2002), which is used to classify the remaining (test) samples. The reported classification accuracy is the mean area under the receiver operating characteristic curve across the training/test set splits.

Tables 1 and 2 show the estimated classification accuracy, for each of the 2 data sets, of the top and bottom genes from the 6 rankings as well as the mean classification accuracy for top and bottom genes from 20 random rankings. Note that the top-ranked gene is always the same for the “Extended SNR.”

Table 1. Classification accuracy (mean area under the receiver operating characteristic curve [AUC] across 10 training/test set splits) for the union of the top- and bottom-$k$ lists obtained from different rankings of the genes in the Boston lung cancer data with respect to their association with the discrimination between patients with good and poor outcome

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 10$</th>
<th>$k = 30$</th>
<th>$k = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExtendedSNR</td>
<td>0.344</td>
<td>0.774</td>
<td>0.837</td>
<td>0.832</td>
</tr>
<tr>
<td>SNR</td>
<td>0.344</td>
<td>0.583</td>
<td>0.606</td>
<td>0.566</td>
</tr>
<tr>
<td>Efron-t</td>
<td>0.344</td>
<td>0.492</td>
<td>0.607</td>
<td>0.572</td>
</tr>
<tr>
<td>SAM</td>
<td>0.344</td>
<td>0.492</td>
<td>0.607</td>
<td>0.572</td>
</tr>
<tr>
<td>CorrelationExtendedSNR</td>
<td>0.344</td>
<td>0.594</td>
<td>0.572</td>
<td>0.594</td>
</tr>
<tr>
<td>MedianAggregated</td>
<td>0.410</td>
<td>0.565</td>
<td>0.617</td>
<td>0.566</td>
</tr>
<tr>
<td>Random</td>
<td><strong>0.464</strong></td>
<td>0.489</td>
<td>0.473</td>
<td>0.478</td>
</tr>
</tbody>
</table>

The best performing method for each $k$ is highlighted in bold.

Table 2. Classification accuracy (mean AUC across 5 training/test set splits) for the union of the top- and bottom-$k$ lists obtained from different rankings of the genes in the Hedenfalk breast cancer data with respect to their association with the discrimination between patients with BRCA1 and BRCA2 mutations

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 10$</th>
<th>$k = 30$</th>
<th>$k = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExtendedSNR</td>
<td><strong>0.600</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>SNR</td>
<td><strong>0.600</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Efron-t</td>
<td>0.500</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>SAM</td>
<td>0.500</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>CorrelationExtended SNR</td>
<td><strong>0.600</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>MedianAggregated</td>
<td><strong>0.600</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Random</td>
<td>0.575</td>
<td>0.808</td>
<td>0.885</td>
<td>0.915</td>
</tr>
</tbody>
</table>

The best performing method for each $k$ is highlighted in bold.
“CorrelationExtendedSNR,” and “SNR” rankings. In the Boston data set, the classification ability of the top-ranked genes in the exchangeability extended list vectors is considerably higher than for the other methods, indicating that the increased stability observed in the previous section does not come at the expense of decreased biological significance. For the Hedenfalk data, all ranking methods (even random selection of genes) provide good classification.

6. DISCUSSION

Univariate analysis of high-throughput data sets often results in a ranking of the variables, which is then interpreted to gain biological knowledge. However, these variable rankings are often highly unstable with respect to small changes in the underlying data set or the ranking method, and therefore, methods for stabilizing the variable ranking and allowing more robust comparison to other lists are much needed. In this paper, we have presented a general framework for robust representation and comparison of variable lists. The framework encompasses both ordered and unordered lists, which can therefore be compared on similar terms. Having a measure of similarity between any pair of variable lists can furthermore enable visualization through, for example, multidimensional scaling to obtain a low-dimensional visual representation of large collections of lists. We have shown that the proposed method provides gene lists with advantageous reproducibility across bootstrapped data sets. These results suggest that the exchangeability concept for random variables may be a suitable tool for quantifying the functional redundancy among genes and incorporating this information into the list representation.

In this paper, we have focused on stability with respect to sampling variations, and the exchangeability matrix was estimated with this in mind. However, we can use the proposed method to quantify the exchangeability and obtain more robust gene lists with respect to other sources of variation (e.g. different array platforms). To do this, we could define the position vectors $S_i$ of the genes from the gene rankings obtained from studying the same samples with the different array platforms and define the exchangeability from these vectors. It could also be possible to combine different exchangeability matrices to potentially account for multiple sources of variation. Another potential extension could incorporate exchangeabilities for larger gene sets than the pairs studied in this paper. This would however be more computationally demanding, and it would require an extension of the general framework to incorporate more general tensors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENT

Conflict of Interest: None declared.

REFERENCES


[Received March 15, 2011; revised June 10, 2011; accepted for publication July 19, 2011]