Integration of Ranked Lists via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies

Shili Lin[1,2,*], Jie Ding[1]

[1]Department of Statistics, [2]Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, USA.

[*]Correspondence to: Shili Lin, Department of Statistics, The Ohio State University, Columbus, OH 43210-1247, USA. Email: shili@stat.ohio-stat.edu.

SUMMARY. One of the major challenges facing researchers studying complex biological systems is integration of data from -omics platforms. Omic-scale data include DNA variations, transcriptom profiles, and RAomics. Selection of an appropriate approach for a data integration task is problem dependent, primarily dictated by the information contained in the data. In situations where modeling of multiple raw data sets jointly might be extremely challenging due to their vast differences, rankings from each data set would provide a commonality based on which results could be integrated. Aggregation of microRNA targets predicted from different computational algorithm is such a problem. Integration of results from multiple mRNA studies based on different platforms is another example that will be discussed. Formulating the problem of integrating ranked lists as minimizing an objective criterion, we explore the usage of a cross-entropy Monte Carlo method for solving such a combinatorial problem. Instead of placing a discrete uniform distribution on all the potential solutions, an iterative importance sampling technique is utilized "to slowly tighten the net" to place most distributional mass on the optimal solution and its neighbors. Extensive simulation studies were performed to assess the performance of the method. With satisfactory simulation results, the method was applied to the microRNA and mRNA problems to illustrate its utility.

KEY WORDS: Combinatorial; Data integration; Kullback-Leibler measure; Importance sampling; Optimization criterion; Prostate cancer; Target prediction.

1 INTRODUCTION

One of the major challenges currently facing researchers studying complex traits is integration of data from various -omics platforms. Omic-scale data include DNA variations (SNP arrays), transcriptom profiles (gene expression arrays), protein-DNA binding (ChIP-chips), and RAomics (microRNA arrays). Each of these contains valuable information on different aspects of the whole biological system, but more importantly, if all information is considered jointly in a truly integrated fashion, then it is anticipated that the whole is greater than the sum of its parts. The problem of data integration may be approached from two different ends, which may be conveniently labeled as "high-level" and "low-level" analysis, respectively, from a broad perspective.

"Low-level" analysis deals with multi-factorial raw data directly. One of the earliest integrated approaches is that of combining DNA variation, gene expression, and phenotypic data to dissect complex traits (Schadt et al., 2005). A popular way for studying the interplay of these diverse data types is through treating the gene expression intensity as a quantitative trait, the eQTL approach (Schadt et al., 2003; Wayne and McIntyre, 2002). A similar eQTL approach was also used to study natural variation in human gene expression and potential regulatory role of SNPs (Morley et al., 2004). Transcription regulation modeling is another area in which gene expression, ChIP-chip, and sequence data have been jointly considered (Bar-Joseph et al., 2003; Liu, Brutlag, and Liu, 2002; Sun, Carroll, and Zhao, 2006).

There are other problems that are more suitable for "high-level" analysis. They include those that have multiple same-type results from different studies and/or different data types.

For example, suppose several studies based on different platforms/data types have been undertaken to find genes that distinguish normal tissues from colon cancer tissues. Since the results are likely to have some commonality but not identical, one would be interested in integrating the results to arrive at some consensus that is more "reliable" than any of the individual result. In this case, a meta-analytic approach (high-level analysis) could be used without going all the way back to modeling the raw data, which could be extremely challenging given their vast differences (e.g., Choi et al., 2003; Fishel et al., 2007; Parmigiani et al., 2004; Rhodes et al., 2002, 2004; Wang et al., 2004). In situations where a "significance" measure is not available or unreliable/inconsistent across platforms, one may essentially be dealing with ranking data (Yuen et al., 2002; Xu et al., 2005), which is the type of data integration problem we take up in the current paper.

The problems of comparing top k lists and rank aggregation have been considered in text mining in recent years (Dwork et al., 2001; Fagin, Kumar, and Sivakumar, 2003). The bioinformatic problems that are the subject area of this article are of similar sort. In fact, DeConde et al. (2006) is one example that deals with rank aggregation in combining results from microarray gene expression data based partly on the algorithms of Dwork et al. (2001) for text mining. We briefly outline the two problems to which our proposed method will be applied. One application problem is on combining microRNA targets from multiple target prediction computational algorithms. microRNAs are short noncoding RNAs whose primary role is either to degrade mRNAs or to repress translations by negatively regulating gene expressions post-transcriptionally through binding to their mRNA targets. Currently, there exist a few computational algorithms for predicting genes whose mRNAs are targets of microRNAs (e.g. Enright et al., 2003; John et al., 2004; Krek et al., 2005; Lewis, Burge, and Bartel, 2003, 2005), but their results may vary widely (`http://www.nyas.org/ebriefreps/main.asp?intSubsectionID=3392`). Therefore, it would be useful to consolidate and filter

through these discrepant results so that researchers may have a greater degree of certainty in the validity of these targets before engaging in costly experiments. Since the sensitivity and specificity of such algorithms are largely unknown, it appears most sensible to treat each list (from one algorithm) simply as a ranked list without trying to attach some sort of "significance" measure to its members.

Another problem we will explore is on integrating results from five microarray studies that aimed at finding genes that were differentially expressed between prostate tumors and normal prostate tissues (Dhanasekaran et al., 2001; Luo et al., 2001; Singh et al., 2002; True et al., 2006; Welsh et al., 2001). Despite their common goal, these five studies differ in the laboratories in which the experiments were carried out, in the materials and in the microarray platforms used, which make it difficult to perform a "low-level" data integration analysis given that the raw measures are not directly comparable. Instead, we will apply our "high-level" meta-analytic method to combine the five lists of top-25 up-regulated genes identified from each of the individual studies. Following the protocol of DeConde et al. (2006) for gene selection within individual studies, the input lists for our analysis are identical to theirs, facilitating comparisons of the results.

The method proposed in the current paper makes use of a global optimization technique, cross-entropy Monte Carlo (CEMC), that searches iteratively for the optimal list that minimizes a criterion. More specifically, we use as our optimization criterion the sum of weighted distances between the candidate (aggregate) list and each of the input ranked lists, although the proposed method is general and is amenable to any other optimization criterion. To measure the distance between two ranked lists, we use both the modified Kendall's tau measure and the Spearman's footrule as described in Fagin et al. (2003) in our examples and applications. However, other measures may be used in their place.

2 METHODS

2.1 Formulation of problem

Let $S_A(k)$ be the ranked list with k elements (e.g., genes), from experiment/dataset/algorithm A, depending on the context of the problem. We assume that all lists as inputs to our meta-analytic method are truncated to have the same size (i.e., same k) for ease of description of the method, although this is not a necessary condition. Hereafter we refer to such lists as top-k lists and drop the k in $S_A(k)$ where ambiguity does not occur. Let $S = \cup_A S_A$ be the union of elements from the lists, which no longer retains any ranking information in the original lists. Suppose there are a total of n elements in the set, indexed from 1 through n without any particular order of preference. Let τ be any size k ordered subset of S. Then, the goal is to find τ^* such that

$$\tau^* = \arg\min \{\Phi(\tau), \tau \subset S = (1, \cdots, n)\} = \arg\min \left\{ \sum_A w_A d(\tau, S_A), \tau \subset S \right\},$$

where w_A is a prespecified weight for A, and d is a distance measure, such as the Spearman's footrule or the modified Kendall's tau distance (see next subsection), between two top-k lists. Such a τ^* is referred to as the top-k list of the aggregate candidate targets set, and $y^* = \Phi(\tau^*)$ is the optimal (minimum) value of the objective function Φ. The weight parameters (w_A's) can be regarded as a means for incorporating prior informations regarding the reliability of an experiment/dataset or its compatibility with the other datasets. The weights will be set to uniform if such information is unavailable, and the results will be referred to as unweighted.

2.2 Two distance measures

Two well-known distances can be used in the computation of the objective criterion, the modified Kendall's tau distance and Spearman's footrule distance (Fagin et al., 2003). For each element $t \in \tau \cup S_A$, let $R_{S_A}(t) = R(t)I(t \in S_A) + (k+1)I(t \notin S_A)$, where for $t \in S_A$, $R(t)$

is its original rank, and I is an indicator function that takes the value of 1 if the argument in the parentheses is true, and 0 otherwise. Similarly, define $R_\tau(t) = R(t)I(t \in \tau) + (k+1)I(t \notin \tau)$ for each $t \in \tau \cup S_A$. With these definitions of notation, the Spearman's footrule distance is then

$$d_F(\tau, S_A) = \sum_{t \in \tau \cup S_A} |R_\tau(t) - R_{S_A}(t)|.$$

For defining the modified Kendall's tau distance $d_K(\tau, S_A)$, we let

$$d_K(t, u) = I([R_{S_A}(t) - R_{S_A}(u)][R_\tau(t) - R_\tau(u)] < 0)$$

for each pair of $t, u \in \tau \cup S_A$. Then

$$d_K(\tau, S_A) = \sum_{t,u \in \tau \cup S_A} d_K(t, u).$$

Note that, in contrast to Spearman's footrule distance, the actual rankings of the elements in the lists do not configure into the definition of Kendall's distance, only their orderings matter.

2.3 Cross entropy Monte Carlo

Obviously, the problem as described in 2.1 is combinatorial in nature as the number of possible τ's is $\binom{n}{k}k!$, and thus an exhaustive search will not be tractable even for moderate size top-k lists. To select the top-k list τ^* from the aggregate gene set through optimizing Φ, we propose to adopt the cross entropy (CE) Monte Carlo (MC) approach (Rubinstein, 1997; Rubinstein and Kroese, 2004), which have been used to solve difficult combinatorial problems. Although CEMC has mainly been used to solve problems in engineering and computer science, it was successfully adapted recently for the tagging SNP selection problem in genomics (Liu, Lin, and Tan, 2006).

Let $\boldsymbol{X} = (X_{jr})_{n \times k}$ be a random matrix with each component variable X taking values of 0 or 1, and with the constraints that each column sums to 1 and each row sums to at most 1.

Let $\boldsymbol{v} = (p_{jr})_{n\times k}$ denote the corresponding probability matrix, with the constrains that each column sums to 1. Then each column variable, $\boldsymbol{X}_r = (X_{1r}, X_{2r}, \cdots, X_{nr})$, follows a multinomial distribution with a sample size of 1 and the probability vector $\boldsymbol{v}_r = (p_{1r}, \cdots, p_{nr})$, and with the above stated constraints on the joint column variables. Thus, the pmf can be specified as follows:

$$\begin{aligned} & P_{\boldsymbol{v}}(\boldsymbol{X} = \boldsymbol{x} = (x_{jr})_{n\times k}) \\ \propto & \prod_{r=1}^{k}\prod_{j=1}^{n} (p_{jr})^{x_{jr}}\, I\left(\sum_{j=1}^{n} x_{jr} = 1, r = 1, \cdots, k; \sum_{r=1}^{k} x_{jr} \leq 1, j = 1, \cdots, n\right). \end{aligned} \tag{1}$$

In this case, a realization of $\boldsymbol{X}$, $\boldsymbol{x}$, uniquely determines the corresponding candidate top-k list without the need to reference the probability matrix. That is,

$$\tau = f(\boldsymbol{x}) = (x_{jr} | x_{jr} = 1, j = 1, \cdots, n, r = 1, \cdots, k).$$

In words, the "1" entries in each of the k columns make up the top-k list, in that order. Given the 1-1 correspondence between τ and $\boldsymbol{x}$, finding τ^* is equivalent to finding $\boldsymbol{x}^*$ that minimizes $\Phi(f(\boldsymbol{x}))$.

An exhaustive search can be viewed as placing a discrete uniform distribution on all the possible candidate lists as each is being evaluated exactly once. On the other hand, the idea of finding $\boldsymbol{x}^*$ using CEMC is to iteratively update the parameter matrix $\boldsymbol{v}$ such that, iteration by iteration, $P_{\boldsymbol{v}}(\boldsymbol{x})$ will place more and more of its probability mass on the $\boldsymbol{x}$'s that are in the "neighborhood" of $\boldsymbol{x}^*$. Loosely speaking, $\boldsymbol{x}$ is called a neighbor of $\boldsymbol{x}^*$ if the corresponding value of the objective function, $y = \Phi(f(\boldsymbol{x}; \boldsymbol{v}))$, is close to the minimum y^*. Suppose $\boldsymbol{v}$ is the current estimate of the parameter matrix. We choose the next parameter update $\boldsymbol{v}'$ to minimize the cross entropy (i.e., Kullback-Leibler measure) $CE(Q^*, P_{\boldsymbol{v}'})$ between the distributions $P_{\boldsymbol{v}'} = P_{\boldsymbol{v}'}(\boldsymbol{x})$ and Q^*, where Q^*, given in the following, is the ideal (but unobtainable) importance sampling distribution for estimating rare probability

$B = P_{\boldsymbol{v}}(\Phi(f(\boldsymbol{x};\boldsymbol{v})) \leq y)$:

$$Q^*(\boldsymbol{x}) = \frac{I(\Phi(f(\boldsymbol{x};\boldsymbol{v}) \leq y) P_{\boldsymbol{v}}(\boldsymbol{x})}{B}.$$

One can show that, after some simple algebra, minimizing $CE(Q^*, P_{\boldsymbol{v}'})$ is equivalent to maximizing

$$\sum_{\boldsymbol{x}} [I(\Phi(f(\boldsymbol{x};\boldsymbol{v})) \leq y) \log P_{\boldsymbol{v}'}(\boldsymbol{x})] P_{\boldsymbol{v}}(\boldsymbol{x}) = E_{\boldsymbol{v}}[I(\Phi(f(\boldsymbol{x};\boldsymbol{v})) \leq y) \log P_{\boldsymbol{v}'}(\boldsymbol{x})], \tag{2}$$

which is now free of the probability, B, to be estimated.

Suppose $\boldsymbol{x}_i = (x_{ijr})_{n\times k}, i = 1, \cdots, N$ is a sample drawn from $P_{\boldsymbol{v}}(\boldsymbol{x})$ with the current parameter specification $\boldsymbol{v}$, with their corresponding candidate top-k lists denoted as $\tau_i = f(\boldsymbol{x}_i), i = 1, \cdots, N$. Then

$$\begin{aligned} \boldsymbol{v}_{\text{new}} &= \arg\max_{\boldsymbol{v}'} \left\{ \frac{1}{N} \sum_{i=1}^{N} I(\Phi(f(\boldsymbol{x}_t;\boldsymbol{v})) \leq y) \log P_{\boldsymbol{v}'}(\boldsymbol{x}_t) \right\} \\ &= \left(\frac{\sum_{i=1}^{N} I(\Phi(\tau_i) \leq y) x_{ijr}}{\sum_{i=1}^{N} I(\Phi(\tau_i) \leq y)} \right)_{j=1,\cdots,n; r=1,\cdots,k,} \end{aligned} \tag{3}$$

can be used in the update for the next parameter matrix $\boldsymbol{v}'$. In practice, a weighted average of $\boldsymbol{v}$ and $\boldsymbol{v}_{\text{new}}$ as $\boldsymbol{v}'$ can better balance the rate of convergence and the chance of not being trapped in a local minimum. Furthermore, the threshold value y will also be updated iteratively, which will be configured into the updating scheme of $\boldsymbol{v}$, as detailed in the specific algorithm below. This exercise will lead to the construction of a sequence, $y_0, y_1, \cdots$, which can be proved to converge to a value (y_∞) close to y^*, following Margolin (2005). Similarly, $\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots$ will converge to $\boldsymbol{v}_\infty$, with the corresponding $P_{\boldsymbol{v}_\infty}(\boldsymbol{x})$ placing most of its probability mass on the $\boldsymbol{x}$'s that satisfy $\Phi(f(\boldsymbol{x};\boldsymbol{v})) \leq y_\infty$.

2.4 The OEA algorithm

We now describe the details of the algorithm. The general idea of CEMC may lead to multiple algorithms for performing the same task. We refer to the current algorithm as the

Order Explicit Algorithm (OEA) since the orders of the elements in the optimal list are explicitly specified in the probability matrix. Here are the five steps of the algorithm.

1. Set $\boldsymbol{v}^0$ with each $p^0_{jr} \in (0,1)$ such that $\sum_j p^0_{jr} = 1, r = 1, \cdots, k$. For instance, $p^0_{jr} = 1/n$ for all j and r indicates that each element in the aggregate set can be selected into each position in the integrated top-k list equally likely. We denote this non-informative initial probability matrix by $\boldsymbol{v}^0_{NI}$. Information from the input lists may be used to construct an informative one. For example, one may base on the number of matches and/or the sum of the rankings to arrive at an initial assignment, which will be called $\boldsymbol{v}^0_I$. Multiple starting points may then be obtained by setting $\boldsymbol{v}^0 = (1-\pi_0)\ \boldsymbol{v}^0_{NI} + \pi_0\ \boldsymbol{v}^0_I$ by varying the weight parameter $\pi_0 \in [0,1]$. Set t =0.

2. Draw a sample $\boldsymbol{x}_i = (x_{ijr})_{n\times k}, i = N_1+1, \cdots, N$, from $P_{\boldsymbol{v}^t}(\boldsymbol{x}),\ N > N_1 \geq 0$. Combined with $\boldsymbol{x}_{(l)}, l = 1, \cdots, N_1$, the N_1 realizations from $P_{\boldsymbol{v}^{t-1}}(\boldsymbol{x})$ that correspond to the N_1 smallest values of the objective function, we have a sample of size N. From this sample we find the corresponding top-k list candidates and their objective values, τ_i, and $\Phi(\tau_i), i = 1, \cdots, N$. Sort them in ascending order such that $\Phi_{(1)} \leq \ldots \leq \Phi_{(N)}$. Let $\lfloor \rho N \rfloor$ be the integer part of ρN ($\rho < 1$), then $y^t = \Phi_{(\lfloor \rho N \rfloor)}$ is the sample ρ-quantile of the objective function.

3. Use the same sample we update the parameter vector $\boldsymbol{v}^{t+1}$ as follows:

$$\boldsymbol{v}^{t+1} = \left(p^{t+1}_{jr}\right)_{n\times k} = (1-\pi)\ \boldsymbol{v}^t + \pi\ \boldsymbol{v}_{\text{new}}, \tag{4}$$

   where $\boldsymbol{v}_{\text{new}}$ is as defined in (3) but based on $\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(N_1)}, \boldsymbol{x}_{N_1+1}, \cdots, \boldsymbol{x}_N$, and $0 < \pi \leq 1$ is the weight parameter.

4. If $||\boldsymbol{v}^{t+1} - \boldsymbol{v}^t|| < \varepsilon$ then go to Step 5; otherwise set $t = t+1$ and go back to step 2. Note that $||.||$ denotes a norm such as the sum of squared component distances. In all

our applications, we chose to use $||\boldsymbol{v}^{t+1} - \boldsymbol{v}^t|| = \frac{1}{nk}\sum_{j,r}|p_{jr}^{t+1} - p_{jr}^t| < \varepsilon$, and $\varepsilon \leq 0.01$.

5. Output $y = \Phi_{(1)}$ and the corresponding ordered subset, which will be taken as the estimate of our top-k list from the aggregate candidate set.

The choices for the running parameters, including N, N_1, ρ, and π, depends on a number of factors, including the size of the top-k list and the number of candidate elements (n) in the aggregate set. These parameters, to a large extend, control the efficiency of the algorithm as well as the ability of the algorithm to escape from a local minimum, and thus should be chosen carefully. Based on the recommendation of Liu et al. (2006) and on experience gained from our analyses, we would suggest setting $N \geq kn$ and $N_1 = \frac{1}{c_1}N$ (c_1 is a constant around 10). The weight parameter π plays a similar role as N_1, and we have found that setting π to be between 0.25-0.5 seems to work well. For ρ, setting it very close to 0 will trap the algorithm in its local minimum, but setting it too large will lead to computational inefficiency. It appears that setting ρ in the range of 0.05-0.1 works adequately. Last, but not the least, is the choice of the starting probability matrix $\boldsymbol{v}^0$. Using the mixture of an initial guess based on information from the input rankings and a non-informative uniform prior appears to be adequate, but other strategies for devising a starting value may prove to be more effective.

3 EXAMPLES

To illustrate the performance of OEA, we consider the following two contrive datasets. The first is a smaller problem where complete enumeration is possible so that the true answer is known. The second is a more realistic dataset in terms of the size k. As such, enumeration of all possible output top-k lists is not feasible, but the input top-k lists are designed in such a way that the answer can be "guessed" to a large extend from the data using intuition.

3.1 Ice cream flavors - a toy example

Suppose there are five ice cream flavors, 1 = Chocolate, 2 = Vanilla, 3 = Strawberry, 4 = Butter Pecan, and 5 = Coffee, available for tasting. Three tasters were asked to rank their favorite top three for a market research project. Suppose the rankings from the three tasters are: $S_{A1} = (1, 2, 3), S_{A2} = (3, 5, 1)$, and $S_{A3} = (1, 3, 5)$. Then the aggregate favorite set is $S = \{1, 2, 3, 5\}$. We wish to find the top-3 list τ^* that minimizes $\Phi(\tau) = d(\tau, S_{A1}) + d(\tau, S_{A2}) + d(\tau, S_{A3})$. Note that we set the w's in the objective function to be uniform since we do not have any prior information to suggest how reliable a taster's choice may be in representing the opinion of the population. In this case, there are 24 possible top-3 rankings. For both the Kendall's tau and the Spearman's footrule distances, $\tau^* = (1, 3, 5)$ achieves the minimum value of the objective function: $\Phi(\tau^* = (1, 3, 5))$ equals to 4 and 8, respectively. To run the CEMC algorithm, we initialized $\boldsymbol{v}^0 = \boldsymbol{v}^0_{NI}$. We set the running parameters to be $N = 20, N_1 = 10, \rho = 0.1$, and $\pi = 0.5$. Note that N is much smaller than what we recommend relative to N_1, but it worked fine for this small example. The OEA converged quickly in this example, taking 9 iterations to reach the stopping rule and the correct identification of τ^* with either distance measure. Figure 1 shows the sequence of the probability matrices from the starting value to the one at convergence.

3.2 A larger test dataset

We now consider the problem of obtaining an integrated top-40 list from three individual ranked lists given in Table 1. As can be seen from the table, there are 65 genes in the aggregate candidate set, with 20 of them (genes 1-20) appear in all three lists, 5 in common in each of the three pairs of the lists, and 30 of them only present in one of the studies. Since the number of potential aggregate top-40 lists is more than 5×10^{65}, it is impossible to evaluate all of them to find the τ^* that optimizes the objective function. However, it

would be quite reasonable to guess that the 35 genes that are present in at least two of the individual lists should be included in the optimal integrated top-40 list, with their ordering being $(1-30, 46-50)$. It is uncertain, though, as to what the last five targets in the optimal top-40 list may be.

Application of OEA, based on either the Spearman's footrule distance or the Kendall's tau distance, yielded τ^*'s that included the partial ranked list $(1-30, 46-50)$ as its top-35 genes under most of the various parameter settings. Specifically, we investigated eight combinations of parameter settings for each of the two distances, performing 10 runs (with 10 random seeds) under each setting. The Monte Carlo sample size was set to be $N = 2000$, or 4000, and $N_1 = N/10$. For π, we set it to be 0.25 or 0.5, and $\rho = 0.1$. Three starting configurations were also explored, with $\pi_0 = 0, 0.25, 0.5$. Results from this large number of runs shew that τ^* as described above were always obtained with informative priors ($\pi_0 > 0$) regardless of the other parameter settings or the distance measure used. For those runs that started with the uninformative prior ($\pi_0 = 0$), there were a few of them that did not reach any of the τ^* when the stopping rule was met. In terms of computational intensity, a typical run that achieved an optimal result took 37 iterations to converge. The number of iterations needed was considerably larger (more than 150 in most cases) for the few runs that did not reach τ^*.

3.2.1 Four permuted datasets

To make the problem harder, we also permuted the ordering of each of the individual top-40 lists. The ordering of the genes in the integrated top-40 list could no longer be guessed, but it is fairly safe to suggest that the 35 genes present in at least two of the studies should still be included in such a list. Four random permutations of the three individual top-40 rankings were generated as four new (harder) datasets, and OEA with 10 random seeds was

applied to each of them. In three of the four datasets, genes $\{1-30, 46-50\}$ were included in almost all of the integrated top-40 lists, especially when informative starting probability matrix was used. These results were obtained under the same parameter settings as in the original (non-permuted) data, except that only Spearman's footrule distance was used. For the fourth permuted set, runs with $N = 2000$ did not yield results as good as those from the other three datasets; increasing N to 4000 led to similar results, though. For these permuted datasets, a typical run took more than twice the number of iterations needed for the original data.

4 APPLICATIONS

We present two examples in this section to demonstrate the utility of OEA for tackling bioinformatic problems. The first problem is on transcriptomics, where we are faced with the problem of integrating lists of genes found to be up-regulated in five prostate cancer studies. The second is RAomics related, in which we are interested in aggregating results from several microRNA target prediction programs.

4.1 Up-regulated mRNA expressions

The first six columns of Table 2 present the rankings and the top-25 ranked genes that were found in DeConde et al. (2006) to be up-regulated in prostate tumors compared to normal prostate tissues from five studies (Dhanasekaran et al., 2001; Luo et al., 2001; Singh et al., 2002; True et al., 2006; Welsh et al., 2001), labeled in the column headings. As indicated earlier, these five studies relied on different technologies, and their results show that they are quite discrepant in the genes selected to be included in the top-25's. As can be seen from the table, the gene list from Luo et al. (2001) is the least common with the other four studies.

Specifically, the average number of commonly selected genes between each of the five studies and the other four are 2.75, 6.75, 7, 5, and 6, for the Luo(L), Welsh(W), Dhana(D), True(T), and Singh(S) studies, respectively. As such, one might wish to down weigh the contribution of the Luo study in forming the integrated list. This would be justifiable if there were known cause of concern, such as if the quality of the data were in doubt or if the tumor samples were known, a priori, to be heterogeneous from those in the other studies. In this illustrative example, in addition to running OEA with uniform weight (i.e., unweighted), we also explore a weighting scheme for the purpose of demonstrating its influence on the results. Specifically, we set $w_L = 0.5$ for Luo and gave the weight of 1.0 to each of the other four studies.

The results given in the last four columns of Table 2 show that all four aggregate lists (Kendall unweighted (KUW) and weighted (KW), and Spearman unweighted (SUW) and weighted (SW)) rank HPN(LWDTS), the only gene present in all the five individual top-25 lists, as their top-ranked gene. Note that each letter in the parentheses after the gene name denotes the corresponding study that contains the gene in its top-25 list. There are three genes present in four, and four genes present in three, of the five individual lists, and they all appear in the top half of the aggregate lists. However, note that genes present in more individual lists do not necessarily have to be ranked above those that appear in fewer individual lists, if their individual rankings are relatively low. For instance, FASN(WDS) is present in three of the individual lists with rankings of 5(W), 3(D), and 9(S). On the other hand, NME1(LWDT) appears in four of the lists, but with lower rankings: 14(L), 12(W), 14(D), and 9(T). As such, FASN is ranked higher in any of the aggregate list than NME1. There are other similarities among the four aggregate lists. For example, three of the four lists have identical rankings for the top four genes, while the fourth list (KW) also has the same four genes, albeit with the rankings switched for GDF15 and FASN. A closer inspection shows that switching the rankings of these two genes in either KUW or KW in fact leads

to the same objective function, a reassuring sign that OEA had performed satisfactorily. Overall, the top 16-genes in all four lists are the same, apart from some shuffling of rankings. The rank correlation between the relevant pairs are quite high: (KUW, SUW) = 0.89, (KW, SW) = 0.84, (KUW, KW) = 0.84, and (SUW, SW) = 0.95.

Despite general similarities in the top rank genes in the four aggregate lists, the difference between the Kendall's tau and the Spearman's footrule, and between the weighted and the unweighted, are apparent, especially in the lower ranked genes. Kendall appears to prefer genes that are ranked highly in an individual list, while Spearman looks for consensus. For example, Kendall unweighted included five genes selected by the Luo study only but were ranked highly there, at 3-7, whereas only two singly ranked genes were selected by Spearman unweighted. As evident from comparing Kendall unweighted and weighted, and Spearman unweighted and weighted, genes that were included in Luo may be pushed down on the aggregate list when weighting is activated, such as NME1 and FMO5. Furthermore, CYP1B, ranked No. 3 in Luo, was the only gene made it to Kendall weighted, compared to five genes in unweighted. All in all, it appears that the top ranked entries are fairly stable regardless of the distance measure or the weighting scheme, while there are more movements in the lower ranked genes, which may not be trustworthy anyway due to information degradation, as to be elaborated further in Discussion.

4.2 microRNA targets

Three top-60 putative target lists of miR-155, $S_{\mathrm{MR}}, S_{\mathrm{TS}}$, and S_{PT}, were obtained from a version of miRanda (MR) (Enright et al., 2003; John et al., 2004), TargetScan (TS) (Lewis et al., 2003, 2005), and PicTar (PT) (Krek et al., 2005), respectively. These lists are given in Table 3. We applied OEA to select from the combined target set S $(= S_{\mathrm{MR}} \cup S_{\mathrm{TS}} \cup S_{\mathrm{PT}})$ the top-60 genes that minimizes the weighted sum of distances between this and each of the

individual top-60 lists. All three algorithms are, in part, based on the concept of sequence complementarity, but it is expected that targets predicted by TS and PT would be more similar due to their common "seed" concept on bases 2-8 of the microRNA. As such, the targets predicted from TS and PT are less independent than those from MR, and this closer correlation needs to be taken into account in some way. To this end, we decided to down weigh the contributions from TS and PT by setting $w_{\mathrm{MR}} = 0.4$, and $w_{\mathrm{TS}} = w_{\mathrm{PT}} = 0.3$. Our results are shown in the last two columns of Table 3. There are nine genes that were commonly predicted by all three algorithms, and they were all among the top genes of the integrated lists except SMARCA4, which was ranked No. 34 based on Spearman's. There are 22 additional genes that are in the top-60 lists of two of the algorithms, but four of them (OLFML3, ETNK2, ETS1 and SIM2) whose individual rankings are low, appear in Spearman's but not in Kendall's list. This is consistent with our observation from the mRNA data that Spearman's prefers consensus while Kendall's would go for those that are highly ranked in individual lists. Nevertheless, as with the mRNA data, the top ranked genes are quite consistent in their relative rankings regardless of the distance measure.

5 DISCUSSION

In this paper, we propose the use of a cross entropy Monte Carlo algorithm for integrating results from multiple studies/algorithms via rank aggregation. Extensive simulation studies were performed to assess the performance of the OEA algorithm. Applications of the method to integrating results from five prostate cancer studies, and to integrating microRNA targets predicted by three computational algorithms, demonstrate the utility of the method. In fact, the cross entropy Monte Carlo methodology with its specific setup as in our examples can be applicable to a wider range of problems. Integration of peptides from different prediction

algorithms in proteomics would be one of such examples. The details of our setup and OEA algorithm are very similar to those contained in a recent paper (Pihur, Datta, and Datta 2007). That paper clearly followed an earlier version of our work presented in 2006 in the International Biometrics Conference (Lin, Ding and Zhou, 2006).

Compared to the results of DeConde et al. (2006) on the same mRNA data, our appear to be more sensible. The top 16 genes in our four aggregate lists are the same, and each is present in at least two of the five individual studies. On the other hand, the top 16 genes in each of the three aggregate lists of DeConde contain quite a number of genes selected by only one study in the top-25 list, and they are ranked in the aggregate lists above some that have appeared in multiple individual lists and have higher individual rankings. Equally counter intuitive is the situation where genes ranked lower in the same studies may receive higher rankings in the aggregate lists. For instance, both FASN and KRT18 were ranked by the same three studies in their top-25 lists with rankings of {5(W), 3(D), 9(S)} and {7(W), 9(D), 11(S)}, respectively. However, KRT18 was ranked higher than FASN, in each of the three aggregate lists, despite the lower rankings in each of the three individual lists. While the higher rankings for KRT18 in MCT and Thurstone might be explained by the relative rankings of these two genes in the other two studies, borrowing the argument put forth by DeConde, the result nevertheless cannot be explained following the same argument for the "majority-rule" algorithm MC4. In fact, switching the positions of FASN and KRT18 would result in an decrease of 3, 11, and 11, for MC4, MCT, and Thurstone, respectively, in the value of the minimization criterion based on Kendall's distance, the same criterion, apart from a scale, used by DeConde to gauge the relative performance of their three algorithms. These, along with other similar observations show that the lists obtained by DeConde using any of the three algorithms are sub-optimal, according to several optimization criteria, and they are inferior to any of those obtained in our study (Figure 2). On the other hand, It is

reassuring to see that each of our four integrated lists minimizes the corresponding criterion. Perhaps this is not surprising since the two MC algorithms (MC4 and MCT) in DeConde are heuristic in nature, whereas Thurstone is model-based, whose assumptions (such as unit variance and independence of genes) may no be valid.

Our CEMC algorithm is iterative in nature, and with a number of tuning parameters. As such, careful selection of starting points and parameter values is essential to minimize the chance of converging to a local, rather than the global, minimum. In practice, multiple runs with various values of the starting matrix and tuning parameters would be highly recommended; the outcome that renders the smallest value of the objective function could then be regarded as the optimal result. This, in spirit, is similar to the recommendation of using multiple starting points when running the iterative EM algorithm (McLachlan and Peel, 2000).

As with any ranking problem, the ranks of elements in a set with high rankings are much more reliable than those ranked in the bottom. In the context of our problem, switching some of the rankings in the bottom would lead to very similar, or even the same, values of the objective function. This is indicative of the flat surface of the objective function around the minimal, and as such, there is very little information to distinguish between lists with rearrangements in the bottom. This is especially true for the criteria based on Spearman's footrule, which may simply select an aggregate rank that is the same as one of the individual rank in those that are ranked low in the list. Therefore, interpretation of the rankings of the elements in the bottom of a ranked list should be done with caution, as information for ranking may have degraded to such an extent that the ranks assigned to those elements may be very inaccurate. In this regard, it would be useful to associate each entry in the ranked list with a confidence (or information content) measure of some sort.

The main purpose of our applications to the microRNA and to the mRNA data is to

demonstrate the utility of the proposed method. The choice of miR-155 and the prostate cancer studies was mainly due to convenience as the rankings from each individual study in both data sets were readily compiled and available to us. There is an added advantage in using the mRNA data as our results can be directly compared to those from the source of the compiled data. For the microRNA application, it seems appropriate to combine the diverse predicted targets from different computational algorithms to arrive at an aggregate list that is more informative for down-stream experiments. We chose to use the predicted targets from three algorithms to illustrate the method and the results, but predicted targets from other programs (e.g., Kiriakidou et. al., 2004; Vatolin et al., 2006; Wang and Wang, 2006; Huynh et al., 2006) can also be integrated. For the mRNA data, we by no means suggest that combining all five studies in this way is the best approach to integrate these results as there may still be heterogeneity in the samples among the studies despite metastatic prostate cancer samples were excluded. Similarly, our choice to down weigh the Luo study should be viewed simply as an illustration on the impact of the weighting scheme rather than as a suggestion that the Luo study was less reliable. We could, as an alternative, choose to weight each study inversely proportional to its effective sample size. Nevertheless, the top genes in the aggregate lists (e.g., HPN, AMACR, FASN) all have been suggested to be involved in prostate cancer development and progression (Klezovitch et al., 2004; Kuefer et al., 2002; Pizer et al., 2001). On another note, although the two applications are on integration of data of the same type, the method may also be applied to aggregate results from studies of different data types or results from different analysis methods of the same data set.

As mentioned earlier, the general framework of CEMC may lead to the proposals of multiple algorithms tailored for specific problems. Although our OEA algorithm appears to work satisfactorily for the problems investigated, other well-designed algorithms may prove to be even more efficient than OEA. Finally, we wish to note that our procedure is applicable to

variable lengths of individual and aggregate lists. However, due to information degradation for ranking those that are not ranked highly, it would not make sense to have a aggregate list longer than any individual list.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants DMS-0112050 (to the MBI) and DMS-0306800 (to S.L.).

REFERENCES

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003) Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, **21**, 1337-1342.

Choi, J.K., Yu, U., Kim, S., and Yo, O.J. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84-i90.

DeConde, R.P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., and Etzioni, R. (2006) Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology*, **5**, article 15.

Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.S., and Chinnaiyan,A.M. (2001) Delineation of prog-nostic biomarkers in prostate cancer. *Nature*, **412**, 822-826.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001) Rank aggregation meth-ods for the web. http://www10.org/cdrom/papers/577/.

Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C., and Marks,D.S. (2003) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.

Fagin,R., Kumar,R., and Sivakumar,D. (2003) Comparing top k lists. *SIAM Journal of Discrete Mathematics*, **17**, 134-160.

Fishel, I., Kaufman, A., and Ruppin, E. (2007) Meta-analysis of gene expression data: a predictor-based approach. *Bioinformatics*, **23**, 1599-1606

Fishel, Irit1; Kaufman, Alon1; Ruppin, Eytan1

John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C., and Marks,D.S. (2004) Human microRNA targets. *PLOS Biology*, **2**, 1862-1879.

Kiriakidou,M., Nelson,P.T., Kouranov,A., Fitziev, P., Bouyioukos,C., Mourelatos,Z., and Hatzigeorgiou,A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes and Development*, **18**, 1165-1178.

Klezovitch, O., Chevillet, J., Mirosevich, J., Roberts, R., Matusik, R., and Vasioukhin, V. (2004) Hepsin promotes prostate cancer and metastasis. *Cancer Cell*, **6**, 185- 195.

Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M., and Rajewsky,N. (2005) Combinatorial microRNA target predictions. *Nature Genetics*, **37**, 495-500.

Kuefer, R., Varambally, S., Zhou, M., Lucas, P.C., Loeffler, M., Wolter, H., Mat-tfeldt, T., Hautmann, R.E., Gschwend, J.E., Barrette, T.R., Dunn, R.L., Chin-naiyan, A.M., and Rubin, M.A. (2002) alpha-Methylacyl-CoA racemase: expression levels of this novel cancer biomarker depend on tumor differentiation. *American Journal of Pathology*, **161**, 841-848.

Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P., and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.

Lewis,B.P., Burge,C.B., and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15-20.

Lin, S., Ding, J., and Zhou, J. (2006) Rank aggregation of putative microRNA targets with cross-entropy Monte Carlo methods. Paper presented in the International Biometrics Conference in June 2006. Montreal, Canada.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, **20**, 835-9.

Liu, Z., Lin, S., and Tan, M. (2006) Genome-wide tagging SNPs with entropy-based Monte Carlo methods. *Journal of Computational Biology*, **13**, 1606-1614.

Luo, J., Duggan, D.J., Chen, Y., Sauvageot, J., Ewing, .M., Bittner, M.L., Trent, J.M., and Isaacs, W.B. (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Research*, **61**, 4683-4688.

Margolin, L. 2005. On the convergence of the cross-entropy method. *Annals of Operations Research*, **134**, 201-214.

McLachlan, G., and Peel, D. (2000) Finite Mixture Models. Wiley, New York.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743-7.

Parmigiani, G., Garrett-Mayer, E.S., Anbazhagan, R., and Garielson, E. (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research*, **10**, 2922-2927.

Pihur, V., Datta, S., and Datta, S. (2007) Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics*, **23**, 1607-1615.

Pizer, E.S., Pflug, B.R., Bova, G.S., Han, W.F., Udan, M.S., and Nelson, J.B. (2001) Increased fatty acid synthase as a therapeutic target in adrogen-independent prostate cancer progression. *Prostate*, **47**, 102-110.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, **62**, 4427- 4433.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Bar-rette, T., Pandey, A., and Chinnaiyan, A.M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the USA*, **101**, 9309-9314.

Rubinstein, R.Y. 1997. Optimization of computer simulation models with rare events. *European Journal of Operations Research*, **99**, 89-112.

Rubinstein, R.Y., and Kroese, D.P. 2004 The Cross-Entropy Method. A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer, UK.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and

Friend, S. H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297-302.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusis, A. J. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, **37**, 710-7.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., DAmico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., and Sellers, W.R. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203-209.

Sun, N., Carroll, R. J., and Zhao, H. (2006) Bayesian error analysis model for reconstructing transcriptional regulatory networks. *Proceedings of the National Academy of Sciences of the USA*, **103**, 7988-93.

True, L., Coleman, I., Hawley, S., Huang, A., Gifford, D., Coleman, R., Beer, T., Gelman, E., Datta, M., Mostaghel, E., Knudsen, B., Lange, P., Vessella, R., Lin, D., Hood, L., and Nelson, P. (2006) A Molecular Correlate to the Gleason Grading System for Prostate Adenocarcinoma. *Proceedings of the National Academy of Sciences of the USA*,

Wang, J., Coombes, K.R., Highsmith, W.E., Keating, M.J., and Abruzzo, L.V. (2004) Differences in gene expression between B-cell chronic lymphcytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics*, **20**, 3166-3178.

Wayne, M. L., and Mcintyre, L. M. (2002) Combining mapping and arraying: An approach

to candidate gene identification. *Proceedings of the National Academy of Sciences of the USA*, **99**, 14903-6.

Welsh, J.B., Sapinoso, L.M., Su, A.I., Kern, S.G., Wang-Rodriguez, J., Moskaluk, C.A., Frierson, H.F.Jr., and Hampton,G.M. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research*, **61**, 5974-5978.

Xu L., Tan A.C., Naiman D.Q., Geman D., and Winslow R.L. (2005) Robust can-cer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, **21**, 3905-3911.

Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J., and Sealfon, S.C. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA mi-croarrays. *Nucleic Acids Research*, **30**, e48.

Table 1: Gene names (renamed as numbers) of top-40 lists from three algorithms.

	Rank														
Algorithm	1	2	$\cdots$	20	21	$\cdots$	25	26	$\cdots$	30	31	$\cdots$	35	36 $\cdots$	40
A1	1	2	$\cdots$	20	21	$\cdots$	25	26	$\cdots$	30	31	$\cdots$	35	36 $\cdots$	40
A2	1	2	$\cdots$	20	21	$\cdots$	25	41	$\cdots$	45	46	$\cdots$	50	51 $\cdots$	55
A3	1	2	$\cdots$	20	56	$\cdots$	60	26	$\cdots$	30	46	$\cdots$	50	61 $\cdots$	65

Table 2: Individual top-25 ranked genes from five prostate cancer studeis and their aggregate ranks from the OEA algorithm based on four optimization criteria.

Rank	Luo	Welsh	Dhana	True	Singh	Kendall		Spearman	
						Unweighted	Weighted	Unweighted	Weighted
1	HPN	HPN	OGT	AMACR	HPN	HPN(LWDTS)	HPN(LWDTS)	HPN(LWDTS)	HPN(LWDTS)
2	AMACR	AMACR	AMACR	HPN	SLC25A6	AMACR(LWDT)	AMACR(LWDT)	AMACR(LWDT)	AMACR(LWDT)
3	CYP1B1	OACT2	FASN	NME2	EEF2	FASN(WDS)	GDF15(WDTS)	FASN(WDS)	FASN(WDS)
4	ATF5	GDF15	HPN	CBX3	SAT	GDF15(WDTS)	FASN(WDS)	GDF15(WDTS)	GDF15(WDTS)
5	BRCA1	FASN	UAP1	GDF15	NME2	UAP1(WDS)	KRT18(WDS)	NME2(TS)	NME2(TS)
6	LGALS3	ANK3	GUCY1A3	MTHFD2	LDHA	OACT2(WD)	EEF2(DTS)	UAP1(WDS)	OACT2(WD)
7	MYC	KRT18	OACT2	MRPL3	CANX	KRT18(WDS)	UAP1(WDS)	OACT2(WD)	KRT18(WDS)
8	PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA	SLC25A6(TS)	NME2(TS)	SLC25A6(TS)	SLC25A6(TS)
9	WT1	GRP58	KRT18	NME1	FASN	NME1(LWDT)	OACT2(WD)	KRT18(WDS)	UAP1(WDS)
10	TFF3	PPIB	EEF2	COX6C	SND1	EEF2(DTS)	GRP58(WS)	EEF2(DTS)	EEF2(DTS)
11	MARCKS	KRT7	STRA13	JTV1	KRT18	STRA13(WD)	NME1(LWDT)	STRA13(WD)	STRA13(WD)
12	OS-9	NME1	ALCAM	CCNG2	RPL15	NME2(TS)	STRA13(WD)	NME1(LWDT)	NME1(LWDT)
13	CCND2	STRA13	GDF15	AP3S1	TNFSF10	CANX(WS)	SLC25A6(TS)	CANX(WS)	SND1(DS)
14	NME1	DAPK1	NME1	EEF2	SERP1	SND1(DS)	CANX(WS)	ALCAM(DS)	ALCAM(DS)
15	DRRK1A	TMEM4	CALR	RAN	GRP58	GRP58(WS)	ALCAM(DS)	GRP58(WS)	GRP58(WS)
16	TRAP1	CANX	SND1	PRKACA	ALCAM	ALCAM(DS)	SND1(DS)	SND1(DS)	CANX(WS)
17	FMO5	TRA1	STAT6	RAD23B	GDF15	PPIB(WD)	PPIB(WD)	FMO5(LD)	SLC19A1(WD)
18	ZHX2	PRSS8	TCEB3	PSAP	TMEM4	TMEM4(WS)	TMEM4(WS)	TMEM4(WS)	TMEM4(WS)
19	RPL36AL	EMTPD6	EIF4A1	CCT2	CCT2	CYP1B1(L)	MTHFD2(TS)	CCT2(TS)	CCT2(TS)
20	ITPR3	PPP1CA	LMAN1	G3BP	SLC39A6	MTHFD2(TS)	MRPL3(WT)	PRKACA(T)	MRPL3(WT)
21	GCSH	ACADSB	MAOA	EPRS	RPL5	ATF5(L)	CCT2(TS)	MTHFD2(TS)	EPRS(T)
22	DDB2	PTPLB	ATP6V0B	CKAP1	RPS13	MRPL3(WT)	SLC19A1(WD)	PTPLB(W)	PPIB(WD)
23	TFCP2	TMEM23	PPIB	LIG3	MTHFD2	BRCA1(L)	JTV1(T)	PPIB(WD)	MTHFD2(TS)
24	TRAM1	MRPL3	FMO5	SNX4	G3BP2	LGALS3(L)	CYP1B1(L)	MRPL3(WT)	FMO5(LD)
25	YTHDF3	SLC19A1	SLC7A5	NSMAF	UAP1	MYC(L)	COX6C(T)	SLC19A1(WD)	SLC7A5(D)

Table 3: Individual top-60 targets from three prediction programs and their aggregate ranks from the OEA algorithm based on two optimization criteria.

Rand	MR	TS	PT	Kendall	Spearman
1	BACH1	ZNF537	ZNF537	ZNF537	ZNF537
2	SEMA5A	BACH1	BACH1	BACH1	BACH1
3	OR1K1	FGF7	IKBKE	FGF7	FGF7
4	JARID2	ZNF652	ASTN2	ASTN2	ASTN2
5	JARID1B	FLJ30435	AKAP10	ZIC3	JARID1B
6	SGKL	RAB11FIP2	MGC13272	SGKL	SGKL
7	OPRM1	ZIC3	FBXO11	RNF123	ZIC3
8	RAPH1	MIDN	KBTBD2	PAPOLA	RAPH1
9	CLCN5	ARID2	COL7A1	TP53INP1	SPI1
10	SPI1	PAPOLA	MAP3K10	SPI1	PAPOLA
11	TP53INP1	KIBRA	RNF123	SOCS1	RNF123
12	STAC	PICALM	FLJ14299	FBXO11	STAC
13	CARHSP1	CAB39	SF3B1	CARHSP1	CARHSP1
14	PSKH1	CEBPB	KPNA1	CEBPB	CEBPB
15	MAP3K14	RNF123	CARHSP1	MAP3K14	MAP3K14
16	HBP1	SOCS1	EHD1	SMARCA4	SOCS1
17	AGTRAP	SOX10	PAPOLA	SEMA5A	AGTRAP
18	KCNN3	SUFU	RAB34	KCNN3	KCNN3
19	TRIM32	MAP3K7IP2	GCN5L2	MLSTD2	TRIM32
20	CSNK1G2	GPM6B	SOCS1	CSNK1G2	CSNK1G2
21	SOCS1	MLR1	ZIC3	OR1K1	FBXO11
22	CYR61	ASTN2	MLSTD2	JARID2	CYR61
23	DUSP14	FBXO11	SALL1	KIBRA	DUSP14
24	MGP	MAP3K14	ACTA1	CSF1R	MGP
25	SLC11A2	C10orf26	LRP1B	C10orf26	SLC11A2
26	FGF7	TOMM20	C1QL2	OPRM1	OPRM1
27	PCDH9	C3orf18	LOC51161	RREB1	PCDH9
28	DHX40	TLE4	CSNK1G2	CLCN5	DHX40
29	KIAA0258	SGKL	FGF7	AICDA	KIAA0258
30	CEBPB	BCAP29	ADD3	JARID1B	OR1K1
31	GCET2	CARHSP1	MYB	FBXO33	GCET2
32	ARL8	MYO1D	OLFML3	EHD1	ARL8
33	SEL1L	CSF1R	ARNTL	PSKH1	CSF1R
34	ZIC3	FLJ37543	COPS3	MGC13272	SMARCA4
35	RREB1	SOX1	SMARCA4	HBP1	RREB1
36	MYPN	UPP2	SPRED1	AGTRAP	MYPN
37	RNF123	TP53INP1	CEBPB	RAPH1	TP53INP1
38	C21orf107	MYO10	ARVCF	STAC	C21orf107
39	STARD6	SPI1	KCNN3	IKBKE	STARD6
40	CDC37	DNCI1	CSF1R	LRP1B	CDC37
41	AICDA	PRO1855	ZNF642	KBTBD2	AICDA
42	SATB2	LOC284058	KIBRA	COL7A1	KIBRA
43	FBXO33	ETS1	YWHAE	TRIM32	FBXO33
44	RPS6KA3	MLSTD2	ETNK2	RAB11FIP2	MLSTD2
45	FOS	GDF6	182-FIP	MAP3K10	FOS
46	BIRC4	KIAA1889	SEPT11	CYR61	BIRC4
47	NKX3-1	SMARCA4	SCG2	AKAP10	NKX3-1
48	EIF2C4	RREB1	UBQLN1	FLJ14299	LRP1B
49	LRP1B	ZNF236	ETS1	KPNA1	ETS1
50	KIAA1411	DNAJB1	SIM2	DUSP14	SIM2
51	CHAF1A	FLJ20273	KIAA0863	RAB34	CHAF1A
52	C10orf26	SIM2	AXOT	SLC11A2	C10orf26
53	SGCZ	FBXO33	WIT-1	ZNF652	SGCZ
54	SDCBP	SBLF	ENTH	SF3B1	SDCBP
55	SMARCA4	SGCB	SKI	FLJ30435	JARID2
56	CUGBP2	AICDA	H3F3A	MIDN	CUGBP2
57	HRPT2	CARD11	MAP3K14	SALL1	HRPT2
58	MEIS1	ETNK2	SLA	PICALM	ETNK2
59	MEF2A	OLFML3	KIAA1274	ARID2	OLFML3
60	SYPL	EHD1	NDFIP1	CAB39	EHD1

$$\boldsymbol{v}^0 = \begin{bmatrix} 0.250 & 0.250 & 0.250 \\ 0.250 & 0.250 & 0.250 \\ 0.250 & 0.250 & 0.250 \\ 0.250 & 0.250 & 0.250 \end{bmatrix} \rightarrow \boldsymbol{v}^1 = \begin{bmatrix} 0.425 & 0.225 & 0.225 \\ 0.125 & 0.175 & 0.375 \\ 0.325 & 0.375 & 0.125 \\ 0.125 & 0.225 & 0.275 \end{bmatrix} \rightarrow \boldsymbol{v}^2 = \begin{bmatrix} 0.613 & 0.213 & 0.113 \\ 0.063 & 0.088 & 0.388 \\ 0.263 & 0.588 & 0.063 \\ 0.063 & 0.113 & 0.438 \end{bmatrix} \rightarrow$$

$$\boldsymbol{v}^3 = \begin{bmatrix} 0.756 & 0.156 & 0.056 \\ 0.031 & 0.044 & 0.294 \\ 0.181 & 0.744 & 0.031 \\ 0.031 & 0.056 & 0.619 \end{bmatrix} \rightarrow \boldsymbol{v}^4 = \begin{bmatrix} 0.878 & 0.078 & 0.028 \\ 0.016 & 0.022 & 0.147 \\ 0.091 & 0.872 & 0.016 \\ 0.016 & 0.028 & 0.809 \end{bmatrix} \rightarrow \boldsymbol{v}^5 = \begin{bmatrix} 0.939 & 0.039 & 0.014 \\ 0.008 & 0.011 & 0.073 \\ 0.045 & 0.936 & 0.008 \\ 0.008 & 0.014 & 0.905 \end{bmatrix} \rightarrow$$

$$\boldsymbol{v}^6 = \begin{bmatrix} 0.970 & 0.020 & 0.007 \\ 0.004 & 0.006 & 0.037 \\ 0.023 & 0.968 & 0.004 \\ 0.004 & 0.007 & 0.952 \end{bmatrix} \rightarrow \boldsymbol{v}^7 = \begin{bmatrix} 0.985 & 0.010 & 0.004 \\ 0.002 & 0.003 & 0.018 \\ 0.011 & 0.984 & 0.002 \\ 0.002 & 0.004 & 0.976 \end{bmatrix} \rightarrow \boldsymbol{v}^8 = \begin{bmatrix} 0.992 & 0.005 & 0.002 \\ 0.001 & 0.001 & 0.009 \\ 0.006 & 0.992 & 0.001 \\ 0.001 & 0.002 & 0.988 \end{bmatrix} \rightarrow$$

$$\boldsymbol{v}^9 = \begin{bmatrix} 0.996 & 0.002 & 0.001 \\ 0.001 & 0.001 & 0.005 \\ 0.003 & 0.996 & 0.001 \\ 0.001 & 0.001 & 0.994 \end{bmatrix}$$

Figure 1: Path of probability matrices - from starting ($\boldsymbol{v}^0$) to convergence ($\boldsymbol{v}^9$). The starting probability matrix is completely uninformative. The matrix at convergence leads to the top-3 list $(1, 3, 5)$ by selecting the gene corresponding to the largest probability in each column. The columns may not sum to exactly 1 due to rounding.

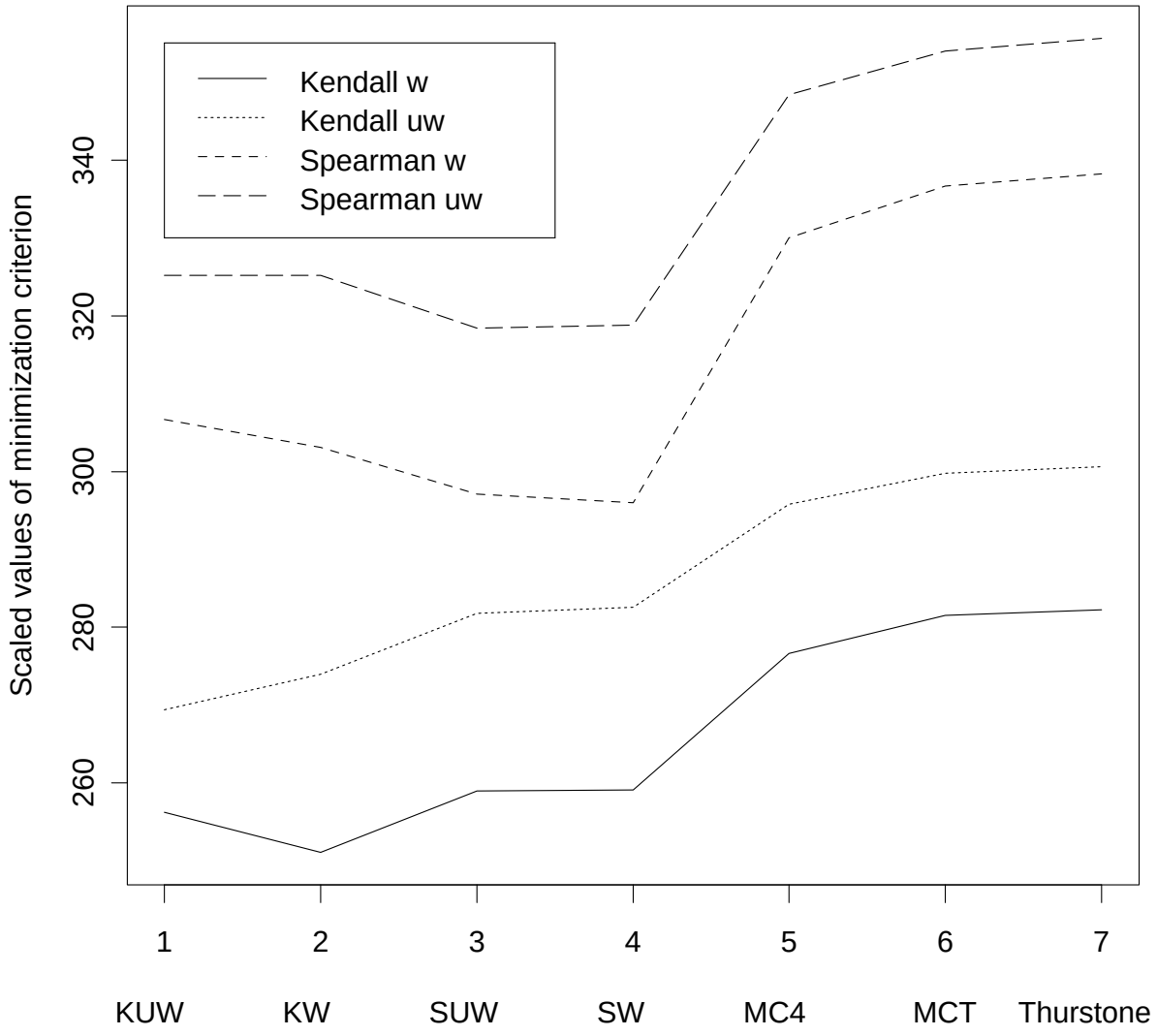

Figure 2: Comparison of seven aggregate ranked lists based on four objective criteria. KUW, KW, SUM, SW are those obtained from the OEA algorithm, while MC4, MCT, and Thurstone were from DeConde et al. (2006). The values of the objective criteria were scaled in the following way: the Kendall uw and Spearman uw criteria were divided by 5, while the Kendall w and Spearman w criteria were divided by 4.5 since the Luo study only received a weight of 0.5.